Promoter motifs facilitate as well as impede the emergence and evolution of gene expression

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. Prokaryotic gene expression requires a stretch of regulatory DNA called a promoter, which contains two promoter motifs called -10 and -35 boxes. However, these motifs may occur in both promoters and non-regulatory DNA in multiple copies. It is possible that these motifs influence both the evolution and emergence of regulatory DNA, but this has never been thoroughly tested. To understand whether and how promoter motifs influence promoter emergence and evolution, we start from 25 DNA sequences enriched with -10 and -35 box motifs (promoter islands), mutagenize the sequences, and measure gene expression driven by 240'000 mutants. We find that the probability that mutations create an active promoter varies more than 200-fold, but is not correlated with the number of promoter motifs. In most mutants (~67%), the weakening, strengthening, or emergence of a promoter is caused by mutations that modify already existing motifs. Notably, creating new motifs can either increase or decrease promoter activity. 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 and YS 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[1]. It is controlled by DNA sequences that encode motifs for transcription factors (TFs) to bind and either activate or repress transcription[2,3]. Mutations to these motifs can modulate gene expression, leading to the evolution of new phenotypes[1,4]. Mutations can also create new regulatory sequences from non-regulatory DNA, a process called *regulatory emergence*[5,6].

We define a *motif* as a DNA sequence computationally predicted to be compatible with TF binding. We specifically distinguish between a motif and a binding site, because binding sites refer to the location of where a TF binds in the genome, rather than a DNA sequence, and correspond to only a fraction of all motif locations in a genome[7–9].

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

[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

evidence shows that motifs can facilitate promoter 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[10]. Likewise, transposable elements contain motifs compatible with the binding of particular TFs[11,12], and these motifs have biased gene regulatory networks towards utilizing these TFs in evolution[12].

On the other hand, there is also evidence that regulatory motifs can impede promoter emergence and evolution. One reason for this is that different regulatory proteins can compete in their binding to shared motifs[13,14]. Another reason is that gene expression can be silenced when regulatory DNA contains multiple motifs for the same TF[1516–18]. 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[16]. 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[19–21].

In prokaryotes the primary regulatory sequences are called promoters[22,23]. The canonical promoter recruits the RNA polymerase subunit, σ70[22,24,25], to transcribe target genes. These promoters consists of two *promoter motifs* called the -10 box and the -35 box[22,26,27] whose sequences are very well characterized (consensus: `TATAAT` and `TTGAAA`, respectively)[27–33]. We also know how the DNA sequences of promoter motifs, as well as their distances (consensus 17±1 bp) influence gene expression[13,22,27,34–36]. Like in humans[37] and invertebrates[38], most prokaryotic promoters contain multiple regulatory motifs[22], and have multiple transcription start sites[39]. 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[13,28].

Recent work shows that promoters can emerge from promoter motifs or adjacent to them[40]. However, encoding -10 and -35 box motifs is insufficient to drive transcription. For instance, the *E. coli* genome contains clusters of -10 and -35 box motifs with little or no promoter activity. Such clusters are called *promoter islands*[18,41–43]. One study found that promoter activity can either emerge from a promoter island when a -35 box motif is mutated to match the consensus sequence, or counterintuitively, when -10 and -35 box motifs are destroyed[18].

Thus, it is neither understood how additional promoter motifs influence promoter activity, nor is it clear when a promoter motif will facilitate or impede de-novo promoter emergence.

Here, we systematically study the ability of -10 and -35 box motifs to influence the evolution and emergence of promoters. To this end, we select 25 fragments from 25 promoter island sequences, with both regulatory and non-regulatory activity, and screen more than 240'000 mutant variants for promoter activity. We find that ~67% of all promoter activity originates from -10 and -35 box motifs, but the likelihood of mutations creating new promoters does not correlate with the number of motifs. Additionally, mutations can create additional -10 and -35 box motifs to either increase or decrease promoter expression. These results highlight the complex interplay of promoter motifs throughout regulatory evolution and the emergence of new phenotypes.

RESULTS

Parent sequences are enriched with motifs for -10 and -35 boxes.

Promoter islands are ~1kbps (kilo base pairs) long[18,41–43], 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 box motifs for further analysis (**Fig 1A**). To this end, we used position-weight matrices (PWMs) – computational tools to identify promoter motifs (Methods)[44]. We refer to these 25 sequences

as *parent sequences* (P1-25). On average, each parent contains ~4.22 -10 box motifs and ~5.14 -35 box motifs on each of its two DNA strands (**Fig S1**).

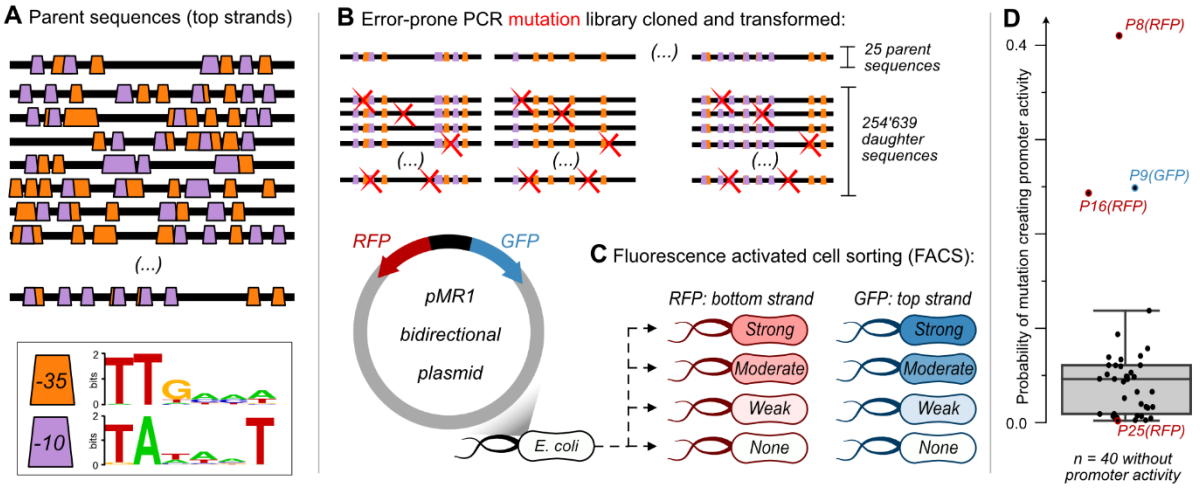


Figure 1. Mutating parent sequences reveals vastly different probabilities of promoter emergence. (A) The location of -10 and -35 boxes in a subset of the parent sequences. See **Fig S1** for the complete set (n=25). Orange trapezoids correspond to -35 box motifs, and magenta trapezoids to -10 box motifs, 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 parent sequences from 25 promoter islands in the *E. coli* genome using 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)** Circles show the probability that mutation creates an active promoter from an inactive parent. Data is shown for each of 40 inactive parents and orientations (i.e., 25 parents ×2 orientations = 50. 50 – 3 top strand promoters – 7 bottom strand promoters = 40.) The height of the box represents the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR.

Inactive promoters vary widely in their potential to become active promoters.

Next we created a mutant library of *daughter* sequences from the 25 parent sequences 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 [30,45]. Specifically, we cloned the library into a bidirectional fluorescence reporter plasmid (pMR1) to measure promoter activity on both DNA strands (top: GFP, bottom: RFP)[46]. 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 parent and orientation (top/bottom) and ~2.5-point mutations per daughter (**Fig S2**). Each parent has 3×150 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 parents.

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 an *active promoter* if its score exceeded 1.5 a.u., as this score lies at the boundary between no fluorescence and weak fluorescence based on the sort-seq bins (methods).

Some sequences in our library are unmutated parent sequences, which allowed us to determine that 8/25 of the parent sequences already encode active promoters before mutagenesis. Specifically, one parent drove expression on the top strand (P22), five did on the bottom strand (P6, P12, P13, P18, P21), and two drove expression on both strands (P19 and P20). See **Fig S4** for fluorescence score distributions for each parent and its daughters, and **Data S2** for all daughter sequence fluorescence scores.

We refer to the remaining 40 parent sequences and orientations as inactive. Although mutating each of the 40 inactive parent sequences could create a promoter, the likelihood P_{new} that a mutation creates an active promoter varies dramatically among parents. For each inactive parent and orientation, **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) and the highest 0.41 (PI7), a 205-fold difference.

These P_{new} values do not significantly correlate with the number of daughter sequences, the GC-contents of the parents, and – most notably – with the number of -10 or -35 box motifs (least squares linear regression, p=0.529, 0.930, 0.852, 0.834 respectively) (**Fig S5**). Thus, the number of promoter motifs does not directly influence the likelihood of promoter emergence.

Promoters primarily emerge and evolve from specific subsets of -10 and -35 box motifs. Next we asked where both active 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[29,30,47], as well as DNA regions where promoters most readily emerge[40]. The calculation effectively converts all 245'639 data points into maps that reveal where active 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[40] (see methods). For example, two such hotspots exist in parent P19, which is an active promoter that drives expression from both DNA strands (**Fig 2B**). One hotspot overlaps with a -35 box motif and the other overlaps with a -10 box motif. 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. Additionally, the number of hotspots that a sequence encodes also correlates modestly with the probability P_{new} (least squares linear regression, $p=1.35\times10^{-4}$, $r^2 = 0.322$) (**Fig S5**). In

other words, the more hotspots a parent has, the more likely it is that mutations create a new promoter from it.

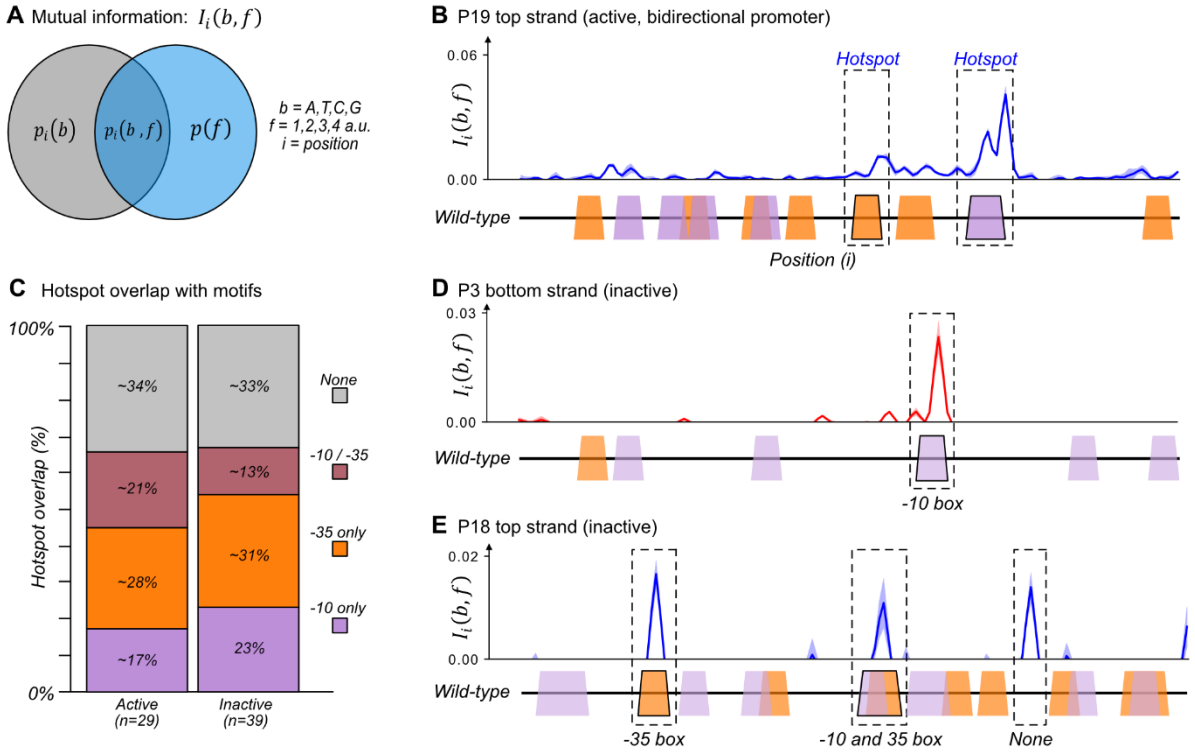


Figure 2. The majority of promoters emerge and evolve within a subset of preexisting promoter motifs. (A) We calculated the mutual information $I_i(b,f)$ between nucleotide identity (b = A,T,C,G) and fluorescence scores rounded to the nearest whole number (f= *1,2,3,4 a.u.*) for each position i in a 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 box motifs. Top: we plot the mutual information $I_i(b,f)$ for P19's top strand. P19 is an active promoter on both DNA strands. 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 box motifs (magenta trapezoids) and -35 box motifs (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 box motifs only (magenta), -35 box motifs only (orange), both -10 and -35 box motifs (red), or with neither (gray). We plot this both for active parents (left) and inactive parents (right). **(D)** Analogous to (B) but for the bottom strand of P3. P3 is an inactive parent sequence. Hotspot overlaps with a -10 box motif. **(E)** Analogous to (B) but for the top strand of P18. P18 is an inactive parent sequence. Hotspots overlap with (from left to right) a -35 box motif, both a -10 and a -35 box motif, and neither (None). **Fig S6** shows analogous mutual information plots for daughters derived from each parent sequence.

Across all 25 parent sequences and for both expression directions (top and bottom strand), we identified a total of 68 hotspots **(Fig 2C**). 29 hotspots occur in the parent sequences with promoter activity. They correspond to locations where mutations either increase or decrease promoter activity. ~17% (5/29) of these hotspots overlap exclusively with -10 box motifs, ~28% (8/29) with -35 box motifs, and ~21% (6/29) overlap with both -10 and -35 box motifs. The remaining ~34% of hotspots overlap with neither a -10 nor a -35 box motif (10/29).

The remaining 39 hotspots occur in inactive parent sequences. Mutations in these locations solely increase fluorescence. Of these 39 hotspots, ~23% (9/39) overlap with -10 box motifs, ~31% (12/39) with -35 box motifs, ~13% with both a -10 and -35 box motifs (5/39), and ~33% with neither motif (13/39) (**Fig 2C-E**). See **Fig S6** for the 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 box motifs that do not overlap with hotspots. Of the 211 -10 box motifs, only 25 (~12%) overlap with a hotspot. Similarly, of the 257 -35 box motifs, only 31 (~12%) overlap with a hotspot. For example, the inactive parent P3 contains five -10 box motifs, but only one overlaps with a hotspot (**Fig 2D**). Similarly, the inactive P18 contains seven -10 and -35 box motifs each, but promoters only emerge in three hotspots (**Fig 2E**). In sum, the majority (~67%) of new promoters evolve and emerge from motifs for -10 and -35 boxes, but the presence of a motif does not indicate that promoter activity can emerge or is encoded within the motif.

Promoters can emerge when mutations create motifs but not by destroying them.
We hypothesized that promoters emerge when mutations create -10 and -35 box motifs. 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 motif in the hotspot. We then asked whether gaining a -10 or -35 box motif leads to a significant increase in fluorescence that indicates the creation of a new promoter (Mann-Whitney U-test, **Fig. S7** and methods)

We identified 5 and 4 hotspots in which gaining a -10 box motif or a -35 box motif creates a de-novo promoter by increasing fluorescence by more than +0.5 a.u. (**Fig 3A**). For ~67% (6/9) of these hotspots, the locations of the newly created -10 and -35 box motifs overlap with preexisting -10 and -35 box motifs.

As examples, we highlight two hotspots involving the creation of a -10 box. Parent P16 gains promoter activity when mutations create a -10 box downstream of a preexisting -35 box motif (**Fig 3B**), increasing median fluorescence by +1.99 a.u. (**Fig 3C**) (two-tailed MWU test, $q=7.64\times10^{-32}$). We call the act of mutations creating a -10 box downstream of a preexisting -35 box motif *Shiko emergence*[40]. The term is a homage to the Sumo exercise Shiko, where the wrestler firmly plants their foot firmly on the floor, and lowers their other foot to a firm position. In contrast, parent P9 gains promoter activity when mutation creates a new -10 box that does not lie downstream of an existing -35 box motif (**Fig 3D,E,** median fluorescence gain +1.8 a.u., two-tailed MWU test, $q=2.83\times10^{-173}$). See **Fig S8** for the other hotspots in which -10 boxes are created.

We also highlight one hotspot where gaining a -35 box creates de-novo promoter activity. Specifically, parent P1 gains promoter activity when mutations create a -35 box upstream of a preexisting -10 box motif **(Fig 3F**), which increases fluorescence by +0.74 a.u. (**Fig 3G**, two-tailed MWU test, $q=4.25\times10^{-42}$). This is also an example of Shiko Emergence (Shiko -35). See **Fig S8** for the other hotspots in which -35 boxes are created. We did not find any hotspots where a -35 box is created without a downstream -10 box motif.

Prompted by an earlier report[18], we also asked whether mutations that destroy -10 or -35 box motifs can increase promoter activity. No such mutations exist in our data (see **Fig 3A**).

To summarize, we found 9 hotspots where mutations created -10 and -35 boxes that lead to de-novo promoter emergence. When -10 boxes are gained, the majority (three of five) form downstream of a preexisting -35 box motif (Shiko -10 emergence). The remaining two -10

boxes do not form near a -35 box motif (gain -10, **Fig 3H**). Four of the four newly formed -35 boxes form upstream of a -10 box motif (Shiko -35). We did not find any hotspots where creating a -35 box without a downstream -10 box motif leads to promoter emergence. In the majority of the 9 hotspots (6 of 9), new -10 or -35 boxes overlap preexisting -10 and -35 box motifs.

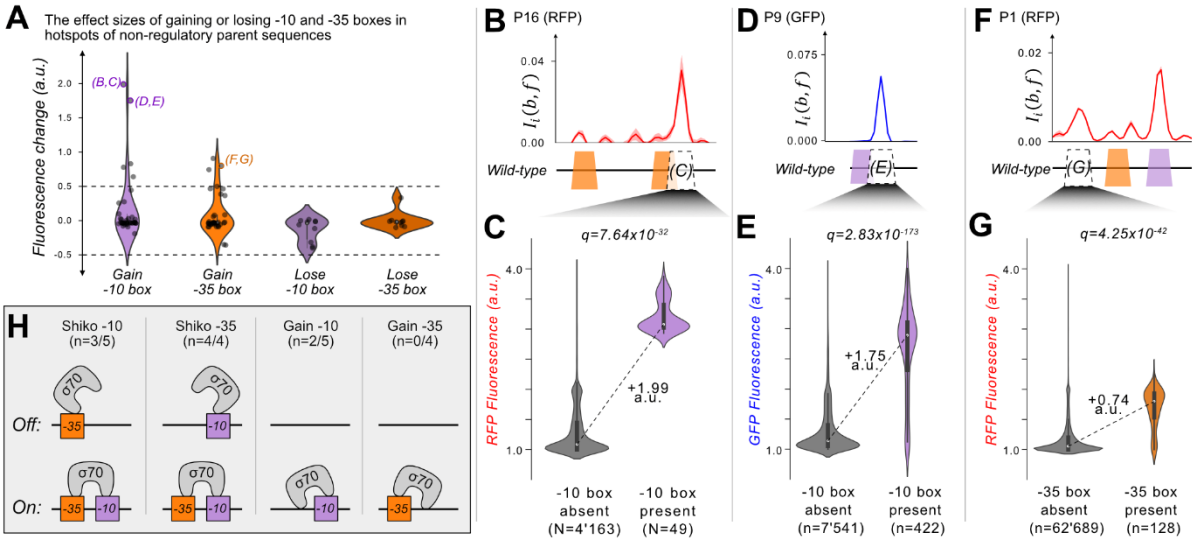


Figure 3. Gaining -10 and -35 boxes that overlap with promoter motifs creates de-novo promoters. (A) The change in fluorescence (arbitrary units, a.u.) when gaining or losing -10 and -35 box motifs in mutual information hotspots of inactive parent sequences (see **Fig S7** for calculation overview). Dotted lines indicate an effect size threshold of ± 0.5 a.u. Each circle corresponds to a gain or loss of a box motif 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. **(B)** The bottom strand of parent P16. 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 box motifs (magenta trapezoids) and -35 box motifs (orange trapezoids) along the wild-type parent sequence. Dotted trapezoid indicates the region of interest analyzed in the subsequent panel. **(C)** Fluorescence scores (a.u.) for daughter sequences without (left) and with (right) a -10 box motif in the region of interest defined in (B). 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, $q=7.64\times10^{-32}$). 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. **(D)** Analogous to (B) but for the top strand of parent P9. **(E)** Analogous to (C) except for the region of interest defined in (D) (two-tailed MWU test, $q=2.83\times10^{-173}$). **(F)** Analogous to (B) but for the bottom strand of parent P1. **(G)** Analogous to (C) except for gaining a -35 box instead of a -10 box in the region of interest defined in (F) (two-tailed MWU test, $q=4.25\times10^{-42}$). **(H)** Promoter emergence models from left to right. Promoters emerge by gaining -10 boxes downstream of preexisting -35 box motifs (Shiko -10), gaining -35 boxes upstream of preexisting -10 box motifs (Shiko -35), gaining a -10 box independent of an upstream -35 box motif, or gaining a -35 box upstream independent of a preexisting -10 box motif. See **Fig S8** for additional examples.

Gaining boxes can increase or decrease promoter activity (by more than one mechanism)

We next asked how gaining and losing -10 and -35 box motifs can change expression driven by active parent sequences. To this end, we repeated the previous analysis but for the active parents.

We first identified 5 and 3 hotspots in which losing a -10 box motif or -35 box motif *decreases* expression (**Fig S9A**). These particular hotspots revealed where promoters are located in each parent. For example, for parent P12 (**Fig S9**), losing a -35 box decreases fluorescence

290 by -1.57 a.u. (**Fig S9**) (two-tailed MWU test, $q=6.28\times10^{-67}$), and losing a -10 box decreases
291 fluorescence by -1.41 a.u. (two-tailed MWU test, $q=6.28\times10^{-91}$). Based on these data and the
292 close spacing (18 bp) of these two motifs, we can infer that they constitute the binding sites
293 for the promoter in P12 (**Fig 4A**). See **Fig. S9** for additional examples. We did not find any
294 hotspots in which destroying a -10 box motif or a -35 box motif increases expression (see **Fig**
295 **S9A**).

296 We also found 4 and 1 hotspot(s) in which gaining a -10 or -35 box further increases
297 expression. More surprisingly, in 4 and 6 hotspots gaining a -10 or -35 box *decreases*
298 expression (n= 4+1+4+6, 15 hotspots with increased or decreased expression) (see **Fig S9A**).
299 To understand how this is possible, we analyzed each parent sequence where this occurs in
300 detail. This analysis allowed us to group such events into three categories. That is, mutation
301 can 1) create a *tandem motif* that lowers expression, 2) shift -10 and -35 boxes to modulate
302 expression, or 3) create additional -10 and -35 box motifs that destroy the original promoters
303 and lower expression (**Fig 4**). Next we briefly describe examples of each category.

304 For 2 of the 15 hotspots (~13%), mutations create a *tandem motif,* where two -10 or -35 box
305 motifs occur in head-to-tail orientation. For example, when mutations of the aforementioned
306 parent sequence P12 create a -35 box motif upstream of the existing -35 box, median
307 fluorescence decreases by 1.1 a.u. (two-tailed MWU test, $q=3.35\times10^{-5}$) (**Fig 4B**). Similarly,
308 when mutations in P12 create a -10 box motif upstream of the mapped -10 box, median
309 fluorescence also decreases by 0.61 a.u. (two-tailed MWU test, $q=4.36\times10^{-4}$) (**Fig 4C**).

310 For 6 of the 15 hotspots (40%), mutations shift the location of -10 and -35 boxes. For example,
311 mutations can shift the -10 box of parent P12 +1 bp away from the -35 box, which increases
312 median expression by +0.78 a.u. (**Fig 4D**) (two-tailed MWU test, $q=5.85\times10^{-13}$). See **Fig S10**
313 for additional hotspots where such shifting occurs (**Fig 4E**).

314 For 3 of the 15 hotspots (20%), we find that mutations can create -10 or -35 box motifs
315 overlapping -10 and -35 boxes from the putative promoters we mapped in **Figure S9**. For
316 example, for parent P22, where our mutational data had allowed us to map the active -35 and
317 -10 boxes (**Fig 4F**), mutations can change the -35 box into a -10 box motif, which decreases
318 expression by 0.90 a.u. (**Fig 4G**) (two-tailed MWU test, $q=7.29\times10^{-7}$). Thus, while the loss in
319 expression appears to be caused by a gain of a motif, this gain also entails of a loss of an
320 existing, active box (**Fig 4H**). See **Fig S10** for additional examples.

321 The remaining 4 of 15 hotspots (~27%) are located in P20 (**Fig. S11**), which encodes a
322 bidirectional promoter. However, because our mutational data did not allow us to confidently
323 map this promoter on either strand of P20, we cannot resolve why gaining -10 or -35 box
324 motifs may increase expression (in 2 of the 4 P20 hotspots) or decrease expression (in the
325 remaining 2 hotspots, **Fig. S11**) of this parent sequence.

326

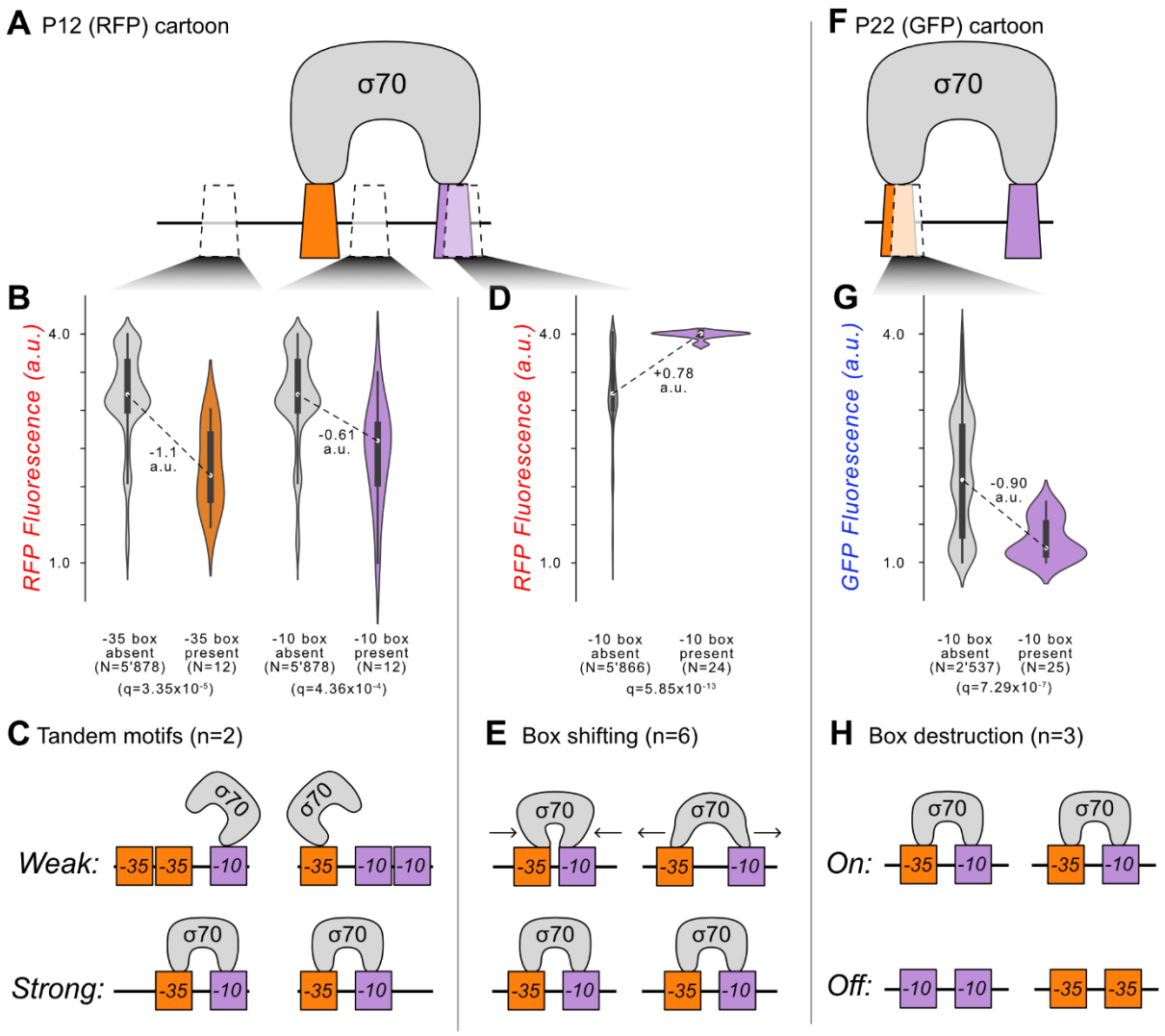


Figure 4. Gaining -10 and -35 boxes modulates promoter activity. (A) A cartoon for the promoter mapped on the bottom strand of P12 based on data from **Fig S9.** Orange trapezoids = -35 boxes, magenta trapezoids = -10 boxes, dashed trapezoids = regions of interest in subsequent figure panels. **(B)** 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. We test this null hypothesis for daughter sequences with vs without a -35 box motif (left comparison) or with vs without a -10 box motif (right comparison) using a Mann-Whitney U (MWU) test. The q-values correspond to Benjamini-Hochberg-corrected p-values (methods; (two-tailed MWU test, $q=3.35\times10^{-5}$ and $q=4.36\times10^{-4}$, respectively). 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. Dotted lines bridging the box plots with numbers underneath correspond to the median change in fluorescence (a.u.). **(C)** A proposed mechanism whereby gaining -35 boxes adjacent to -35 boxes or -10 boxes adjacent to -10 boxes lowers expression. **(D)** Analogous to (B) but for the presence or absence of a -10 box shifted +1 bp downstream from the modeled -10 box (two-tailed MWU test, $q=5.85\times10^{-13}$). **E)** A proposed mechanism where mutations shift -10 and -35 boxes either closer or further apart to modulate promoter strength. See **Fig S10** for 5 additional examples. **(F)** Analogous to (A) but for the promoter mapped on the top strand of P20. **(G)** Analogous to (B) but for the presence or absence of a -10 box motif overlapping with the modeled -35 box (two-tailed MWU test, q=7.29×10-7). **(H)** A proposed mechanism whereby mutations create new -10 or -35 box motifs in the locations of existing promoter binding sites, destroying the promoter. See **Fig S10** for two additional examples.

DISCUSSION

We found that sequences enriched with -10 and -35 box motifs without promoter activity vary dramatically (more than 200-fold) in the probability P_{new} that mutations create active promoters in them (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 box motifs (least squares linear regression, p=0.852, 0.834 respectively). In addition, hotspots of promoter emergence coincide with only ~12% of the -10 and -35 box motifs. However, P_{new} correlates with the number of mutual information hotspots (least squares linear regression, $p=1.35\times10^{-4}$, $r^2 = 0.322$, see **Fig S5**). For instance, three parental sequences with high P_{new} (P8, P9, and P16, see **Fig 1D**) encode multiple hotspots (3,3, and 5 hotspots respectively, see **Fig S6**). In each of these hotspots, promoters emerge when mutations create new -10 and -35 boxes (see **Fig 3** and **Fig S8**). Sequences with the lowest emergence probability, such as the bottom strand of P25, contain no hotspots (**Fig S6**). Therefore, promoter emergence is best explained by how readily mutations create *additional* promoter motifs, not by the number of preexisting -10 and -35 box motifs. Additionally, ~78% (7/9 hotspots) of emergence hotspots, gain new -10 and -35 boxes adjacent to corresponding -10 and -35 box motifs (Shiko Emergence, see **Fig 3H**), suggesting that the location of new and preexisting motifs also influences promoter emergence.

-10 and -35 box motifs, however, influence where promoters emerge. We found that the majority (~67%) of all mutual information hotspots overlap with -10 or -35 box motifs (see **Fig 2C**). This can partially be explained by our finding that mutations create additional -10 and -35 boxes upstream or downstream of -35 and -10 box motifs respectively, to build promoters (see **Fig 3**). Additionally, mutations within the promoter motifs can shift their locations, change a -35 box into a -10 box motif, or vice versa (see **Fig 3** and **Fig S8**). This finding may help to explain why a previous study concluded that destroying promoter motifs creates promoter activity[18]. Specifically, destroying a -10 or -35 box motif can create a de-novo promoter, but only if this destruction simultaneously creates a new -10 or -35 box (see **Fig 3A**).

When asking how mutations change the activity of already active parent sequences, we found that their ability to shift the locations of -10 and -35 boxes is especially important. For example, in 6 hotspots, mutations that shift -10 and -35 boxes either increase or decrease expression (see **Fig 4E**). This shifting also explains why ~67% of hotspots in active sequences overlap with existing -10 and -35 box motifs. This finding is consistent with previous studies, which demonstrate that the distance between -10 and -35 boxes is an integral component of promoter strength, and that deviations from the 17±1 bp spacing can result in lower expression[13,27,34].

Promoter activity can be further modulated by the creation of tandem motifs in parent P12. Such tandem motifs do not abolish gene expression but decrease it by about one half (**Fig 4B** and **Fig 4C**). One possible explanation for this halving is that RNA polymerase can bind to both parts of the tandem motif (i.e. both -10 boxes or both -35 boxes), but only binding to one of these two yields promoter activity. To understand the biophysical mechanism behind this phenomenon is a task for future work. Regardless of this mechanism, the observation underlines that DNA sequences flanking a promoter can be important for protein binding[36], as has also been observed in eukaryotic transcriptional regulation[48–51]. Collectively, the shifting of -10 and -35 boxes, and the creation of tandem motifs 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.

Overall, our study demonstrates that mutations to -10 and -35 box motifs are the primary paths to create new promoters and to modulate the activity of existing promoters. On the one hand, this is exciting for studying de-novo genes and the evolution of novel gene expression, because it suggests that promoter emergence and evolution may be predictable[52,53]. On the other hand, our observation that promoter motifs influence where new regulatory sequences can emerge, leads to many exciting insights about regulatory emergence throughout the genome, plasmids, and transposable elements.

For instance, new enhancer sequences in eukaryotes primarily emerged from specific families of transposable elements[6,54]. Furthermore, transposons encoding particular motifs influence which TFs are utilized in a gene regulatory network[12]. Collectively, observations like these suggest that the identity of a DNA sequence heavily influences how gene regulation emerges and evolves, and thereby which novel phenotypes are most likely to arise by chance.

DATA AVAILABILITY

Data S1 is a list of the DNA sequences for the primers and promoter island parent sequences in this study. Data stored as an Excel spreadsheet.

Data S2 is a dataframe with all of the daughter sequences and their respective fluorescence scores after cleaning the dataset. Data stored as a csv file.

Data S3 is a dataframe with each identified hotspot and whether it overlaps with a -10 or -35 box. Data stored as a csv file.

Data S4 is a dataframe with the output from the gain-loss analysis for both the top and bottom strand. Data stored as a csv file.

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

A Jupyter notebook with all of the python scripts to recreate the analysis and figures in this study, as well as Data S1-S5 are available at the Github repository: https://github.com/tfuqua95/promoter_islands

The 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)

To create PWMs for -10 and -35 box motifs, we obtained a list of -10 and -35 box sequences from Regulon DB[35], and converted them into PWMs using `Biopython.motifs` from the `Biopython` package[55] (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 motif, we used the standard practice of setting a threshold equal to the information content of the PWM[44]. 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 motifs,

438 respectively. In **Fig S9** we also plot "low-affinity" -10 and -35 box motifs. For these motifs, we
439 halved the thresholds (1.99 and 1.70 bits, respectively).

440 **Amplifying parent sequences and plasmid MR1 with PCR**
441 We acquired the genomic coordinates for the promoter island sequences from ref. [41]. We used
442 a polymerase chain (PCR) reaction to amplify 78 parent sequences from the genome, using
443 Q5 polymerase (NEB, USA product #M0491). The reaction amplifies each parent sequence
444 from the genome, and additionally concatenates short sequences to the 5' and 3' ends of the
445 inserts (5'-GGCTGAATTC…insert…GGATCCTTGC-3'). These overhangs allowed us to carry out
446 the error-prone PCR with a single set of primers. See **Data S1** for the list of primers and **Fig**
447 **S12** for an overview of the PCR.

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

460 To create the mutagenesis library of parent sequences, we pooled the 78 PCR products such
461 that each product had the same concentration in the pool. We then performed a single error-
462 prone polymerase chain reaction on the pooled products to create a mutagenesis library. To
463 this end, we used GoTaq polymerase (Promega, USA, product #M3001). Specifically, we
464 combined 1 uL of two PCR primers complementary to both the pMR1 plasmid and to the
465 sequences concatenated to the ends of the initial PCR (1 uL forward and 1 uL reverse primer,
466 each at a 100 uMol l^{-1}) (see **Data S1** and **Fig S12**), 10 uL of GoTaq reaction buffer, 1 uL 10
467 mM dNTPs (Thermo Scientific, USA, product #R0191), 1 uL of pooled DNA, and 1 uL of GoTaq
468 polymerase. We added water to a total of 98 uL, then added 2 uL of 15 mMol $MnCl_2$ to the
469 reaction. We then excised and purified the product as described for the parent sequences.

470 We additionally used Q5 polymerase to create and amplify linear copies of pMR1 (see **Data**
471 **S1** for primer sequences and **Fig S12**). We carried out the necessary PCR as described for
472 the parent sequences but for an extension time of 2 minutes and 30 seconds.

473 **Molecular cloning**
474 We cloned the error-prone PCR library into pMR1 at the EcoRI and BamHI sites using
475 NEBuilder (New England Biolabs, USA, product #E2621) as follows: 100 ng of linear pMR1,
476 25 ng insert, 5 uL NEBuilder mastermix, and water to 10 uL total volume. To increase the
477 reaction efficiency, we added 1 uL of T4 DNA ligase buffer (New England Biolabs, USA,
478 product #B0202S). Using a thermal cycler (C1000 Touch Thermal Cycler, Bio-Rad, USA) we
479 incubated each assembly reaction for 1 hour at 50ºC.

480 Subsequently, we electroporated our product directly into *E. coli* DH5α electrocompetent cells
481 (Takara, Japan, product #9027) by adding 2 uL of (unpurified) product to 2× 50 uL aliquots of
482 *E. coli* (100 uL total) in a 2 mm electroporation cuvette (Cell Projects, England, product #EP-
483 202), and electroporated the cells using a Bio-Rad MicroPulser (Bio-Rad, USA). For recovery,

484 we then added 1 mL of Super Optimal Broth with Catabolite Repression Medium (SOC,
485 provided with the electrocompetent cells (Takara, Japan, product #9027), shaking the cells at
486 230 RPM at 37ºC for 1.5 hours (Infors HT, Switzerland, Multitron).

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

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

503 **Sort-Seq**
504 To perform sort-seq, we first added 100 uL of the glycerol stock of our mutagenesis library to
505 1 mL of LB-chloramphenicol (100 ug/ml). We then allowed the resulting culture to shake at
506 230 RPM and 37ºC overnight (~16 hours) on an Infors HT (Multitron, Switzerland) shaking
507 incubator. Subsequently, we centrifuged the culture and washed the cell pellet twice with
508 Dulbeco's Phosphate Buffered Saline (PBS) (Sigma, USA, D8537). We then resuspended the
509 pellet in PBS.

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

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

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

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

526 We defined the bins for GFP analogously, but with GFP and RFP controls swapped, with the
527 following exception: Because the GFP positive control produces a bimodal FITC-H distribution,

528 we defined the lower bound of bin #4 for green fluorescence as the peak of the higher mode
529 of this distribution. See **Fig S2** for an illustration of the bins.

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

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

546 After ~16 hours of overnight culture, we created glycerol stocks for each culture (8 cultures
547 total) as described in "Molecular Cloning." From the remaining culture volume, we isolated the
548 plasmids using a Qiagen QIAprep Spin Miniprep kit (Qiagen, Germany, product #27104) as
549 described by the manufacturer. We amplified the plasmid inserts using PCR as described in
550 "Molecular Cloning", using primers (**Data S1**) that add a unique multiplexing barcode onto the
551 5'-end of each PCR product from each bin. In addition, we isolated the inserts from the
552 unsorted library, and endowed them too with a unique barcode. This procedure yielded 9
553 (=8+1) sequence samples, each with a unique barcode. We then pooled equimolar amounts
554 of these samples, and sequenced them using Illumina paired-end sequencing (Eurofins
555 GmbH, Germany).

556 **Processing sequencing reads**

557 Using Flash2[56], we merged the paired-end reads that resulted from Illumina sequencing. Each
558 paired-end read contains palindromic cut-sites for 5'-EcoRI (`GAATTC`) and 3'-BamHI
559 (`GGATCC`). To orient all of the paired-end reads in the same direction, we searched each
560 paired-end read for both cut-sites, discarding any sequence that did not encode both, and took
561 the reverse complement of the read if the BamHI site was upstream of the EcoRI site. We then
562 identified the tagmented barcode sequence upstream of the EcoRI site, classifying from which
563 fluorescence bin the read originated. We then cropped the sequences to remove the barcodes,
564 BamHI, and EcoRI sites, and counted the total number of reads within every bin. We refer to
565 each of these unique paired-end read sequences as *daughter sequences.*

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

571 To map each daughter sequence to its respective template parent sequence, we calculated
572 the Hamming distance (i.e. the number of nucleotide differences between two sequences)
573 between each daughter sequence and the 78 parent sequences, and identified the parent as

the sequence with the smallest Hamming distance. To minimize our likelihood of mapping the daughter sequences to the wrong parent sequence, we removed daughter sequences from the dataset for which the closest Hamming distance was greater than 10. Because the error-prone PCR created different numbers of daughter sequences per parent sequence, we also removed all parent and their respective daughter sequences if there were less than 2'000 unique daughter sequences. After these calculations and filtration steps, our dataset contained 245'639 daughter sequences derived from 25 parent sequences.

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

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

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

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

Mutual information

To calculate mutual information between nucleotides and fluorescence scores, we first randomly split the dataset into three equal-sized pseudo-replicates (r1, r2, and r3) to minimize the effects of sampling biases as well as sorting errors when calculating mutual information. For each pseudo-replicate, we calculated the mutual information $I_i(b,f)$ between the nucleotide identity b and fluorescence score f score at each position i ($1<i<150$), of every daughter sequence using equation (2), as defined in ref. [30]:

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

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.).

ACKNOWLEDGEMENTS AND CONTRIBUTIONS

The European Molecular Biology Organization (EMBO) supports TF with a postdoctoral fellowship (ALTF 963-2021). 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

1. Prud'homme, B., Gompel, N. & Carroll, S. B. Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences* **104**, 8605–8612 (2007).

2. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**, 318–356 (1961).

669 3. Seshasayee, A. S. N., Sivaraman, K. & Luscombe, N. M. An overview of prokaryotic
670 transcription factors : a summary of function and occurrence in bacterial genomes. *Subcell Biochem*
671 **52**, 7–23 (2011).

672 4. Fuqua, T. *et al.* Dense and pleiotropic regulatory information in a developmental enhancer.
673 *Nature* **587**, 235–239 (2020).

674 5. Rebeiz, M., Jikomes, N., Kassner, V. A. & Carroll, S. B. Evolutionary origin of a novel gene
675 expression pattern through co-option of the latent activities of existing regulatory sequences.
676 *Proceedings of the National Academy of Sciences* **108**, 10036–10043 (2011).

677 6. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).

678 7. Krieger, G., Lupo, O., Wittkopp, P. & Barkai, N. Evolution of transcription factor binding
679 through sequence variations and turnover of binding sites. *Genome Res.* **32**, 1099–1111 (2022).

680 8. Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor–DNA binding: beyond binding site
681 motifs. *Curr Opin Genet Dev* **43**, 110–119 (2017).

682 9. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif
683 syntax. *Nat Genet* **53**, 354–366 (2021).

684 10. Galupa, R. *et al.* Enhancer architecture and chromatin accessibility constrain phenotypic
685 space during Drosophila development. *Developmental Cell* **58**, 51-62.e4 (2023).

686 11. Villanueva-Cañas, J. L., Horvath, V., Aguilera, L. & González, J. Diverse families of transposable
687 elements affect the transcriptional regulation of stress-response genes in Drosophila melanogaster.
688 *Nucleic Acids Res* **47**, 6842–6857 (2019).

689 12. Lynch, V. J. *et al.* Ancient Transposable Elements Transformed the Uterine Regulatory
690 Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *Cell Reports* **10**, 551–
691 561 (2015).

692 13. Lagator, M. *et al.* Predicting bacterial promoter function and evolution from random
693 sequences. *eLife* **11**, e64543.

694 14. Chauhan, V. *et al.* Analytical kinetic model of native tandem promoters in E. coli. *PLOS*
695 *Computational Biology* **18**, e1009824 (2022).

696 15. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify cis-
697 regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A*
698 **99**, 757–762 (2002).

699 16. Loell, K. J. *et al.* Transcription factor interactions explain the context-dependent activity of
700 CRX binding sites. *PLoS Comput Biol* **20**, e1011802 (2024).

701 17. Perales, M. *et al.* Threshold-dependent transcriptional discrimination underlies stem cell
702 homeostasis. *Proc Natl Acad Sci U S A* **113**, E6298–E6306 (2016).

703 18. Bykov, A. *et al.* Excessive Promoters as Silencers of Genes Horizontally Acquired by
704 Escherichia coli. *Front Mol Biosci* **7**, 28 (2020).

705 19. Crampton, N., Bonass, W. A., Kirkham, J., Rivetti, C. & Thomson, N. H. Collision events
706 between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic*
707 *Acids Research* **34**, 5416–5425 (2006).

20. Hobson, D. J., Wei, W., Steinmetz, L. M. & Svejstrup, J. Q. RNA polymerase II collision interrupts convergent transcription. *Mol Cell* **48**, 365–374 (2012).

21. Wang, L., Watters, J. W., Ju, X., Lu, G. & Liu, S. Head-on and co-directional RNA polymerase collisions orchestrate bidirectional transcription termination. *Molecular Cell* **83**, 1153-1164.e4 (2023).

22. Huerta, A. M. & Collado-Vides, J. Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. *Journal of Molecular Biology* **333**, 261–278 (2003).

23. van Hijum, S. A. F. T., Medema, M. H. & Kuipers, O. P. Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation. *Microbiology and Molecular Biology Reviews* **73**, 481–509 (2009).

24. Burgess, R. R., Travers, A. A., Dunn, J. J. & Bautz, E. K. Factor stimulating transcription by RNA polymerase. *Nature* **221**, 43–46 (1969).

25. Paget, M. S. & Helmann, J. D. The σ70family of sigma factors. *Genome Biology* **4**, 203 (2003).

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

27. Urtecho, G., Tripp, A. D., Insigne, K. D., Kim, H. & Kosuri, S. Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry* **58**, 1539–1551 (2019).

28. Urtecho, G. *et al.* Genome-wide Functional Characterization of Escherichia coli Promoters and Sequence Elements Encoding Their Regulation. *eLife* **12**, (2023).

29. Ireland, W. T. *et al.* Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *eLife* **9**, e55308 (2020).

30. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* **107**, 9158–9163 (2010).

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

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

33. Warman, E. A. *et al.* Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA-sequence symmetry. *Nat Microbiol* **6**, 746–756 (2021).

34. Klein, C. A., Teufel, M., Weile, C. J. & Sobetzko, P. The bacterial promoter spacer modulates promoter strength and timing by length, TG-motifs and DNA supercoiling sensitivity. *Sci Rep* **11**, 24399 (2021).

35. Tierrafría, V. H. *et al.* RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microbial Genomics* **8**, 000833.

36. Mitchell, J. E., Zheng, D., Busby, S. J. W. & Minchin, S. D. Identification and analysis of 'extended –10' promoters in Escherichia coli. *Nucleic Acids Res* **31**, 4689–4695 (2003).

747 37. Gotea, V. *et al.* Homotypic clusters of transcription factor binding sites are a key component
748 of human promoters and enhancers. *Genome Res* **20**, 565–577 (2010).

749 38. Lifanov, A. P., Makeev, V. J., Nazina, A. G. & Papatsenko, D. A. Homotypic Regulatory Clusters
750 in Drosophila. *Genome Res* **13**, 579–588 (2003).

751 39. Mendoza-Vargas, A. *et al.* Genome-Wide Identification of Transcription Start Sites, Promoters
752 and Transcription Factor Binding Sites in E. coli. *PLoS One* **4**, e7526 (2009).

753 40. Timothy Fuqua & Andreas Wagner. Mobile DNA is replete with hotspots for the de novo
754 emergence of gene regulation. *bioRxiv* 2023.10.22.563463 (2023) doi:10.1101/2023.10.22.563463.

755 41. Panyukov, V. V. & Ozoline, O. N. Promoters of Escherichia coli versus Promoter Islands:
756 Function and Structure Comparison. *PLOS ONE* **8**, e62601 (2013).

757 42. Purtov, Y. A. *et al.* Promoter islands as a platform for interaction with nucleoid proteins and
758 transcription factors. *J Bioinform Comput Biol* **12**, 1441006 (2014).

759 43. Shavkunov, K. S., Masulis, I. S., Tutukina, M. N., Deev, A. A. & Ozoline, O. N. Gains and
760 unexpected lessons from genome-scale promoter mapping. *Nucleic Acids Res* **37**, 4919–4931 (2009).

761 44. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant
762 alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).

763 45. Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-
764 scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).

765 46. Westmann, C. A., Alves, L. de F., Silva-Rocha, R. & Guazzaroni, M.-E. Mining Novel
766 Constitutive Promoter Elements in Soil Metagenomic Libraries in Escherichia coli. *Frontiers in*
767 *Microbiology* **9**, (2018).

768 47. Belliveau, N. M. *et al.* Systematic approach for dissecting the molecular mechanisms of
769 transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences* **115**, E4796–
770 E4805 (2018).

771 48. Afek, A., Schipper, J. L., Horton, J., Gordân, R. & Lukatsky, D. B. Protein−DNA binding in the
772 absence of specific base-pair recognition. *Proceedings of the National Academy of Sciences* **111**,
773 17140–17145 (2014).

774 49. Gordân, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding
775 specificity of bHLH transcription factors through DNA shape. *Cell Rep* **3**, 1093–1104 (2013).

776 50. Chiu, T.-P., Li, J., Jiang, Y. & Rohs, R. It is in the flanks: Conformational flexibility of
777 transcription factor binding sites. *Biophysical Journal* **121**, 3765–3767 (2022).

778 51. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325–328
779 (2015).

780 52. Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. *Nat Ecol Evol* **1**, 1–9 (2017).

781 53. Stern, D. L. & Orgogozo, V. The loci of evolution: how predictable is genetic evolution?
782 *Evolution* **62**, 2155–2177 (2008).

783 54. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature*
784 **515**, 355–364 (2014).

785 55. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular
786 biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

787 56. FLASH: fast length adjustment of short reads to improve genome assemblies | Bioinformatics
788 | Oxford Academic. https://academic.oup.com/bioinformatics/article/27/21/2957/217265.

789

SUPPLEMENTARY FIGURES

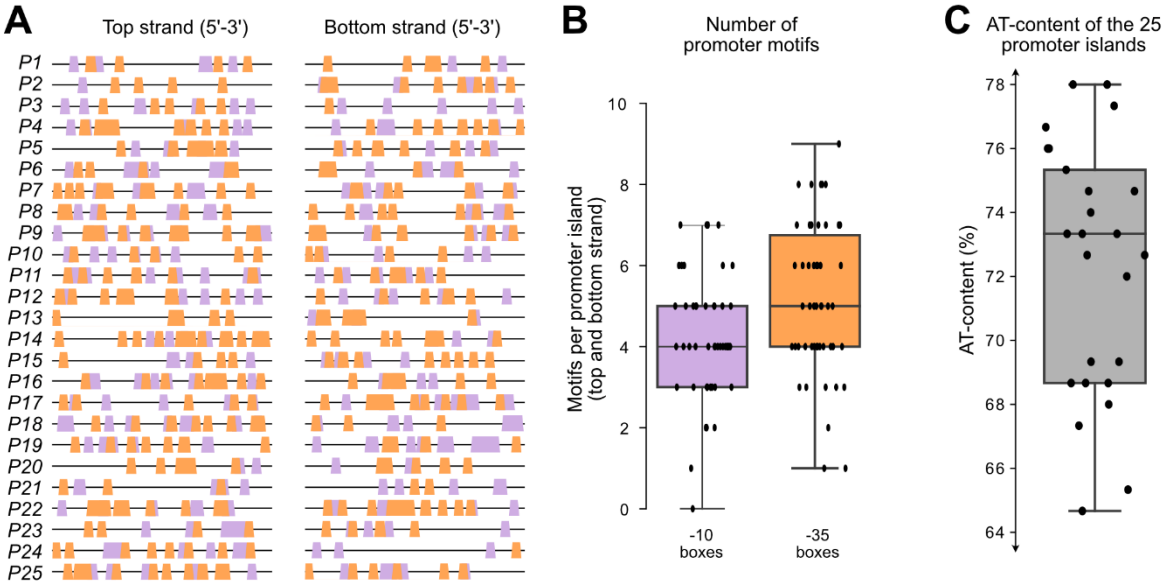


Figure S1. Promoter island sequences. (A) The locations and arrangements of -10 and -35 box motifs located across both the top and bottom strands of the 150 base pair DNA sequences called parent sequences (P1-P25). Orange trapezoids: -35 box motifs. Magenta trapezoids: 10 box motifs. Left: top strand. Right: bottom strand. Note: the plot is shown from 5'-3' for both strands. **(B)** Distributions for the number of motifs found in the 25 promoter island sequences. We identified motifs with position-weight matrices (PWMs, methods). Magenta: the number of -10 box motifs in each promoter island regardless of orientation (n=50). Orange: the number of -35 box motifs in each promoter island, in both orientations (n=50). **(C)** Distribution of AT-content for the 25 parent sequences. Boxes in B) and C) represent the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR

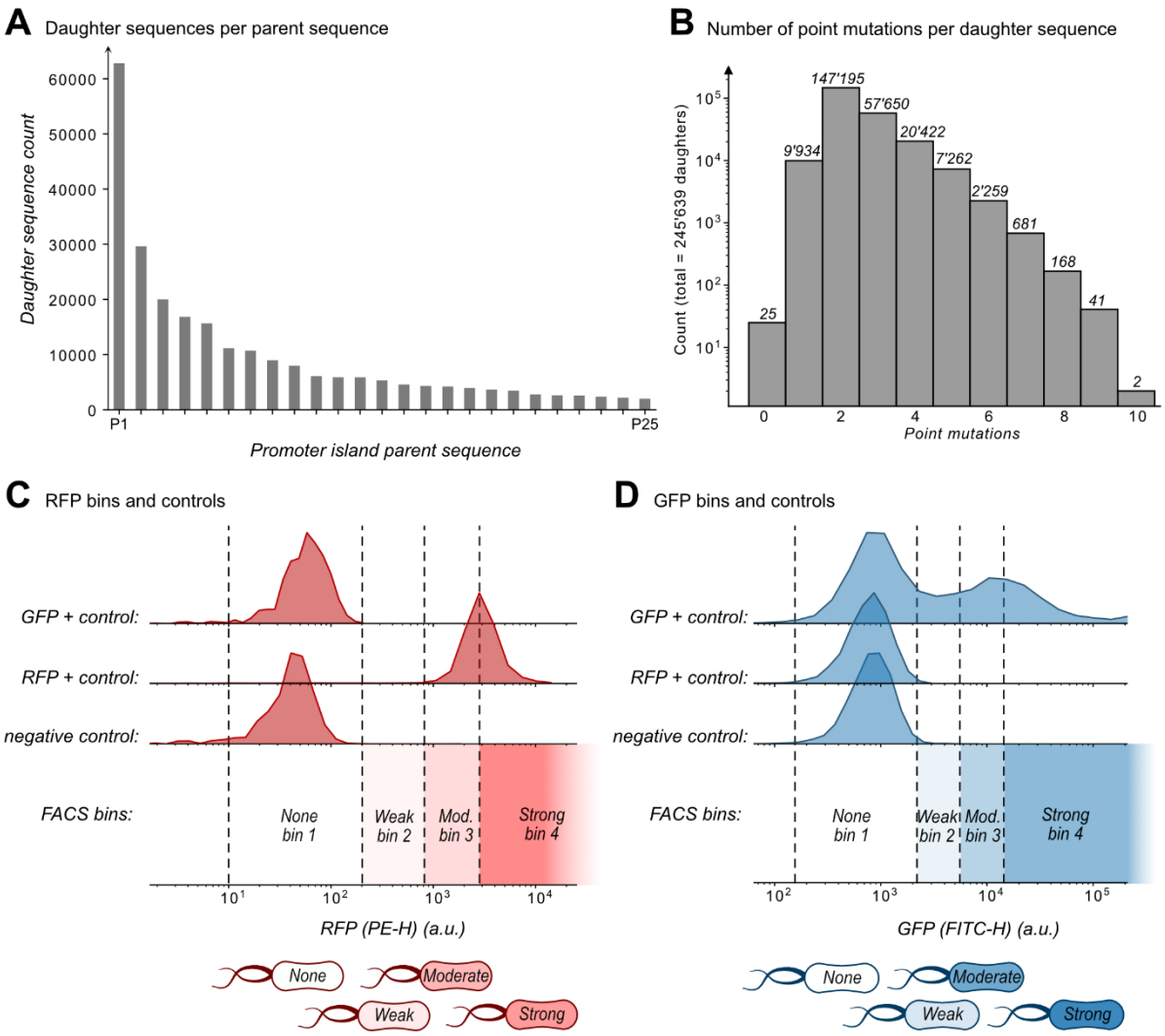
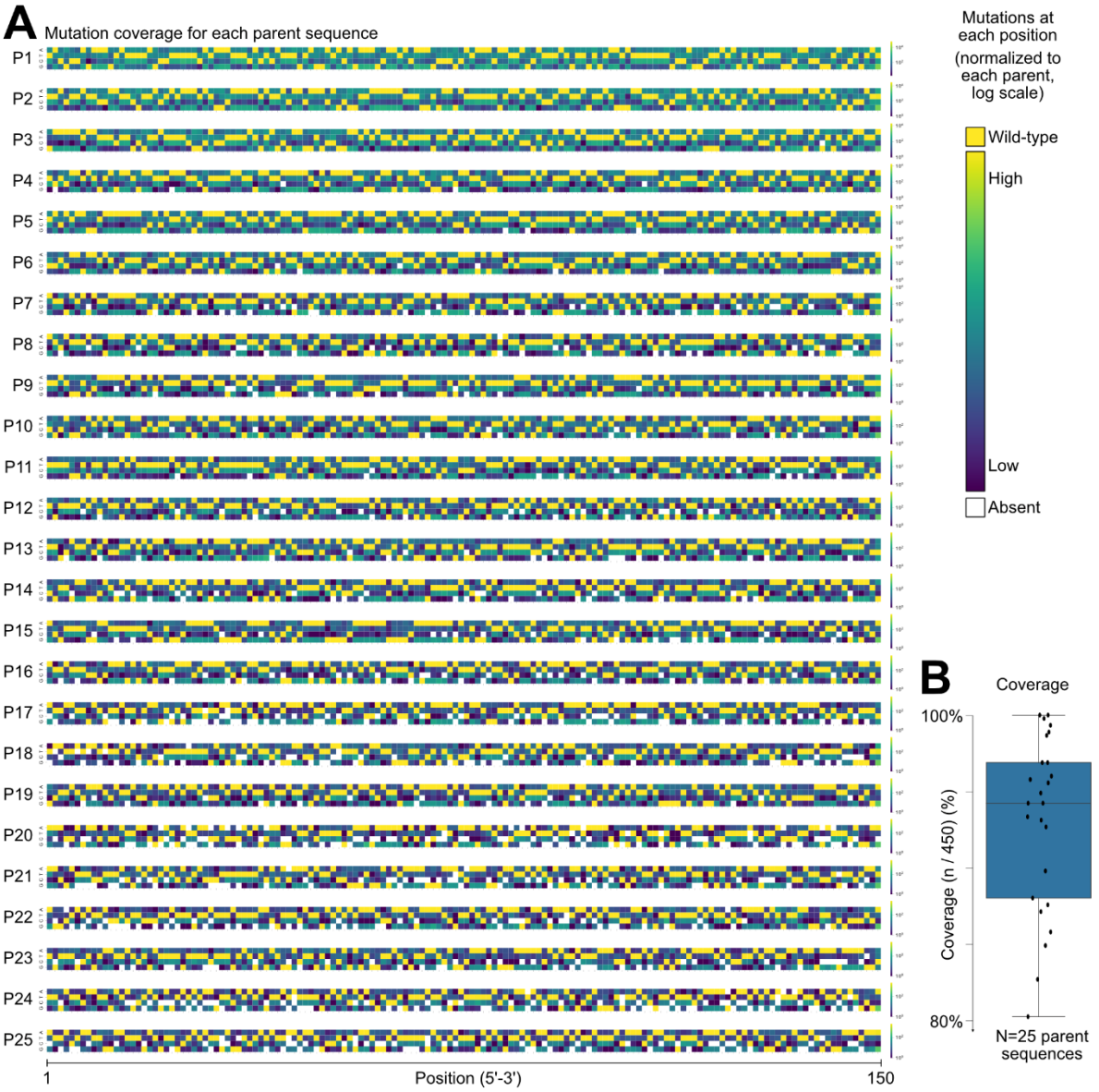


Figure S2. Mutagenesis library and sort-seq bins. (A) A histogram with the number of unique daughter sequences per parent sequence (N=245'639). **(B)** The number of point mutations per daughter sequence (N=245'639). The number above each bin is the number of daughters in each bin. **(C)** We defined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



818

819 **Figure S3. Mutational coverage for the parent sequences. (A)** Heatmaps depicting the frequency of each
820 nucleotide (A,T,C,G, y-axis) at each position (5'-3', x-axis) for all of the daughter sequences of each parent
821 sequence (P1-P25). Daughter sequences contain 1-10 point mutations (see **Fig S2B**). The color of each square in
822 the heatmap corresponds to the frequency of each nucleotide occurring in the library at its respective position,
823 where less frequent nucleotides are shown in dark magenta. Highly frequent nucleotides are in yellow (log-scale).
824 These include the nucleotides of the parent sequence, due to our modest mutation rate (~2.0 point mutations per
825 daughter sequence). Light gray squares indicate mutations absent from the library. **(B)** We calculate coverage as
826 the percentage of all possible mutations of the parent sequences (A,T,C,G and positions 1 through 150, 3×150 =
827 450 total neighboring mutations). In the boxplot, the box represents the interquartile range (IQR) and the center
828 line the median. Whiskers correspond to 1.5×IQR.

829

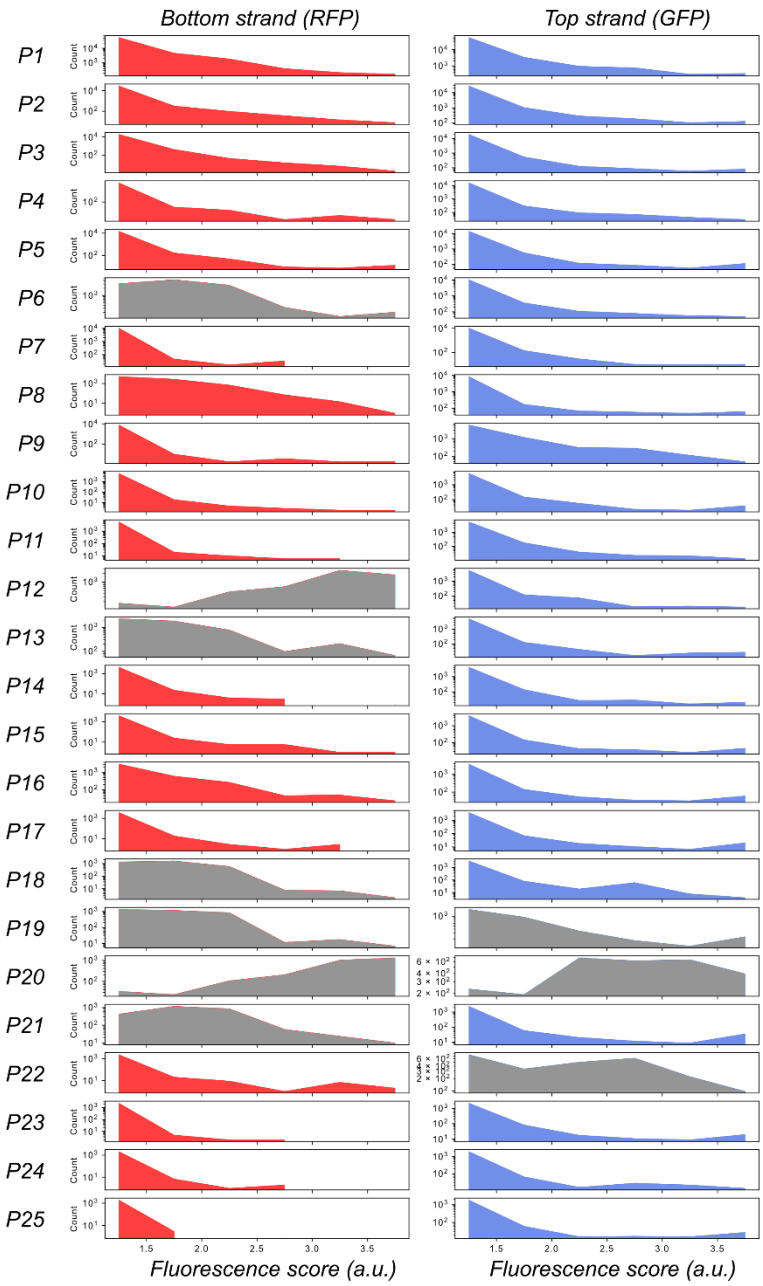


Figure S4. Histogram of the fluorescence score distributions for each promoter island and genetic orientation. Rows: parent sequences (P1-P25). Left column: bottom strand (RFP fluorescence scores); Right column: top strand (GFP fluorescence scores.) Each panel is a histogram of the frequency of fluorescence scores (arbitrary units, a.u.) from the sort-seq experiment. Gray histograms indicate those parent sequences / orientations that already encode promoter activity. Note: daughter sequence counts vary among parents, and y-axis is unique to each panel.

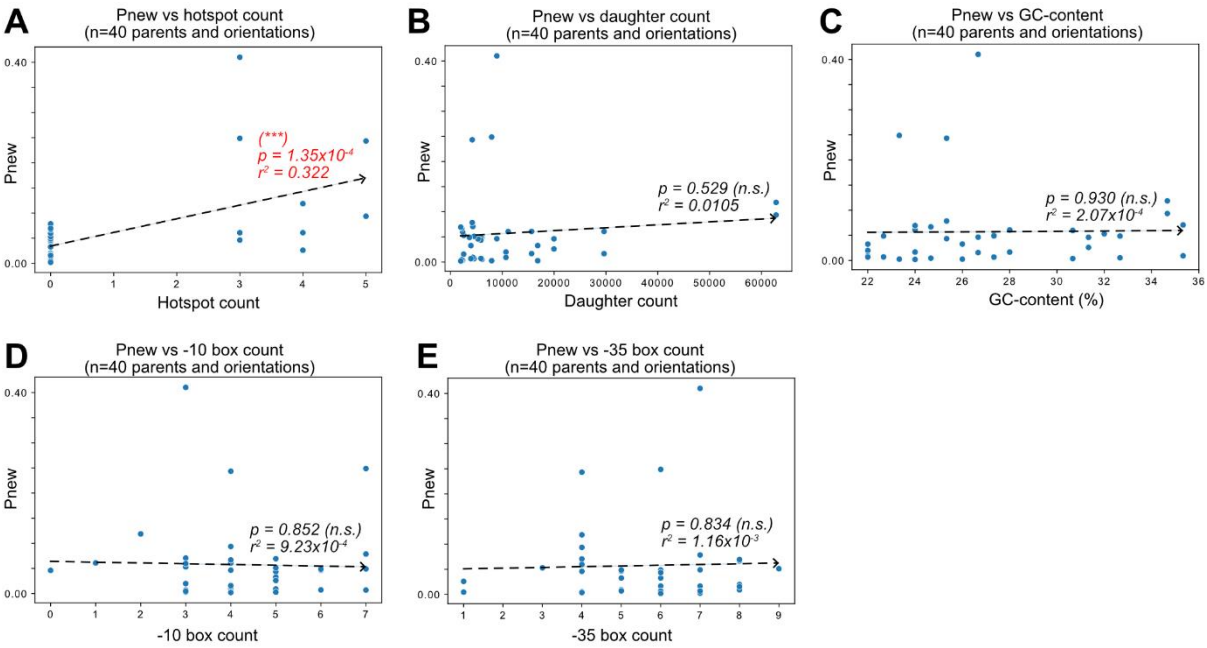


Figure S5. Number of hotspots correlates with P_{new}, but numbers of -10 and -35 box motifs do not. (A) Scatterplots comparing the number of mutual information hotspots (see **Fig 2**) and P_{new} (see **Fig 1**). P_{new} is the ratio of daughter sequences with a fluorescence score greater than 1.5 a.u. to the total number of daughter sequences for each parent. Each point represents a parent / orientation without regulatory activity (n=40). The dotted 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) ($p=1.35\times10^{-4}$, $r^2=0.322$). **(B)** Analogous to (A) but comparing the number of daughter sequences for each parent with P_{new} ($p=0.529$, $r^2=0.0105$, n.s.). **(C)** Analogous to (A) but comparing the GC-content of each parent sequence with P_{new} ($p=0.930$, $r^2=2.07\times10^{-4}$, not significant (n.s.)). **(D)** Analogous to (A) but comparing the number of -10 box motifs in each parent sequence and orientation with P_{new} ($p=0.852$, $r^2=9.23\times10^{-4}$, n.s.). **(E)** Analogous to (A) but comparing the number of -35 box motifs in each parent sequence and orientation with P_{new} ($p=0.834$, $r^2=1.16\times10^{-3}$, n.s.).

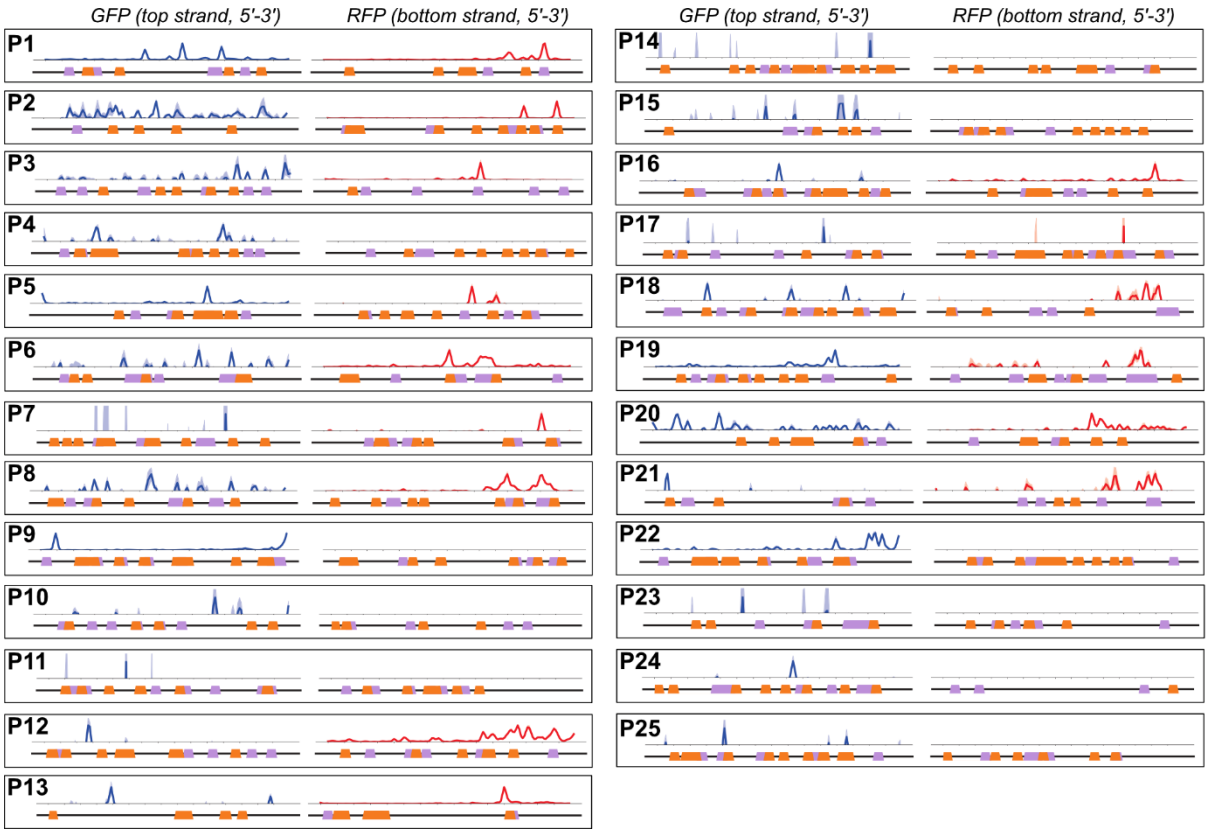


Figure S6. Mutual information and promoter motifs in the promoter islands. Each panel corresponds to a unique promoter island sequence, with its numerical identifier in bold. Within each panel there are two plots, corresponding to the top strand (left), and the bottom strand (right), both shown from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

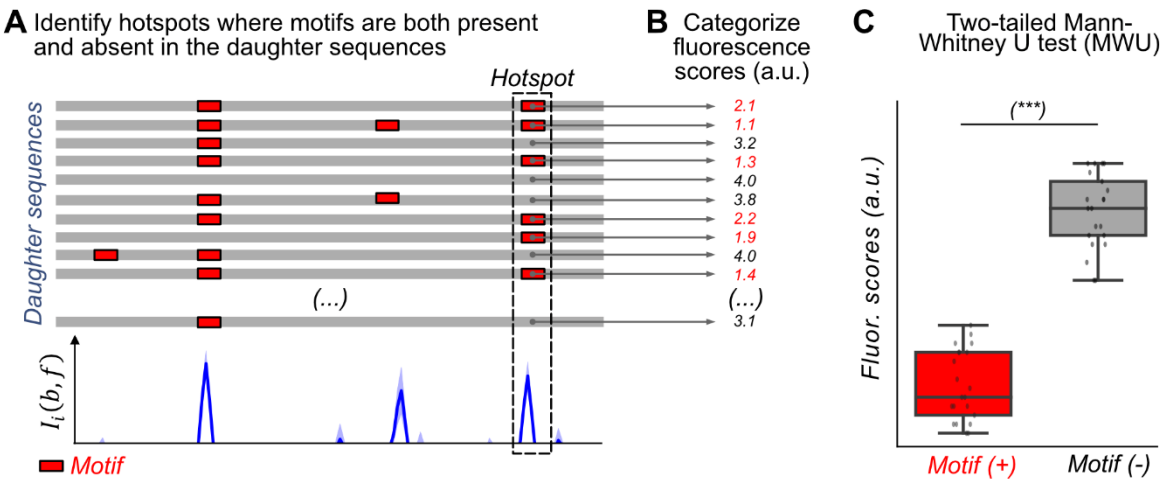
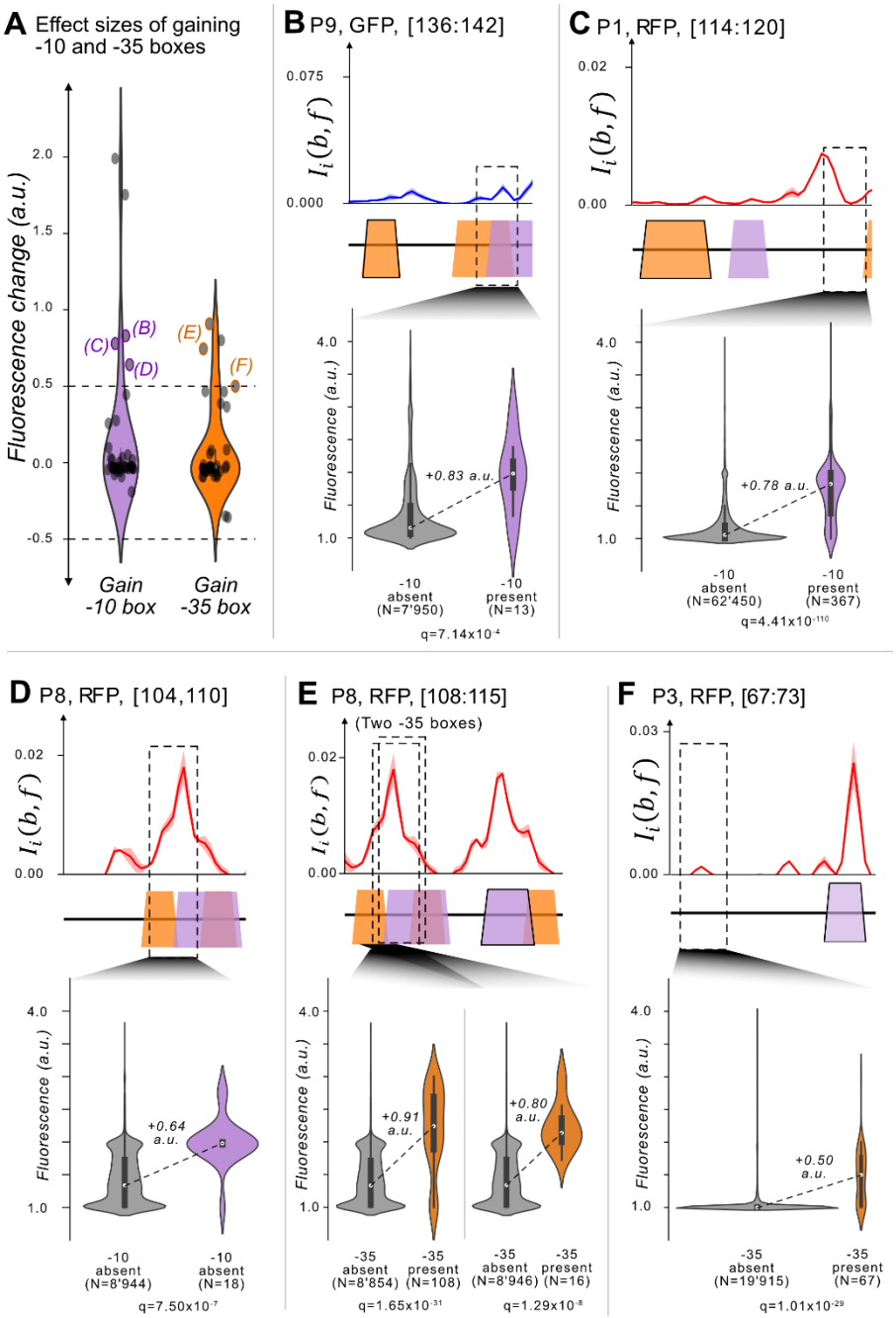


Figure S7. Identifying where motifs are gained and lost in hotspots that are associated with changes in fluorescence. These panels do not show data from any parent sequence or hotspot, but show hypothetical data to illustrate how we identified the associations of base identity and fluorescence in our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



871

872 **Figure S8. Additional examples of which gaining -10 and -35 boxes creates de-novo promoters.** (**A**) The
873 change in fluorescence (arbitrary units, a.u.) when gaining or losing -10 and -35 boxes in hotspots in the non-
874 regulatory parent sequences. Dotted lines indicate an effect size threshold of ± 0.5 a.u. Each point corresponds to
875 a gain or loss of a motif in a mutual information hotspot. Outlined points with letters in parenthesis highlight the
876 corresponding figure panels. See **Fig 3A** for the same plot. The areas of the violin plots are the kernel density
877 estimates (KDE) of each distribution. **(B)** Top: the top strand of parent P9. Dashed rectangle indicates the region
878 of interest analyzed in the subsequent figure panel. We plot the mutual information $I_i(b,f)$ between nucleotide
879 identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 standard
880 deviation when the dataset is randomly split into three equally sized subsets (methods). Middle: position-weight
881 matrix (PWM) predictions for the -10 box motifs (magenta trapezoids) and -35 box motifs (orange trapezoids) along
882 the wild-type parent sequence. Bottom: fluorescence scores (a.u.) for daughter sequences with vs without a -10
883 box in the region of interest at position 136:142. We tested the null hypothesis that the fluorescence scores have
884 the same central tendency using a two-sided Mann-Whitney U (MWU) test ($q=7.14\times10^{-4}$). The areas of the violin
885 plots are the kernel density estimates (KDE) of each distribution. Within each violin plot is a box plot. The box
886 represent the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR. **(C)**

Analogous to (B) except for gaining a -10 box on the bottom strand of parent P1 at position 114:120 (two-tailed MWU test, $q=4.41\times10^{-110}$). **(D)** Analogous to (B) except for gaining a -10 box on the bottom strand of parent P8 at position 104:110 (two-tailed MWU test, $q=7.50\times10^{-7}$). **(E)** Analogous to (B) except for gaining two -35 boxes on the bottom strand of parent P8 at position 108:115 (two-tailed MWU tests, $q=1.65\times10^{-31}$ and $q=1.29\times10^{-8}$, left and right -35 box, respectively). **(F)** Analogous to (B) except for gaining a -35 box on the top strand of parent P3 at position 16:22 (two-tailed MWU test, $q=9.62\times10^{-15}$).

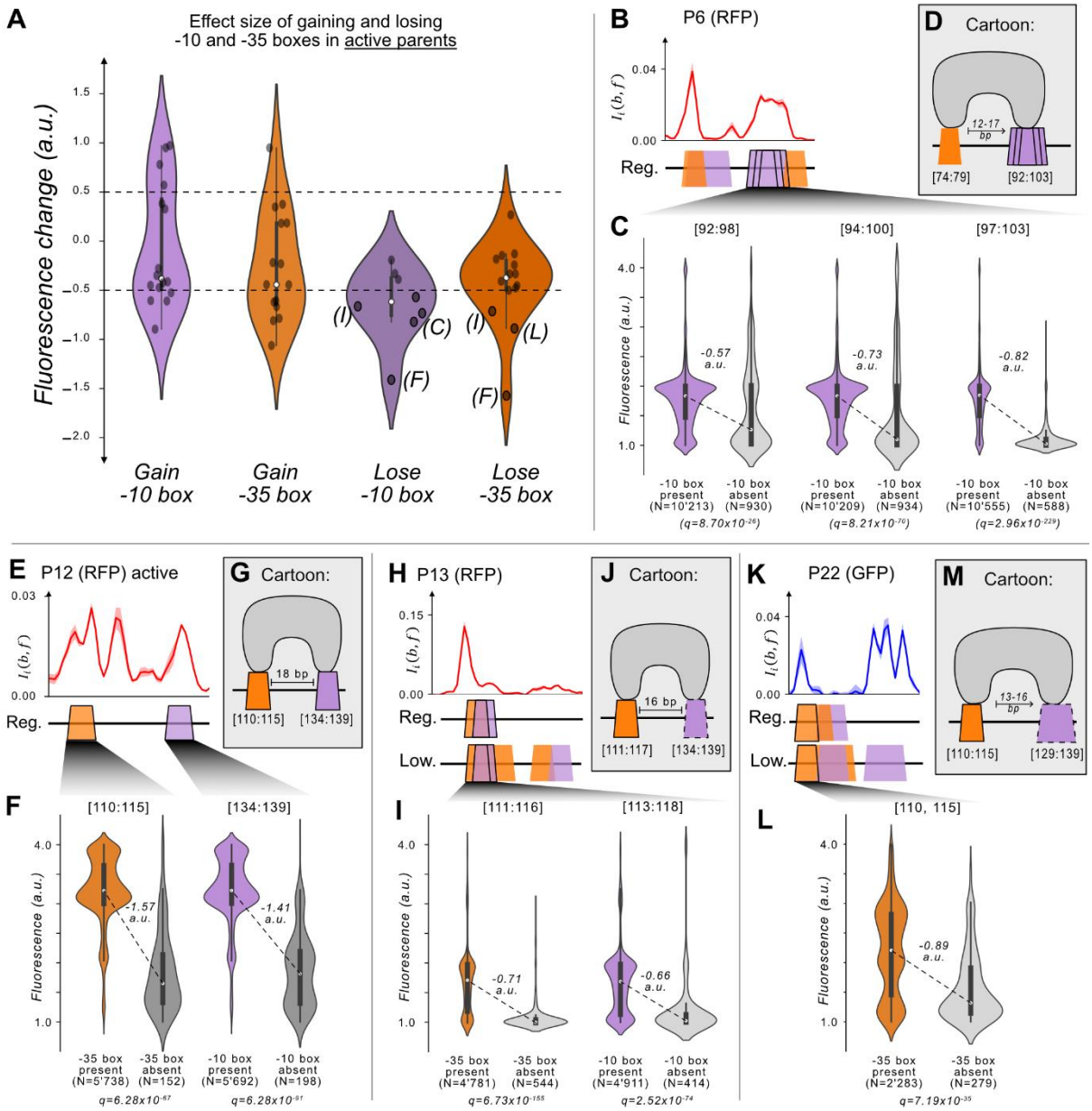


Figure S9. Mapping promoters in active parents. (**A**) The change in fluorescence (arbitrary units, a.u.) when gaining or losing -10 and -35 boxes in hotspots in the active parent sequences. Dotted lines indicate an effect size threshold of ± 0.5 a.u. Each point corresponds to a gain or loss of a -10 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. **(B)** Top: the bottom strand of parent P6. 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 box motifs (magenta trapezoids) and -35 box motifs (orange trapezoids) along the wild-type parent sequence. Outlined trapezoids with shadows indicate regions of interest for the subsequent figure panel. **(C)** We test the null hypothesis that the fluorescence scores have the same central tendency for daughter sequences with vs without -10 boxes using a two-sided Mann-Whitney U (MWU) test. We tested the gain of -10 boxes in three locations. Left: position 92:98 ($q=8.70\times10^{-25}$), middle: position 94:100 ($q=8.21\times10^{-70}$), right: position 97:103 ($q=2.96\times10^{-229}$). **(D)** We combine the information from the previous panels (B-C) to generate a cartoon model for P6. Solid trapezoids correspond to validated binding sites in the previous panel. Trapezoids without borders correspond to PWM predicted motifs. Dashed trapezoids refer to inferred low-affinity motifs (see methods). Orange trapezoids: -35 boxes, magenta trapezoids = -10 boxes. **(E)** Analogous to (B) but for the bottom strand of P12. **(F)** Analogous to (C) but for losing a -35 box and a -10 box (two-tailed MWU test, $q=6.29\times10^{-67}$ and $q=5.28\times10^{-91}$, respectively). **(G)**

Analogous to (D) but a cartoon model for P12 based on panels (E-F). **(H)** Analogous to (B) but for the bottom strand of P13. **(I)** Analogous to (C) but for losing a -35 box and a -10 box (two-tailed MWU test, $q=6.73\times10^{-155}$ and $q=2.52\times10^{-74}$, respectively). **(J)** Analogous to (D) but a cartoon model for P13 based on panels (H-I). **(K)** Analogous to (B) but for the top strand of P22. **(L)** Analogous to (C) but for losing a -35 box (two-tailed MWU test, $q=7.19\times10^{-35}$). **(M)** Analogous to (D) but a cartoon model for P22 based on panels (K-L).

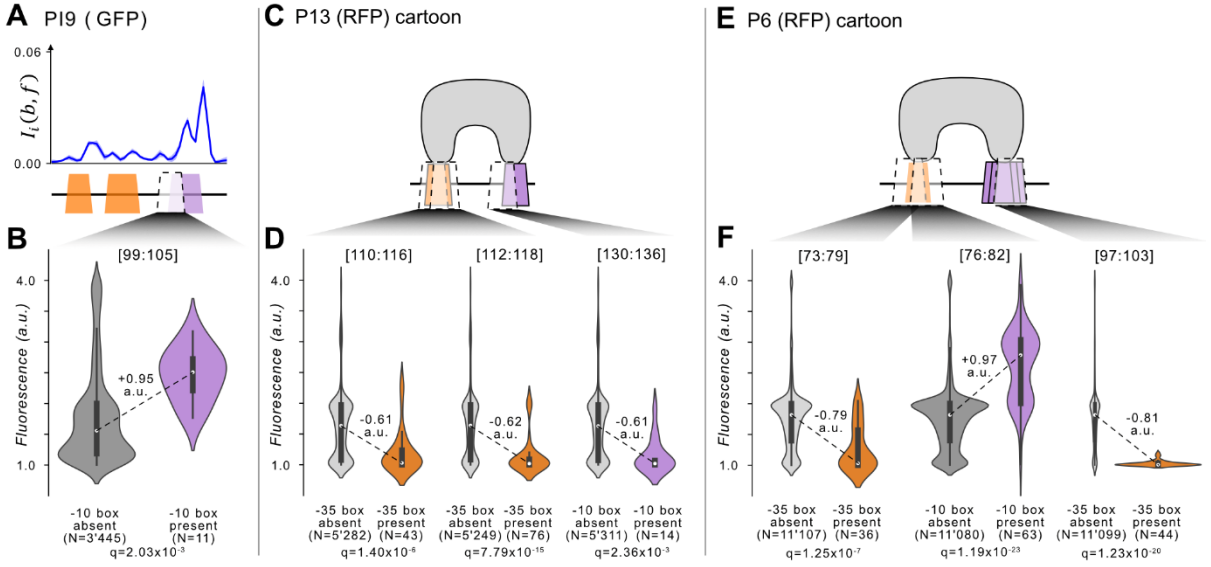


Figure S10. Additional examples of mutations modulating promoter activity. (**A**) Top: the top strand of parent P19. We plot the mutual information $I_i(b,f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (methods). Middle: position-weight matrix (PWM) predictions for the -10 box motifs (magenta trapezoids) and -35 box motifs (orange trapezoids) along the wild-type parent sequence. Dashed trapezoids with shadows indicate regions of interest for the subsequent figure panel. **(B)** We test the null hypothesis that the fluorescence scores have the same central tendency for daughter sequences with vs without -10 boxes using a two-sided Mann-Whitney U (MWU) test ($q=2.03\times10^{-3}$). 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. **(C)** A cartoon for the promoter mapped on the bottom strand of P13 based on data from **Fig S9**. Orange trapezoids = -35 boxes, magenta trapezoids = -10 boxes, dashed trapezoids = regions of interest in subsequent figure panels. **(D)** Analogous to (B) but for gaining (left) a -35 box, (middle) a -35 box, (right) a -10 box (two-sided MWU test, $q=1.40\times10^{-6}$, $q=7.79\times10^{-15}$, $q=2.36\times10^{-3}$, respectively). **(E)** Analogous to (C) but for the bottom strand of P6. **(F)** Analogous to (B) but for gaining (left) a -35 box, (middle) a -10 box, (right) a -35 box (two-sided MWU test, $q=1.25\times10^{-7}$, $q=1.19\times10^{-23}$, $q=1.23\times10^{-20}$, respectively).

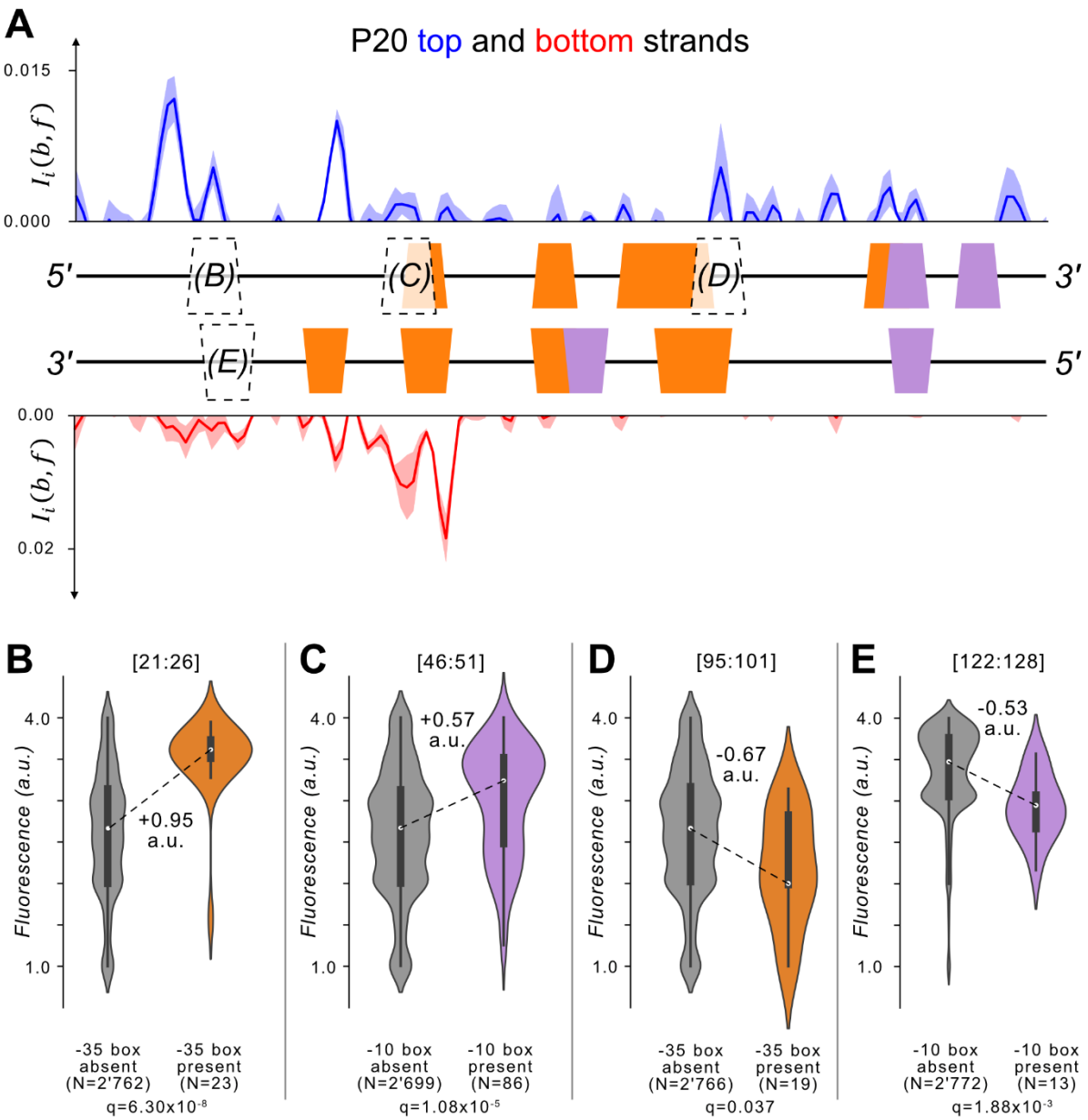


Figure S11. Parent P20. **(A)** We plot the mutual information for the top (blue) and bottom (red) strands of parent P20. Top: the top strand of parent P20. We plot the mutual information $I_i(b,f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded 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 box motifs (magenta trapezoids) and -35 box motifs (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 the bottom strand of parent P20 as described for the top strand. **(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 21:26 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. **(C)** Analogous to (B) but for gaining a -10 box at position 46:51 (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, $q=1.88\times10^{-3}$).

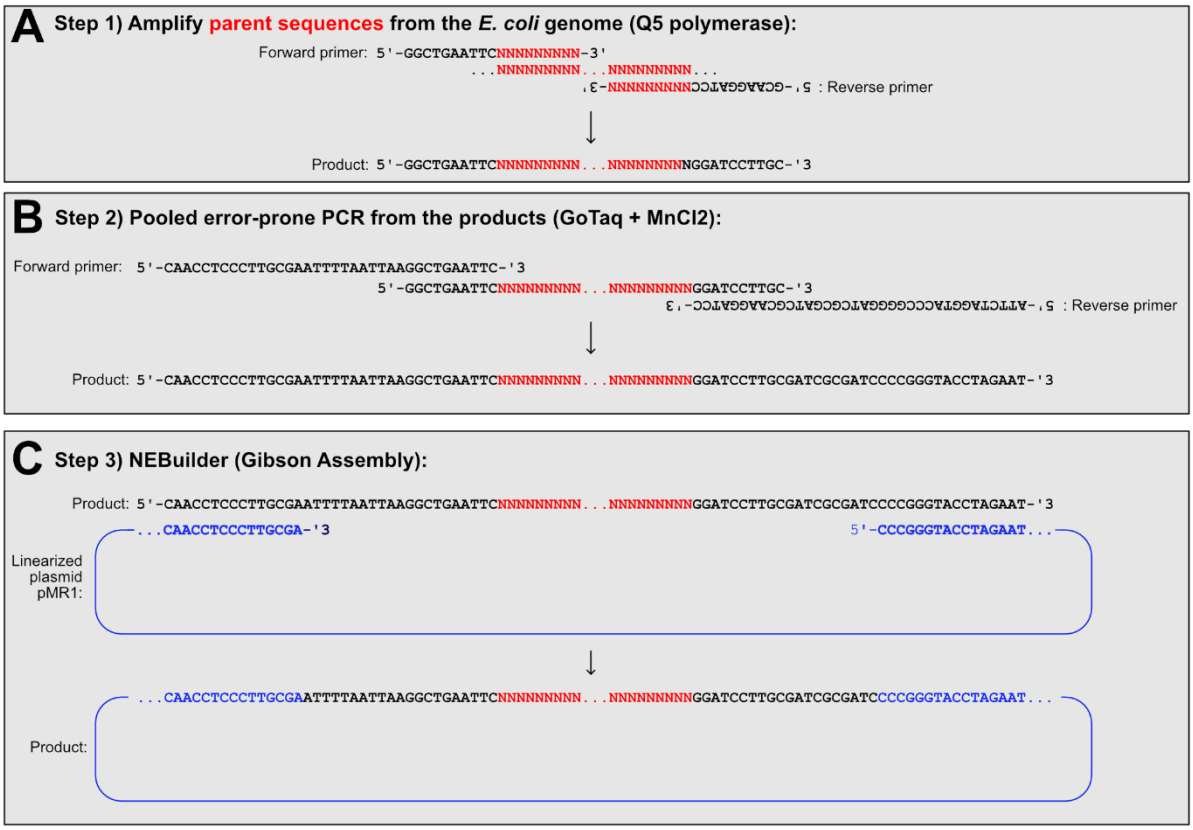


Fig S12. Molecular cloning parent sequences. (A) We first amplify the parent sequences from the *E. coli* genome using Q5 polymerase. Top: the forward and reverse primers contain constant overhang regions, and unique sequences (red N's) unique to the parent sequence. Bottom: the product is a unique parent 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.