Mobile DNA is replete with hotspots for the de novo emergence of gene regulation

Timothy Fuqua[12], Andreas Wagner[123*]

ABSTRACT

DNA mutations that create new gene expression are important raw materials for Darwinian evolution. One potential source of new gene regulation is mobile DNA, which can sometimes drive the expression of genes near its insertion site in a genome through outward-directed promoters. However, we do not know how frequent this ability is, nor how frequently mobile DNA may evolve such promoters *de novo*. Here we address these questions for the insertion sequence family IS3, the most abundant family of a simple form of prokaryotic mobile DNA. First, we estimate that at least 30% of IS3 sequences harbor outward-directed promoters. Second, we combine high-throughput mutagenesis with a massively parallel reporter assay to show that single point mutations suffice to create outward-directed promoters in all the IS3 sequences we studied. We found that in 5.6% of 18'607 mutant IS3 sequences, promoter activity emerged de novo. Promoters preferentially arise at emergence hotspots in each IS3 sequence. These hotspots overlap with promoter motifs that already exist or are newly created by mutation. One common route to promoter activity is gaining a -10 box downstream of an existing -35 box, which we call "Shiko Emergence." Overall, we show that mobile DNA has a high latent potential to drive new gene expression. This makes mobile DNA ideal for domestication by its host organism. It also raises intriguing questions about how this potential has evolved.

INTRODUCTION

Mutations that affect gene regulation have led to many evolutionary adaptations and innovations[1–3], from antibiotic resistant pathogens[4] to the limbless body plan of snakes[5]. Many such mutations affect cis-regulatory elements (CREs), non-coding DNA sequences that bind proteins necessary to express genes[2]. Mutations that alter existing CREs can lead to new gene expression via single nucleotide changes[1], DNA duplications[6,7], or DNA expansions within a CRE[8].

New CREs can also emerge *de novo* through mutations that co-opt DNA with different, non-regulatory functions, or that create CREs from sequences with no prior functions[3,9–11]. De novo CREs have emerged from genomic repeat regions[12–14], sequences surrounding ancient genes[8], and even from random DNA[10,13,15]. In one key study, de novo CREs emerged repeatedly under directed evolution in specific locations within three randomly generated DNA sequences[14]. These observations and theoretical work[15] suggest the existence of DNA hotspots that readily evolve CRE activity. However, beyond anecdotal evidence, we do not know how frequent such hotspots are, and what constitutes a hotspot for de novo CRE emergence.

[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

De novo CREs can also emerge from mobile DNA, by fortuitously encoding transcription factor binding sites[16,17] or *outward-directed promoters*[18–21] – CREs that transcribe genes adjacent to a mobile DNA's genomic integration site[22–24]. In eukaryotes, many young CREs originate from remnants of mobile DNA called short-interspersed nuclear elements (SINEs) and long terminal repeats (LTRs)[16]. A subset of CREs called enhancers have also been demonstrated to originate from SINEs[18] in humans, cows, and dogs. This mode of origin is not rare: Some 44% of primate-specific enhancers may have been co-opted from mobile DNA[25].

In prokaryotes, the simplest and most abundant kind of mobile DNA is the insertion sequence (IS)[26]. It consists of a gene encoding transposase – responsible for its mobility – that is flanked by direct and terminal inverted repeats[22–24,27]. ISs fall into multiple families. One especially abundant family is the IS3 family, with hundreds of known members[28]. In multiple directed evolution experiments, an IS3 member with an outward-directed promoter repeatedly integrated at key genomic locations to selectively increase the expression of various genes[29–31].

The cis-regulatory potential of the IS3 family is unknown, because it remains unclear whether IS3s commonly possess outward-directed promoters, and how easily mutations in ISs can create de novo CRE activity. Here, using a combination of computational analysis and experimental validation on selected IS3s, we first estimated that at least 30% of known IS3 family members encode outward-directed promoters. We then introduced more than 18'000 mutations into parts of five IS3s that do not already drive outward-directed gene expression and assessed the ability of these mutations to create CRE activity. We found that for each of the five IS3s, de novo CRE activity can emerge from single point mutations. We additionally identify hotspots within these IS3s where CRE activity is most likely to emerge, noting that most hotspots overlap with promoter motifs or regions where mutations form a new promoter motif. Overall, our results suggest that mobile DNA has a high latent cis-regulatory potential, begging the question how this potential evolved.

RESULTS

IS3s are enriched with outward-directed promoter signatures.

Four IS3 family members have been serendipitously observed to have outward-directed promoters[24]. However, it is unclear how frequent outward-directed promoters are among the more than 700 characterized IS3 sequences[28]. To find out, we first used established computational tools called position-weight matrices (PWMs) to identify protein binding sites in DNA (**Fig 1a**). Briefly, PWMs are statistical representations of all possible DNA sequences ("sites") a protein can bind to, based on experimentally validated binding sequences. A PWM encodes the frequencies of allowed nucleotides at each position of a binding site as a logarithmically transformed score in information-theoretical units (bits)[32]. Given a query DNA sequence, we can use PWMs to assess whether the query exceeds a threshold of similarity to known protein binding sites, in which case we classify the query as a binding motif.

A canonical prokaryotic promoter contains two AT-rich sites called the -35 and -10 boxes spaced 14-17 base pairs (bp) apart[33]. These boxes bind a subunit of RNA polymerase called the σ_{70} factor[15]. We used PWMs for the -10 and -35 boxes to search for promoters in 706 IS3 sequences (**Fig 1b**) and refer to any

positive match as a promoter signature (see methods). We found 2'209 such promoter signatures, 1'428 of which (~65%) occurred on the bottom strand of DNA.

To locate these promoter signatures in the 706 IS3s, we partitioned the total length of each IS3 sequence into 20 equidistant bins (median length 63.5 bp), and counted the promoter signatures in each bin across all 706 IS3s (**Fig 1c**). We found that on both strands, promoter signatures are significantly non-uniformly distributed across the IS3s (K.S. test, top strand $p=9.61\times10^{-5}$, bottom strand p=0.0112, Methods), because the distal-most 10 percent of IS3s – the region where functional outward-directed promoters can occur (Fig S2) – contain more promoter signatures than the rest (Fig 1c).

We then asked whether this distribution of promoter signatures could occur by chance alone, given the sequence composition of our IS elements. To answer this question, we scrambled each of the 706 IS sequences within each bin, searched for promoter signatures, and plotted their locations (**Figure S1a,b**) (Methods). We found no significant difference between the mean number of signatures per bin in the wild-type vs. the scrambled sequences (**Figure S1c**) (two-tailed t-tests, top: p=0.752, bottom: p=0.161). Overall, this analysis demonstrates that the non-uniform distribution of promoter signatures results from the DNA sequence composition of ISs.

At least 30% of IS3s encode outward-directed promoter activity.

To validate our computational predictions experimentally, we randomly selected five IS3s with predicted outward-directed promoter signatures, and five predicted not to have outward-directed promoter signatures. We then synthesized 120 bp from the termini of these sequences that include the promoter signature (if present), cloned them into a fluorescence reporter plasmid (pMR1), transformed *E. coli* with the resulting plasmid, and measured fluorescence with a flow cytometer (BD Biosciences, FACSAria III). We compared the fluorescence levels to both a control that lacked an insert in the plasmid entirely, and another that encoded a promoter oriented in the opposite direction (**Fig S2**) (see methods).

We found that all tested IS3s (5/5) predicted to have outward-directed promoters indeed drove reporter gene expression (**Fig 1d**). Two of the five IS3s not predicted to have outward-directed promoters also exhibited higher fluorescence than both controls (**Fig 1e**). Taken together, these observations suggest that the promoter signatures we identified within 150 bp from the ends of IS3s indeed constitute outward-directed promoters. Based on our data, we estimate that ~19% (413) of the identified 2'209 signatures are outward-directed promoters. They occur on 215 IS3 sequences. Given our low false positive rate, we thus estimate that at least ~30% (215/706) of all IS3s may encode outward-directed promoters. This number may be an underestimate, because our experiments also reveal a high false negative rate for promoter prediction (Fig 1e). That is, computational analysis fails to identify promoter signatures for some IS3 sequences that actually drive gene expression (see Fig 1e).

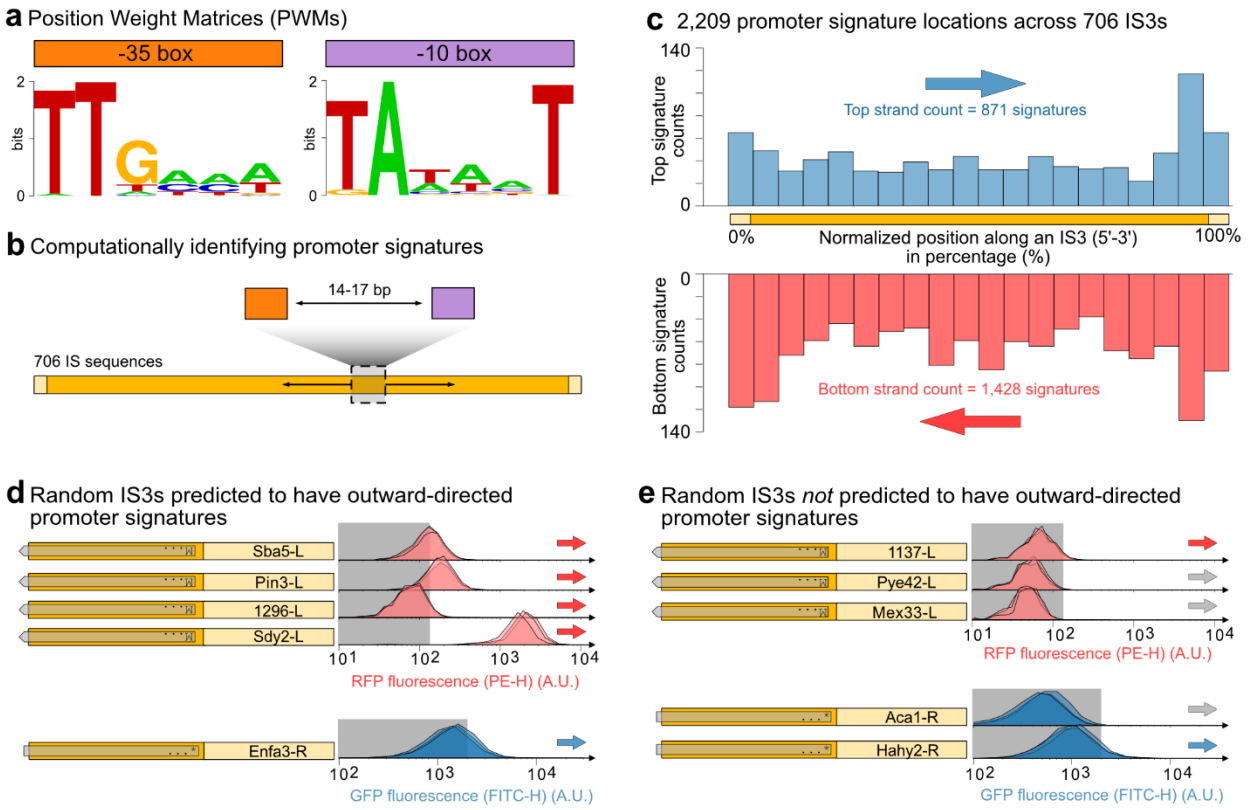

Figure 1 IS3s preferentially encode promoter signatures close to their ends. (a) Sequence logos derived from position weight matrices (PWMs) depicting the likelihood of a base being present at each position of a protein, in this case, the σ_{70} factor. The taller the letter at each position is, the more likely it is that the corresponding nucleotide is present at the position in the binding site. Left: PWM logo for the -35 box. Right: PWM logo for the -10 box. **(b)** To computationally identify promoter signatures in the ISs, we searched for -35 and -10 boxes using PWMs spaced 14-17 base pairs (bp) apart in both the top and the bottom strands of 706 IS3s. **(c)** We plotted the identified promoter signatures as histograms with a fixed bin width of 5% of IS3 length (20 bins in total). The top and bottom histograms correspond to promoter signatures on the top and bottom strand of the IS3s, respectively. **(d-e)** Flow cytometry plots depicting the distribution of fluorescence as arbitrary units (a.u.) for genetic constructs cloned into a reporter plasmid (pMR1). Depending on the orientation, the fluorescence readout is either for RFP (red histograms, bottom strand) or GFP (blue histograms, top strand). Shaded regions correspond to readouts produced by negative controls (see methods). Blue and red arrows indicate high fluorescence, and gray arrows indicate fluorescence indistinguishable from controls. **(d)** Fluorescence readouts for five randomly selected flanking IS3 sequences predicted to have outward-directed promoter activity, and **(e)** for five randomly selected predicted not to have outward promoter signatures.

Promoters readily emerge from non-regulatory mobile DNA.

Because all mobile DNA is continuously exposed to mutation pressure during its evolution in a host genome, we wished to find out if and how mutations can create promoters de novo in IS3s without outward-directed promoter activity. To choose an appropriate size of IS3 DNA fragments for mutagenesis, we first performed an experiment that inserted a functional promoter at varying positions across an IS3 backbone, and asked whether this promoter could drive detectable gene expression of an adjacent reporter gene. This experiment shows that the promoter needs to occur within 150 bp from the reporter gene to drive detectable expression (**Fig S2**). This observation motivates our choice to study de-novo

promoters that originate from the ends of IS3 sequences, because only from there can they drive ectopic gene expression.

We amplified five 150 bp sequences from the ends of three *E.coli* IS3s without detectable promoter activity, which we call 1L, 1R, 2R, 3L, and 3R (see methods), and refer to them as parent sequences (**Fig S4a**). We pooled these parent sequences, and created from them a deep mutational scanning library of daughter sequences via error-prone polymerase chain reaction (PCR) (**Fig 2a**). We cloned this library into the pMR1 plasmid[34], and transformed it into *E.coli* cells (**Fig 2b**).

The pMR1 plasmid encodes a gene encoding green fluorescent protein (GFP) on the top strand, downstream of each mutagenized IS fragment, and a gene encoding a red fluorescent protein (RFP, mCherry) on the bottom strand, upstream of each mutagenized IS fragment. The pMR1 plasmid thus allowed us to simultaneously measure promoter activity resulting from mutations on both strands of the IS3s.

To measure reporter expression driven by each daughter sequence, we used Sort-Seq[35–40] (**Fig 2c**). Specifically, we separated bacterial cells with a cell sorter (BD Biosciences, FACSAria III) into four fluorescence bins according to whether their daughter sequence drives no, low, medium, or high fluorescence. We sequenced the plasmid inserts from the subpopulation of each bin (**Fig S4**). We performed Sort-Seq in three technical replicates (r1, r2, r3), calculating a fluorescence score for each sequence and triplicate that ranged between one (no expression) and four (highest expression) (**Fig S4**). We report the fluorescence score of each sequence as an average over these triplicates. This average score strongly correlated with the score from each technical replicate (Pearson R, GFP: $R_{mean:r1}$ = 0.93, $R_{mean:r2}$ = 0.95, $R_{mrean:r3}$ = 0.94. RFP: $R_{mean:r1}$ = 0.94, $R_{mean:r2}$ = 0.95, $R_{mean:r3}$ = 0.95) (**Fig S4**) (see methods).

This massively parallel reporter assay identified 18’607 unique daughter sequences, with a mean of 3’721 daughter sequences per parent sequence (1L = 3’162, 1R = 6’354, 2R = 4’130, 3L = 3’027, 3R = 1’934) (**Fig S4**). The daughter sequences harbored a median of 2.0 point-mutations (standard deviation 1.18) relative to their respective parent. Despite this small number of mutations, every parent gave rise to multiple daughter sequences with active de novo promoters. More specifically, we discovered 1’047 daughter sequences (5.6%) that showed promoter activity. Of these, 584 (~3.1%) expressed GFP (top strand), and 472 (~2.5%) expressed RFP (bottom strand). The strength of de novo promoters follows a broad tailed distribution, with many more strong promoters emerging than would be expected from a normal distribution (**Fig S4**).

Parent sequences substantially differ in their potential to evolve new promoters.

We then asked whether promoters emerged with equal likelihoods among the different parent sequences. To answer this question, we computed for each parent sequence the probability P_{new} that mutations create a new promoter, and did so separately for promoters emerging on the top and the bottom strands (**Fig 2d,e**).

We found that P_{new} varied 11.5-fold among parent sequences, from P_{new}=0.02 to P_{new}=0.23. In addition, promoters are 1.2-fold (584/472) more likely to emerge on the top (GFP) strand than on the bottom (RFP) strand. For the parent sequence with the lowest P_{new} (sequence 2R), 11% and 2% of daughter sequences

drove new expression on the top and bottom strands respectively. Conversely, for the parent sequence with the highest P_{new} (3R), 23% of daughter sequences drove new gene expression on the top strand and 3% on the bottom strand. Thus, relative to the other parent sequences, some parent sequences are biased towards evolving promoter activity, while others are biased against evolving promoter activity (**Fig 2d,e**).

The frequency of de novo promoters and their strengths increases with the number of mutations.
We next asked how the strength of a de novo promoter is related to the number of mutations in the daughter sequence harboring it. To this end, we grouped the daughter sequences into categories encoding 1, 2, 3, and 4 or more (4+) mutations, and calculated the frequencies of the different promoter strengths from each category (none, weak, medium, and high) (**Fig 2f**). The frequency of daughter sequences encoding promoters increased with the number of mutations (1 mutation: 7.8%, 2: 8.2%, 3: 9.9%, 4+: 14%). The vast majority (~96%) of daughter sequences with only a single mutation and a new promoter drove weak expression, and fewer than 1% drove high expression (none: 2'855 daughter sequences, weak: 233, moderate: 8, high: 2).

Conversely, for emergent promoters with 4+ mutations, 69% of the promoters produced weak expression and 11% high expression (none: 1'080, weak: 117, moderate: 35, high: 18). Thus, both the frequency of de novo promoters and the frequency of strong de novo promoters increases with the number of mutations.

Because single mutations are more frequent than double, triple etc. mutations, they arguably provide the easiest evolutionary route to new promoters. To better understand how single mutations creating new promoters are distributed within daughter sequences, we generated genotype-phenotype maps for all 1'549 daughter sequences that differed only by a single mutation from their respective parent (**Fig 1g,h**) (3L = 300, 3R = 303, 1L = 323, 1R = 278, 2R = 345). This figure once again illustrates that the vast majority (~96%) of promoter-creating mutations produce weak promoters.

Because weak emergent promoters are so widespread in the single mutants, we first focused our analysis on them. To start, we found that with one exception, these mutations are uniformly distributed throughout each sequence (**Fig S4a**). The exception is the parent sequence 3R(+) (K.S. test, p-value = 0.007), in which a cluster of weak-promoter-originating mutations exists in the final 50 bp of the parental 3R(+) sequence.

The parent sequences have an average AT-content of ~54.9% (3L: 54.0%, 3R: 50.7%, 1L: 56.7%, 1R:61.3%, 2R: 50.0%), which is higher than that of the *E. coli* genome (~50.8%)[41]. Because the -35 and -10 box consensus sequences are also AT-rich (`TTGAAA` and `TATAAT` respectively), we hypothesized that weak promoters often emerge by increasing the binding scores for either -35 or -10 boxes. To test this hypothesis, we calculated the changes in PWM scores of both -10 and -35 boxes before and after each mutation (**Fig S4b,c**). We found that -10 box scores are ~1.5 times (14.2% vs 9.3%) more likely to increase when a single mutation creates a weak promoter than when it creates no promoter (chi-squared test, 4 d.f., p = 0.014). We refer to the latter class of mutations as *promoter-neutral*. In contrast -35 box scores are not more likely to increase when a mutation creates a new promoter than when it does not (weak promoters: 12.8% vs promoter-neutral: 14.1%). Additionally, for promoter-creating mutations, -10 and -35 box scores did not change 72.5% and 74.7% of the time, respectively. For promoter-neutral mutations,

-10 and -35 box scores did not change 78% and 66.2% of the time. Overall, these observations suggest that single mutations that create weak promoters do sometimes but not always exert their effect through canonical promoter motifs.

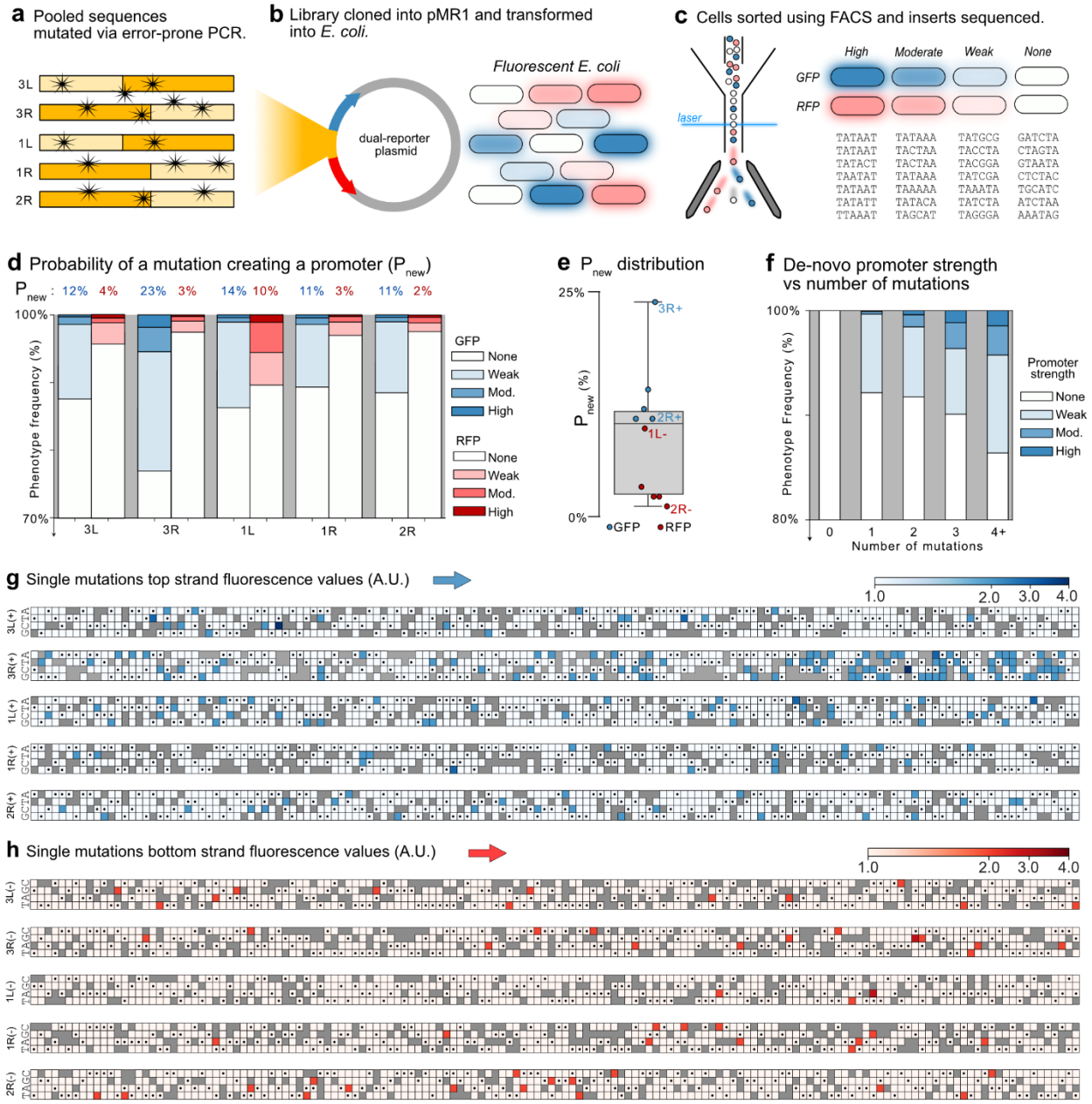

Figure 2 Promoters emerge de novo with different probabilities for different parent sequences. (a) We pooled parental IS fragments for error-prone PCR, **(b)** cloned the mutagenized library into the dual reporter plasmid pMR1, and transformed the resulting plasmid library into *E. coli*. Inserts with promoter activity fluoresce green or red (shown as blue or red here), depending on the orientation of the newly created promoter, and with different intensities based on the promoter strength. **(c)** We sorted bacteria using fluorescence activated cell sorting (FACS) into four bins for each fluorescence color, corresponding to none, low, medium, and high fluorescence for both GFP and RFP (thus 8 bins total). We isolated inserts from cells in each bin and sequenced

them using Illumina sequencing. **(d)** Percentages at the top of the figure: for each parent sequence, the probability of a mutation creating an active promoter de novo (P_{new}). For each parent sequence (x-axis), the height of the vertical bars shows the percent of mutations creating promoters with expression strengths in each of four color-coded categories (color legend, blue: GFP, red: RFP). Note: the y-axis begins at 70%. **(e)** The probability of a mutation creating an active promoter de novo in the parent sequences (P_{new}) for both the top strand (blue: GFP) and bottom strand (red: RFP). **(f)** Percent of de novo promoters in each strength category (white to blue, see color legend) based on the number of mutations. Note: the y-axis begins at 80%. **(g,h)** Single mutations observed for each parental IS fragment (rows) and each nucleotide position (columns), together with the new gene expression they drive (blue or red, see color legend). Gray boxes indicate that no mutagenized fragment harbors the indicated nucleotide. Boxes with black circles indicate the wild-type sequence. Sequences are shown from the 5' to the 3' end. **(g)** Expression level of top DNA strand (blue, darker: higher expression). **(h)** Expression level of bottom DNA strand (red, darker: higher expression).

Mutual information identifies promoter emergence hotspots.

We next asked whether some regions within parent sequences may be biased towards evolving promoter activity through any number of mutations. To this end, we calculated, for each position (i) in each parental sequence, the mutual information (I_i) between the identity of the nucleotide in each daughter sequence and the level of gene expression (fluorescence) associated with the nucleotide[35,38,40] **(Fig 3a)**. The essence of this mutual information calculation is to divide the joint probability of a position having a particular base b *and* fluorescence score f, $p_i(b, f)$, by the product of the individual probabilities $p_i(b) \times p(f)$. This calculation can be visualized with a Venn-diagram, where the individual probabilities $p_i(b)$ and $p_i(f)$ are represented as circles, and the joint probability $p_i(b, f)$ is the area which the circles overlap (**Fig 3b**). The higher the mutual information at position i, the more that position contributes to fluorescence changes, revealing where mutations are likely to create new binding sites for transcription factors and σ factors (**Fig 3c**) (Methods).

The total mutual information, i.e. (I_i) added across all positions, indicates how likely mutations in a parent are to create promoter activity overall. This total mutual information varies substantially among parent sequences (**Fig 3d**). For the five parent sequences and both orientations, the total mutual information has a median of ~0.037 bits per parent sequence. However, two parent sequences in particular: 1L(-) and 3R(+) have exceptionally higher total mutual information, with 0.405 and 0.161 bits respectively (~4.4× and ~10.9× higher than the median of all parent sequences). These sequences also have the highest promoter emergence probability P_{new} on either DNA strand (10% and 23%, see figure 2d). In contrast, sequence 2R(-) with the least mutual information (0.00860 bits), did not harbor any emergence hotspots and had the lowest P_{new} of 2%.

Studying the distribution of mutual information within each parent sequence, we found that mutations increase fluorescence in clusters. We refer to these clusters as promoter emergence hotspots, and identified 18 such hotspots. Each parent harbored 0-4 hotspots (**Fig 3e** and **Fig S5**). To identify pertinent sequence signatures within each hotspot, we overlayed the mutual information with the location of PWM-predicted -10 and -35 boxes in the parental sequences. The majority (10 of 18) hotspots overlap with existing -10 or -35 boxes. This overlap can be explained by our analysis in Figure S5, where a subset of single mutations create weak promoters by increasing binding scores of these boxes.

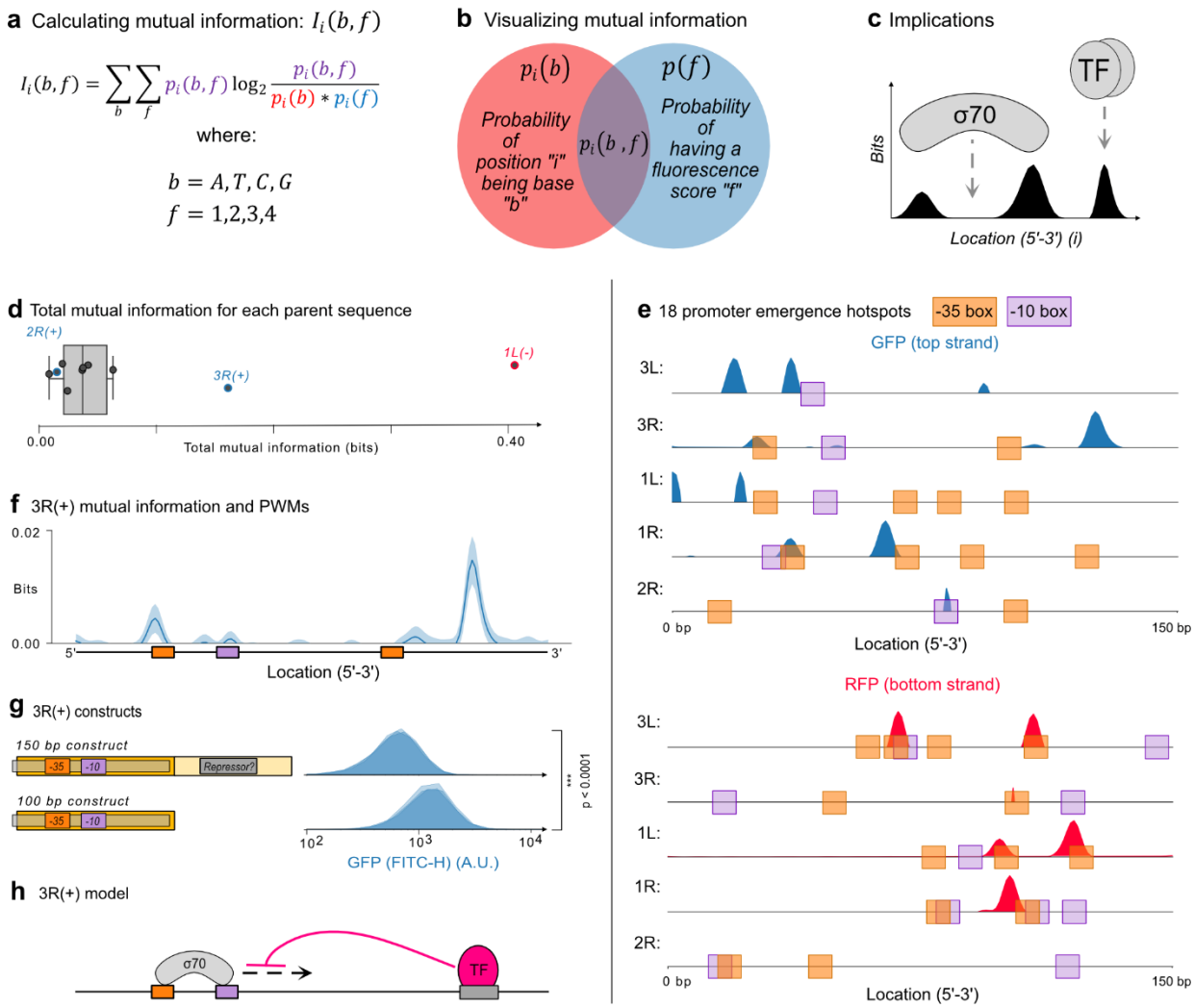

Figure 3. Mutual information reveals hotspots for de novo promoter emergence. (a) We calculated the mutual information $I_i(b,f)$ for every position *(i)* in each parent DNA sequence, and for every possible base $(b = A, T, C, G)$ and fluorescence value *(f=1,2,3,4)* with the equation shown here (see methods). **(b)** The components of the equation can be illustrated with a Venn-diagram, where the red circle corresponds to the probability $p_i(b)$ of a position i encoding base b, and the blue circle to the probability $p(f)$ of being associated with a fluorescence score *(f).* The intersection of the two circles in magenta corresponds to the joint probability $p_i(b,f)$ of position *(i)* encoding base (b) and having a score of *(f).* **(c)** Peaks of mutual information have been used in previous studies to map protein binding sites on DNA[35,40]. **(d)** Distribution of total mutual information for all five parental sequences on both top and bottom strands. **(e)** Mutual information for the five parent sequences in both top and bottom orientations. Mutual information reveals 18 regions (histograms) in the parent sequences where daughter sequences are mutated to create fluorescence activity. We refer to these as promoter emergence hotspots. Orange and magenta boxes correspond to PWM-predicted -35 and -10 boxes present in the parent sequences. Note: to illustrate the location of the hotspots within the parent sequences, the y-axis scale differs among parent sequence. See Figure S5 for a figure with identical y-axis scales. **(f)** Mutual information for GFP expression of parent 3R(+). We subsampled 50 percent of the data 30 times, and computed the mutual information for each subsample (see Methods). The solid line and the light blue region indicate the average and ±1 standard deviation over all subsamples. The locations of PWM-predicted promoter elements in the wild-type parent sequences are shown below the horizontal axis (orange = -35, magenta = -10). **(g)** Reporter constructs and their fluorescence readout measured with a flow cytometer. Top: wild-type 150 bp 3R(+) parent sequence. Bottom: 100 bp 3R(+) parent sequence without inverted repeat and candidate repressor site. Median fluorescence 1,212 arbitrary units (a.u.) vs 638 a.u., respectively; two-tailed t-test, $p=4.3\times10^{-188}$. **(h)** Model in which a downstream repressor blocks activity of the promoter on 3R(+).

Sequence 3R(+) encodes both a -10 and a -35 box spaced 15 bp apart, each overlapping with one hotspot (**Fig 3f**). However, 3R(+) also contains an additional hotspot downstream of these boxes that is not similar to a -10 or -35 box. Because repression is a frequent mode of bacterial gene regulation, we hypothesized that this hotspot may be a repressor binding site. To test this hypothesis, we measured reporter fluorescence driven by both the wild-type 150 bp construct and a shortened version of the construct that did not contain the candidate repressor binding site (**Fig 3g**). We found that this shorter construct drives almost twice as much expression than the wild-type (1,212 a.u. vs 638 a.u., two-tailed t-test, $p=4.3\times10^{-188}$). Thus, sequence 3R(+) already contains a functional promoter (**Fig 3h**). Remarkably, this promoter is located inside of the transposase coding sequence of the IS3 from which 3R(+) is derived.

This experiment demonstrates one avenue by which mutations can create promoter activity without creating new -10 and -35 boxes, namely the inactivation of a DNA sequence that represses transcription. This DNA sequence overlaps with one hotspot that do not overlap with -10 and -35 boxes in the parent sequences. It underscores our observation (Fig. S5) that de novo promoters emerge not just by creating canonical promoter motifs.

Emergent promoter activity is associated with gaining -10 boxes and Shiko Emergence.

We hypothesized that some of the remaining hotspots corresponded to regions where new -10 and -35 boxes appear upon mutation in the daughter sequences. To test this hypothesis, we computationally searched for regions in each parent sequence that gained -10 and -35 boxes from mutations that are associated with significant increases in fluorescence (Methods). We found that the largest of these changes in fluorescence occurred when mutations created new -10 boxes in three parental sequences (1L-, 1R+, and 3L-) (**Fig 4**). These parental sequences each harbor two promoter emergence hotspots close to one another (six hotspots total, i.e., two hotspots in each of three IS sequences). We refer to the left hotspot as hot_L and the right hotspot as hot_R. We briefly describe promoter emergence for these IS3 sequences.

For the parent sequence 1L(-), both hot_L and hot_R overlap with -35 boxes **(Fig 4a)**. Mutations in the daughter sequences create -10 boxes in three regions associated with increased fluorescence: one within hot_L and two within hot_R (**Fig S6**). The largest increase in fluorescence occurs when a -10 box appears in hot_R. This occurs in 149 daughters. It increases median promoter activity by ~196%, a highly significant change (1.10 a.u. → 3.26 a.u., MWU test, $q=1.52\times10^{-93}$) (**Fig 4b**). Gaining -10 boxes in the other two regions increases fluorescence by ~106% (91 daughters, 1.10 a.u.→ 2.27 a.u., MWU test, $q=1.73\times10^{-46}$) and by 80% (31 daughters, 1.10 → 1.98 a.u., MWU, $q=8.97\times10^{-15}$) (**Fig S6a-c**).

For the parent sequence 1R(+), hot_L overlaps with a preexisting -35 box (**Fig 4c**). Mutations in the daughter sequences create -10 boxes in both hot_L and hot_R (**Fig S6**). In 14 daughter sequences, mutations create a -10 box in hot_L which increases the median reporter fluorescence by ~144% (1.11 a.u. → 2.71 a.u., MWU test, $q=4.56\times10^{-9}$) (**Fig S6d,e**). In 20 daughter sequences, mutations create a -10 box in hot_R. This new -10 box is associated with a median reporter expression increase by ~170% (1.11 a.u. → 3.00 a.u., MWU test, $q=3.17\times10^{-13}$) (**Fig 4d**).

Finally, for the parent sequence 3L(-), hot_L overlaps with a -35 box (**Fig 4e**). Mutations in 15 daughter sequences create a -10 box in hot_R, which increases the median reporter fluorescence by ~91% (1.04 a.u. → 1.99 a.u., MWU test, $q=3.37\times10^{-11}$) (**Fig 4f**).

The largest increases in fluorescence occur when the -10 box is formed downstream of preexisting -35 boxes. We call this path to promoter activity *Shiko Emergence* (**Fig 4g**), a homage to the Sumo exercise, "Shiko," where the wrestler firmly plants one foot on the ground and slowly lowers their opposing foot to a firm stance (**Fig 4h**). We did not observe clear evidence of Shiko Emergence where a -35 box was gained upstream of an existing -10 box, nor did we observe it in any other parental sequence.

Gaining -10 boxes can create bidirectional promoters with unequal strengths.

1L(-) increases RFP expression by 196% when gaining a -10 box on the bottom strand (Shiko Emergence), but surprisingly, this gain also increases expression from the opposite (GFP) strand by 16% (1.13→1.31 a.u., MWU test $q=6.87\times10^{-13}$) (**Fig 4i**). Prompted by this observation, we asked whether promoter activity is generally gained on both strands when ISs gain -10 boxes. Indeed, we found that 3 of 6 gained -10 boxes associated with new promoters also increase promoter activity on the opposite strand, albeit less strongly (**Fig S7**). Specifically, gaining a -10 box increases promoter activity on the same strand as the -10 box by 196%, 144%, and 106%, whereas it increases promoter activity on the opposite strand by 16%, 18%, and 14% respectively. For two IS3 sequences that gain bidirectional promoter activity, the promoter emergence hotspots overlap on the top and bottom strands (**Fig S7**).

Because we showed that creating -10 boxes on one strand can create promoter activity on both strands (bidirectional promoters), we also asked more generally how frequently mutations in our parent sequences give rise to bidirectional promoters (**Fig 5a**). A plot of green versus red fluorescence driven from the same daughter sequence shows an L-shaped distribution, with fluorescence primarily being either red or green. We observed that emergent bidirectional promoters with *strong* expression (fluorescence $\geq$ 2.0 a.u.) on both strands are rare. Specifically, on average only ~2.6 of ~3'721 daughter sequences per parent harbor strong bidirectional promoters (3L: 2 sequences, 1L: 6, 2R: 3, 3R: 2, 1R: 0). Sequences driving GFP fluorescence at more than 2.0 a.u. drove an average RFP expression of merely 1.1 a.u.. Conversely, sequences driving RFP fluorescence at more than 2.0 a.u. drove average GFP expression of merely 1.2 a.u.

Accounting for promoter emergence hotspots.

To summarize, using the mutual information (see Figure 3) we identified 18 promoter emergence hotspots in 5 IS3 sequences. For 13 of 18 of these hotspots, we found likely explanations for promoter emergence. First, we identified 8 hotspots that gain -10 boxes associated with increased fluorescence. In 2 of 8 of these hotspots, this results in the emergence of bidirectional promoters. The strongest new promoters emerge specifically when the new -10 box is formed downstream of a preexisting -35 box, which we call Shiko Emergence. Second, we identified 1 hotspot associated with a DNA sequence which represses a functional promoter sequence in 3R(+). Finally, we identified 4 hotspots which overlapped to some degree with preexisting -10 or -35 boxes in the parent IS3 sequences. The remaining 5 of 18 hotspots have eluded characterization.

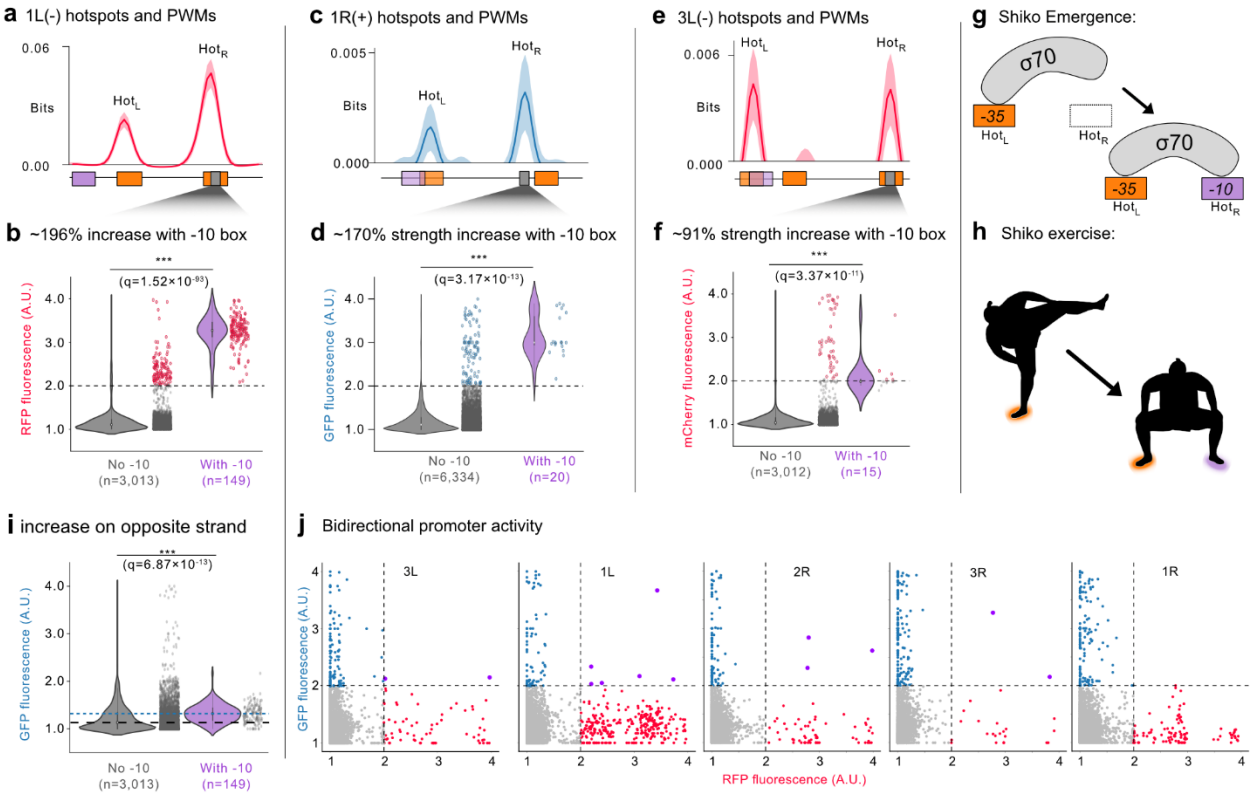

Figure 4. Shiko emergence and bidirectional promoter activity. (a) Promoter emergence hotspots (Hot$_L$ and Hot$_R$) for parent sequence 1L(-) (see also Figures S7 and S8). Solid line: mutual information. Shaded area: ±1 standard deviation (methods). Orange: -35 boxes, magenta: -10 boxes, gray: region of interest (ROI). We compared for sequence 1L(-) mutational data indicating gains of -10 boxes in **(b)** region 119:125 of 1L(-) (grey region in panel a). For this panel (and similar panels below), we plot the fluorescence values of all daughter sequences, splitting the daughter sequences into two groups. Left: those that do not gain a -10 box in the region of interest by mutation. Right: those that gain a -10 box in the region of interest. We tested the null hypothesis of indistinguishable fluorescence for the two categories with a Mann-Whitney U test, and corrected all p-values with a Benjamini-Hochberg correction to compute a q-value, where q<0.05 indicates a significant association between gaining a -10 box and increased promoter strength at a false discovery rate of 0.05. We added a dotted line at a fluorescence of 2.0 arbitrary units (a.u.), above which we consider a promoter to have weak activity, and colored each data point above this value. **(c)** Analogous to a), but for promoter emergence hotspots of parent sequence 1R(+). **(d)** Analogous to b, but for fluorescence values of all daughter sequences of 1R(+) that have or have not acquired a -10 box at region 62:68. **(e)** Analogous to a, but for promoter emergence hotspots for parent sequence 3L(-). **(f)** Analogous to b, but for fluorescence values of all daughter sequences of 3L(-) that have or have not acquired a -10 box at region 107:113. **(g)** Shiko Emergence: creation of a -10 box downstream of a preexisting -35 box. **(h)** The Sumo exercise "Shiko," where the Sumo wrestler has one foot firmly planted on the ground and slowly lowers their opposing foot to a firm stance. **(i)** Analogous to b, but for fluorescence values of all daughter sequences of 3L(-) on the *opposite* strand of the DNA that have or have not acquired a -10 box at region 119:125. Additionally, the black dashed and blue dotted lines correspond to the median fluorescence values of daughter sequences without and with the -10 box on the opposite strand, respectively. **(j)** Each scatterplot shows, for one parental sequence, the red and green fluorescence scores of each daughter sequence. Dotted lines separate fluorescence levels below and above 2.0 arbitrary units (a.u.). We additionally color the daughter sequence depending on their fluorescence. Red: RFP $\geq$ 2.0. Blue: GFP $\geq$ 2.0. Magenta: RFP and GFP $\geq$ 2.0.

DISCUSSION

In this study, we used computational predictions and experiments to estimate that at least 30% (215/706) of naturally occurring IS3 sequences harbor outward-directed promoters capable of driving the expression of adjacent genes. We then demonstrated with a massively parallel reporter assay that IS3 sequences without promoter activity can evolve promoters through one or few mutations. Many such mutations occur in hotspots of promoter emergence. Multiple hotspots involve alteration or creation of -10 or -35 boxes.

The naturally occurring outward-directed promoters we identified in wild-type IS3 sequences preferentially occur at their ends (see Fig. 1c). This preference can mostly be explained by biases in sequence composition, because it also occurs when IS3 sequences are subdivided into multiple sequence windows (bins) whose sequence is randomized. Such randomization additionally suggests that purifying selection is not removing promoters from the interior of IS3 sequences, otherwise the scrambling would have caused the incidence of such internal promoters to increase.

While most mobile DNA insertions into a genome may be deleterious, some may be beneficial because they drive the fortuitous expression of an adjacent gene. ISs that can drive such beneficial expression may be preferentially preserved in evolution. In fact, one IS3 with an outward-directed promoter is employed by its host as a mobile promoter to increase the expression of various genes during multiple directed evolution experiments[29–31]. ISs outside the IS3 family found in pathogenic strains of bacteria have also been used by their hosts for their mobile and outward-directed promoter activity to evolve antibiotic resistances[42]. These findings are also consistent with the observation that prokaryotes and eukaryotes opportunistically use mobile DNA to evolve new cis-regulatory activity[16–25]. The high incidence of outward-facing promoters we detect in wild-type mobile DNA suggests that the expression of adjacent genes may be beneficial or at least tolerated more often than hitherto assumed.

When subjecting multiple parental IS3 sequences to mutagenesis, we found that all of them (10/10) could evolve promoter activity from single mutations. In addition, single mutations that create new promoters are frequent. Specifically, among 1’549 IS3 daughter sequences with single point mutations, ~7.8% had acquired the ability to drive gene expression. This is important because single mutations provide the easiest route towards new promoters. Although most new promoters created by single mutations are weak, such weak expression can be an essential step towards further evolutionary adaptation if it affects a beneficial gene. We found that the incidence of strong promoters increases with the number of mutations an IS3 experiences, such that daughter sequences with four or more mutations were 24 times more likely to encode strong promoters than daughter sequences with single mutations (1.44% vs 0.06%). This means that once a weak promoter is created, further mutations and selection can easily enhance the expression strength of this promoter.

Comparing our observations with limited previous work suggests that the regulatory potential of both existing and mutated IS3 sequences is higher than expected by chance, i.e., from random sequences. One previous study synthesized 40 randomly generated 103 bp parent sequences (with a fixed GC-content of 50%), and tested their ability to evolve promoters de novo. This study found that ~10% (4/40) of random sequences already encoded promoter activity, and in ~60% (23/40) single point mutations sufficed to

create an active promoter [10]. Another study used a thermodynamic model to computationally sample the entire genotypic space of a 115 bp-long DNA sequence for constitutive promoter activity (4^{115}~4×10^{69} possible sequences). It estimated that ~20% of these sequences encode promoter activity, ~80% could gain promoter activity from single point mutations, and ~1.5% of all single point mutations create promoter activity[15].

By comparison, IS3s are 1.5-3 times more likely (30% vs 10-20%) to encode outward-facing promoters than random sequences. In addition, 1.25-1.67 times as many IS3s (100% vs 60-80%) can evolve promoters from single mutations compared to random sequences. Moreover ~5 times as many (7.8% vs 1.5%) single mutations in IS3s create functional promoters compared to the average random sequence. Collectively, these observations suggest that IS3s may be biased towards evolving promoter activity compared to random sequences. We caution that this assertion may be confounded by differences in study designs, such as differences in the size, base composition, and number of mutations per parent sequences. However, if true, it raises the question whether natural selection played a role in creating this potential.

Mutual information has been used previously to map existing promoter architecture[35], and our work shows that it can also help to identify potential future cis-regulatory architectures, which have been called cryptic promoters[43], cryptic low-affinity TF binding sites[44], proto-enhancers[8], etc. Specifically, mutual information helped us to identify 18 hotspots of new promoter emergence, where mutations in our 5 IS3 sequences are especially likely to give rise to new promoters (Figure 3).

We found a likely explanation for promoter emergence in 13 of 18 hotspots. In 8 of these 13 hotspots mutations create new -10 boxes that increase expression from the same strand as the -10 box. (In 2 of these 8 hotspots we additionally observed expression from the opposite strand.) The strongest new promoters emerge from these hotspots when the new -10 box lies downstream of a preexisting -35 box, a process we call Shiko Emergence. In the absence of an adjacent upstream -35 box, the new -10 box creates weaker promoters. This observation is consistent with previous estimates that ~20% of putative E. coli promoters do not encode -35 elements[45].

An additional 4 of the 13 hotspots we characterized overlap with preexisting -10 or -35 boxes, such that the alteration of a canonical promoter is likely responsible for de novo promoter emergence.

Finally, one of the 13 hotspots we characterized overlaps with a stretch of DNA that represses expression. The remaining 5 of 18 (~28%) hotspots have eluded characterization. These hotspots may correspond to binding sites for transcription factors that have been previously demonstrated to bind to IS3s[24,46–48], other RNA polymerase sigma subunits[49], non-canonical promoter motifs such as an “extended” -10 box[50], a recently characterized bidirectional promoter motif[51], and possibly other unknown paths to promoter emergence.

We found that de novo promoters can be bidirectional, but they almost exclusively drive stronger expression in one direction than the other. Specifically, only a tiny fraction (~0.07%, 13/18’607) of daughter sequences harbor bidirectional promoters that drive higher than 2 a.u. of fluorescence in both orientations simultaneously. Our observation suggests that such strong bidirectional promoters are difficult to create de novo. In contrast, a much greater fraction (~19%) of all promoters in the *E. coli*

genome are bidirectional[51]. This contrast suggests that bidirectional promoters are subject to positive selection for their bidirectionality, possibly through the coordinated expression of adjacent genes they allow.

Mobile DNA can be opportunistically used by its host to evolve novel gene regulation[18–21]. In eukaryotes, this is observed when cancer cells rewire gene regulatory networks[52]. In prokaryotes, this has been experimentally demonstrated in the evolution of antibiotic resistance[29–31,42,53]. The kind of mobile DNA we study here is well-suited for such co-option, because at least 30% of ISs can already drive the expression of a nearby gene, and even those that do not can acquire this ability through one or few mutations. This latent potential for new gene regulation raises intriguing questions about its evolutionary origins.

DATA AVAILABILITY

Supplemental Table 1 contains all relevant DNA sequences and primers used in this study.

Supplemental Table 2 contains Excel spreadsheets with the data used in each of the figure panels.

Supplemental Table 3 contains a csv file with each unique daughter sequence, its respective parent sequence, and the GFP and RFP fluorescence scores from the sort-seq experiment.

Supplemental Table 4 contains information on the number of events sorted for each replicate and day during fluorescence activated cell sorting (FACS).

Supplemental Table 5 contains an example dataset to illustrate how we calculated final fluorescence scores.

Supplemental Table 6 contains a csv file with the regions of interest and their respective associations with gaining -10 or -35 boxes and changing fluorescence. The table additionally includes the raw p-values and the corrected q-values.

Python scripts, an anaconda environment with the packages and versions from this study, and supplemental tables can be found on Github:
https://github.com/tfuqua95/promoter_emergence_mobile_DNA

Raw sequencing reads are accessible from the Sequence Read Archive (SRA) with the accession number (PRJNA1021969):
https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1021969

MATERIALS AND METHODS

DNA sequences

We acquired al IS3 sequences from the ISfinder database[28] (https://isfinder.biotoul.fr). See **Supplemental Table 1** for a list of primers and DNA sequences.

Molecular Cloning

To 1) linearize the pMR1 plasmid (cutting); 2) amplify DNA synthesized by Integrated DNA Technologies (IDT), USA); and 3) amplify the inserts from the *E. coli* genome, we used a high-fidelity Q5 polymerase (NEB, USA product #M0491). For each polymerase chain reaction (PCR), we added 1 uL of each primer at a concentration of 100 uMol, 5 uL of the provided Q5 reaction buffer, 1uL of template DNA, 1 uL 10 mM dNTPs (Thermo Scientific, USA, product #R0191), 1 uL of Q5 polymerase, and molecular grade water (AppliChem, Germany, product #A7398) to a volume of 50 uL per reaction. In a thermal cycler (C1000 Touch Thermal Cycler, Bio-Rad, USA) we performed each PCR for 30 cycles, annealing at 55°C for 30 seconds, and extending for 30 seconds at 72°C. See **Supplemental Table 1** for a list of primers and DNA sequences. We separated the PCR products by size using gel electrophoresis, isolating the band of interest with a scalpel, and purifying the product with a Qiagen QIAquick Gel Purification Kit (Qiagen, Germany, product #28706). We carried out the gel purification according to the manufacturer's instructions, apart from the final elution step, where we eluted all samples with 30 uL of H20 instead of 50 uL of Elution Buffer to increase the DNA concentration. We estimated the concentrations of each purified product using a Nanodrop One spectrophotometer (Thermo Scientific, USA).

We cloned the inserts for measuring reporter activity in this study into the pMR1 dual reporter plasmid[34] between the EcoRI (`GAATTC`) and the BamHI (`GGATCC`) restriction sites. To clone the inserts, we used the NEBuilder kit (New England Biolabs [NEB], USA, product #E2621). Specifically, in a PCR tube, we added 100 ng of linearized pMR1 plasmid (plasmid linearized via PCR), 25 ng of the insert, 5 uL of the provided NEBuilder mastermix, and molecular grade water (AppliChem, Germany, product #A7398) to a volume of 10 uL. We incubated the reaction for 1 hour at 50°C in a thermal cycler (Bio-Rad C1000 Touch Thermal Cycler, Bio-Rad, USA).

We immediately transformed the cloned products into *E. coli* DH5α electrocompetent cells (Takara, Japan, product #9027), adding 2 uL of the product to 100 uL of electrocompetent cells. We then electroporated the cells with a Bio-Rad MicroPulser (Bio-Rad, USA) and 2mm electroporation cuvettes (Cell Projects, England, product #EP-202). We allowed the transformed bacteria to recover in 1 mL of the "Super Optimal Broth with Catabolite Repression Medium" (SOC) medium provided with the electrocompetent cells, and incubated the bacteria at 37°C, shaking at 230 RPM for 1.5 hours (Infors HT, Switzerland, Multitron). After the incubation, we plated 5 uL of the bacteria onto a standard petri dish using glass beads on LB-Agar medium supplemented with 100 ug/ml of chloramphenicol. With the remaining ~995 uL of the bacteria culture, we transferred the bacteria to a 50 mL tube, and added 9 uL of LB-chloramphenicol (100 ug/ml) for a total volume of ~10 uL. We incubated the culture overnight at 37°C shaking at 230 rpm. The following morning, we combined 1 mL of the culture with 667 uL of 60% (weight / volume) glycerol, and stored the library at -80°C until needed. To verify the sequence of a cloned insert, we randomly selected three colonies from the LB-agar plate and sequenced using Sanger sequencing (MicroSynth, Switzerland).

Control sequences

We created three control plasmids to identify confounding factors contributing to IS-driven gene expression through fluorescence activated cell sorting (FACS, see Figure S1c and S3b,c). The first is a GFP-positive control, for which we cloned the bba_j23110 promoter oriented towards the GFP coding sequence of pMR1. The second is an RFP-positive control, which also harbors the bba_J23110 promoter,

but we cloned it in the opposite direction to face the RFP coding sequence of pMR1. The third control is an empty pMR1 plasmid without an insert between the BamHI and EcoRI cut sites. We cloned these inserts and transformed the products as described in the "Molecular Cloning" section.

Cytometry plots

We analyzed the flow cytometry data from .fcs files using the software `FlowCal`[54]. We prepared all plots using the python libraries `seaborn`[55] and `matplotlib`[56]. Data and software version numbers are available on the GitHub repository: https://github.com/tfuqua95/promoter_emergence_mobile_DNA

Error-prone PCR

To create the mutagenesis library, we prepared a 100 uL GoTaq (Promega, USA, product #M3001) polymerase chain reaction (PCR). For this reaction, we added 1 uL of the forward and reverse primer at a concentration of 100 uMol, 20 uL of GoTaq reaction buffer, 1uL of template DNA, 1 uL 10 mM dNTPs (Thermo Scientific, USA, product #R0191), 1 uL of GoTaq polymerase, 1 uL of 15 mMol $MnCl_2$, and molecular grade water (AppliChem, Germany, product #A7398) to a volume of 50 uL per reaction. For the template DNA, we combined an equimolar ratio of each parent sequence. See **Supplemental Table 1** for a list of primers and DNA sequences.

In a thermal cycler (C1000 Touch Thermal Cycler, Bio-Rad, USA) we performed each PCR for 30 cycles, annealing at 55°C for 30 seconds, and extending for 30 seconds at 72°C. We separated the PCR products by size using gel electrophoresis, selecting the band of interest with a scalpel, and purifying the product with a Qiagen QIAquick Gel Purification Kit (Qiagen, Netherlands, product #28706) according to the manufacturer's instructions. We only deviated from the protocol at the final elution step, where we eluted all samples with 30 uL of H_2O instead of 50 uL of TE buffer. We verified the concentrations of each purified product using a Nanodrop One spectrophotometer (Thermo Scientific, USA). We then cleaned the product and transformed it into *E. coli* , as described in "Molecular Cloning".

Because we pooled the template sequences at the beginning of the reaction, the library contained different amounts of mutant daughter sequences for each parent template sequence (Figure S3d). Because of this amplification bias, we excluded the parent sequence 2L from the analysis in this study. For future studies, we recommend carrying out individual error-prone PCR reactions per parent sequence, and then pooling the products after purification.

Fluorescence activated cell sorting (FACS)

We inoculated 100 uL of the error-prone PCR library glycerol stock (see sections "Error-prone PCR" and "Molecular Cloning") into a 1 mL LB-chloramphenicol solution (100 ug/ml chloramphenicol), and let the resulting culture grow overnight at 37°C, with shaking at 230 rpm (Infors HT, Switzerland, Multitron). The following morning, we washed the culture twice in Dulbeco's Phosphate Buffered Saline (PBS) (Sigma, USA, D8537) before sorting cells with an Aria III fluorescence activated cell sorter (BD Biosciences, USA) into eight fluorescence bins (GFP and RFP: none, low, medium, and high). To detect and measure GFP fluorescence, we used a 488 nm laser, measuring fluorescein height (FITC-H) at 750 volts. For RFP, we used a 633 nm laser, measuring phycoerythrin height (PE-H) at 510 volts.

To draw the fluorescence gates, we defined fluorescence bin boundaries based on fluorescence measurements from the following three control plasmids. GFP-control: bba j23110 promoter oriented towards GFP. RFP-control: bba j23110 promoter oriented towards RFP. Negative control: empty pMR1 plasmid. See also Figure S3b,c and "Control Sequences".

We define the lower boundary of bin #1 (none, i.e. no expression) for green fluorescence, as the minimum of (i) the lowest value of measured green fluorescence for the negative control (empty pMR1) and (ii) the lowest value of measured green fluorescence for the positive control, but for the opposing fluorophore (RFP). We define a lower boundary for the lowest fluorescence bin to prevent artefacts that may arise when a cell sorter sorts various debris into the lowest bin, including but not limited to salts, empty droplets, or bacterial waste. We define analogously the upper boundary of bin #1 as the maximum of (i) the highest value of measured green fluorescence for the negative control (empty pMR1) and (ii) the highest value of measured green fluorescence for the positive control, but for the opposing fluorophore (RFP). We define the lower and higher boundaries of bin #1 for red fluorescence analogously, but with switched roles for GFP and RFP.

We defined the lower boundary of bin #4 (high, i.e., highest expression) as the mean fluorescence of the respective (green or red) positive control. Because this was the bin with the highest fluorescence, we did not define an upper bound for bin #4.

To define bins #2 and #3, we divided the interval between the lower boundary of bin #4 and the upper boundary of bin #1 in half, and set the upper bound of bin #2 and the lower bound of bin #3 to this half-way point. See Figure S1c and S3b,c for the division of all bins.

We sorted the mutagenesis library over two consecutive days. After sorting at the end of the first day, we added 1mL of SOC medium (Sigma, USA, product #CMR002K) without antibiotics to the sorted cultures, and let the cells recover for two hours at 37°C, with shaking at 230 rpm (Infors HT, Switzerland, Multitron). Afterwards, we filled the cultures with LB-Chloramphenicol (100 ug/ml chloramphenicol) to 10 mL and let the cultures grow overnight, incubating and shaking them as just described.

To ensure that we had sorted each genotype into the appropriate fluorescence bin, we repeated the sorting on the following day using the same procedure. For example, if we had sorted cells that fluoresce at low levels into bin #2 on the first day, we sorted daughter cells from this culture on the second day only into bin #2, i.e., allowing only cells whose fluorescence falls into the boundaries of this bin to be considered for the next analysis step (DNA sequencing). This re-sorting step ensures that we only sequence genotypes that are sorted into the same fluorescence bin after both consecutive days, lowering the possibility of sorting errors. To further minimize these sorting errors and to estimate the variance in fluorescence levels, we also sorted cells into three technical triplicates (r1, r2, r3, see Figure S3g,h) on the second day. In the context of the example, this means that on the second day, we sorted the culture from bin #2 into bin #2 three times, i.e., in three replicate sorting experiments (r1-2, r2-2, r3-2). **Table S4** describes the number of cells and replicates sorted into each bin.

After the second round of sorting, we once again allowed cells to recover in SOC and grew the cultures overnight, as previously described for day 1. The following morning, we created a glycerol stock by adding

1 mL of the culture and 667 uL of 60% glycerol (weight by volume) to a cryotube and stored the cultures at -80°C. We prepared the remaining culture for DNA isolation and Illumina sequencing (see Illumina Sequencing).

To summarize, from a single mutagenesis library of bacterial cells, we sorted bacteria into 24 individual cultures, where 12 cultures correspond to green-sorted bins (GFP) and the other 12 to red-sorted bins (RFP). For both green and red fluorescence, we sorted cultures into three replicates (r1, r2, and r3), each of which we binned into four fluorescence levels (none, low, medium, and high, corresponding to bin#1, #2, #3, and #4 respectively).

Illumina Sequencing

From each sorted culture (see "Fluorescence activated cell sorting" section), we isolated plasmids using a Qiagen QIAprep Spin Miniprep Kit (Qiagen, Germany, product #27104), following the manufacturer's instructions apart from eluting the DNA in 30 uL of H_2O instead of 50 uL of Elution Buffer. From the isolated plasmids, we PCR-amplified the plasmids' inserts using Q5 polymerase (NEB, USA product #M0491) (see "Molecular Cloning" for protocol). We multiplexed the forward primer for each PCR with a unique barcode for each bin and replicate (r1-bin1-GFP, r2-bin1-GFP, r3-bin1-GFP, r1-bin2-GFP, … , r3-bin4-RFP.). In addition, we also isolated plasmids from the unsorted library and PCR-amplified their inserts with their own unique barcoded primers (24 + 1 = 25 total PCRs). See **Table S1** for a list of primers and barcodes.

We separated the resulting PCR products by size using gel electrophoresis, selecting the band of interest using a scalpel, and purifying the product with a Qiagen QIAquick Gel Purification Kit (Qiagen, Netherlands, product #28706) according to the manufacturer's instructions. We only deviated from these instructions in the final elution step, where we eluted all samples with 30 uL instead of 50 uL of the provided elution buffer. We verified the concentrations of each purified product using a Nanodrop One spectrophotometer (Thermo Scientific, USA). We then pooled the barcoded samples and sent them for Illumina paired-end read sequencing (Eurofins GmbH, Germany).

Processing sequencing results

We merged paired-end reads using Flash2[57]. Paired-end reads can be sequenced in either genetic orientation, which can result in ambiguous read orientations. To avoid such ambiguities, we took advantage of the fact that all our inserts were cloned between the palindromic 5'-EcoRI (`GAATTC`) and 3'-BamHI (`GGATCC`) restriction sites of pMR1. We searched for both sites in each paired-end read and discarded any paired-end reads that did not encode both sites. If the BamHI site was upstream of the EcoRI site, we used the reverse complement of the paired-end read for further analysis. The result was that all the paired-end reads are in the same orientation and contain both restriction sites. We then searched for the barcode upstream of the EcoRI site in each paired-end read, used it to identify the bin from which the read originated , and cropped the EcoRI and BamHI sites from each read. We counted the number of reads within each bin, and then created a table in which the first column contains a list of unique sequences. Further columns contain the number of reads associated with the unique sequences in different fluorescence bin (**Supplemental Table 3**). We henceforth refer to each unique paired-end read as a "daughter sequence."

We next removed any daughter sequence with a length different from 150 bp to focus on point mutations rather than insertions and deletions during the analysis. For each daughter sequence, we then calculated the Hamming Distance between the daughter sequence and each of the wild-type "parent sequences", i.e., the number of nucleotide differences between these sequences. We assigned the daughter sequence to the parent sequence with the lowest Hamming Distance.

We determined fluorescence scores that indicate how strongly each daughter sequence drives the expression of RFP and GFP. To this end, we first calculated a fluorescent score (F_{rep}) for each of our three technical replicates (r1, r2, and r3) with equation (1):

$$F_{rep} = \frac{\sum_1^4 (f \times Reads_f)}{\sum_1^4 (Reads_f)} \qquad (\mathbf{1})$$

In this equation, f corresponds to the different fluorescence bins (none, low, medium, and high), which we integer-encoded as $f = 1, 2, 3, 4$, respectively. $Reads_f$ corresponds to the number of reads within each fluorescence bin f. As an example, **Table S5** shows the number of read counts of a specific sequence in each bin of replicate r1**,** which yields a final green fluorescence score of F_{r1}=(1×49) + (2×4) + (3×3) + (4×0) / (49 + 4 + 3 + 0)=1.179 arbitrary units (a.u.) of fluorescence.

We calculated F_{rep} for each technical replicate and each sequence, and averaged these replicate scores to compute a final fluorescence score. In addition to sequences and read counts, **Supplemental Table 3** also provides these scores. We additionally calculated the standard deviation between the three replicates, and compared the fluorescence scores among replicates using a Pearson correlation coefficient (see Figure S3g,h).

We next filtered our data for quality control, removing daughter sequences from further data analysis that 1) are not also found in the unsorted library; 2) did not have at least one read in each of the replicates (r1, r2, r3); 3) are matched to a parent sequence with a Hamming distance larger than 10; 4) have a total number of fewer than 10 reads in all bins; 5) have a standard deviation between the three replicate fluorescence scores F_{rep} greater than 0.3.

After this filtering step, 18'607 unique daughter sequences remained for further analysis, with a mean of 3'721 daughter sequences per parent sequence (3L = 3'027, 3R=1'934, 1L = 3'162, 1R = 6'354, 2R = 4'130). See also Figure S3 for pertinent data.

Kolmogorov-Smirnov tests

We used a Kolmogorov-Smirnov (KS) test for two analyses. The first (Figure 1c) compares the distribution of promoter signatures along 706 IS3s to a uniform distributions on both the top and bottom DNA strand. It tests the null hypothesis that this distribution is a uniform distribution. For this test, we created a list of promoter signature locations that are normalized for IS3 length, where each data point is the location of an individual promoter signature along one IS3 element, and all data points lie in the interval (0,1). We created individual lists of promoter signatures both for the top and the bottom strand. To generate null uniform distributions, we used the `uniform` function from `scipy.stats` to generate a list of numbers between 0 and 1, with the length of these lists equaling the total number of promoter signatures on the top or bottom strands (871 and 1428 promoter signatures, respectively). We then compared the actual

distributions of top or bottom promoter signatures with their respective null distributions using the `kstest` function from `scipy.stats`.

For our second analysis, we examined, for each parent sequence, the locations where weak promoter activity emerges from a single point mutation in a daughter sequence (Figure 2f, g, Figure S4a), and tested the null hypothesis that the locations of these mutations follow a uniform distribution. For each parent sequence, we encoded these locations as a list of integers between 1 and 150 (the length of the mutagenized parent sequence). For the purpose of this analysis, we considered a sequence to have weak promoter activity if its fluorescence score, rounded to the nearest integer, equals two.

To generate the required uniform distribution , we used the `uniform` function from `scipy.stats` to generate a sample of uniformly distributed integers in the interval (1,150), with a sample size identical to a parent's number of daughter sequences with weak promoter activity. We then compared the actual distribution with the null distribution using the `kstest` function from `scipy.stats`.

Position weight matrices (PWMs)

We obtained the PWMs for the -10 and -35 sites as a list of -10 and -35 sequences from Regulon DB[58]. We converted the list of -10 and -35 sequences into a PWM using the `Biopython.motifs` package[59]. To calculate a PWM, we needed to provide a background nucleotide composition . Because we aimed to use the PWMs for many different kinds of sequences, we set this background composition to equal 25% each for A, T, C, and G.

From a query sequence, a PWM returns a score in bits. The higher the score is, the higher is the likelihood that the query sequence binds the protein of interest. Because PWM scores can vary widely among different query sequences, it is not always clear when a PWM score is high enough that the query can be classified as a bona fide transcription factor binding motif. In our study, unless otherwise specified, we used the well-established "Patser threshold" for this purpose, which equals the information content of a motif[32]. For PWMs used in this study, the -35 box has an information content of 3.39 bits, and for the -10 box 3.98 bits. We classified query sequences with a score greater than or equal to these thresholds as binding motifs.

When searching for promoter signatures in 706 IS3s, we first searched for -35 boxes using the -35 box PWM and the `motifs.search` function in `Biopython`[59]. The function identifies both the location and score of all motifs above the specified threshold in the query sequence. If we found a -35 motif, we then searched for -10 boxes downstream of the -35-motif, using the -10 box PWM. If the sequence also encoded a sufficiently high-scoring -10 motif 15-17 downstream of the -35 motif, we classified the sequence as having a promoter signature.

To calculate how PWM scores both the -10 and -35 boxes change in response to single mutations, we first calculated the total PWM scores for both -35 and -10 boxes in the wild-type parent sequences. We then isolated a list of daughter sequences with single mutations that created weak promoter activity (**Fig. S4b**), and a list of daughter sequences with single mutations that did not create promoter activity (**Fig. S4c**). For each subset of sequences, we calculated the PWM scores again. We then quantified the differences in the scores before and after the mutation, and created the contingency tables in Figure S4b-c, classifying a

mutation as either increasing, decreasing, or not changing the PWM score for both the -10 and -35 boxes. Because we were calculating the differences in scores, and not necessarily looking for the gain or loss of binding sites, we lowered the PWM threshold values for the -35 box (3.39 bits) and the -10 box (3.98 bits) to 0.00 bits each while searching for motifs.

Scrambling IS3 sequences and comparing their promoter signatures

To scramble the IS3 sequences, we first partitioned each IS3 into 20 equal-sized bins, rounding to the nearest whole nucleotide (see Figure S1). We shuffled the sequences in each bin using the `shuffle` function from the python `random` module. The function employs the Fisher-Yates shuffle algorithm, which starts at the last nucleotide in the bin, randomly selects an index within the unshuffled part of the sequence, swaps the nucleotide at that index with the current nucleotide, and repeats this process until it has gone through the entire length of the DNA sequence. The algorithm ensures that each nucleotide has an equal probability of being placed in any position in the shuffled DNA sequence.

Association between the gain and loss of -10/-35 boxes and significant changes in fluorescence.

For the analyses of Figures 4, S6, and S7, we computationally searched for regions in each parent sequence that gained or lost -10 and -35 boxes through mutations that are also associated with significant fluorescence increases.

To search for these regions, we moved a sliding window of length 6 bp through the parent sequence (-10 and -35 boxes have a length of 6 base pairs). Within this window, we searched for either a -10 or -35 box motif in all of the parents' mutant daughter sequences, as described in "Position Weight Matrices". If the sequences in the sliding window contained a -35 or a -10 motif above the Patser Threshold[32] (-35 box = 3.39 bits, -10 box = 3.98 bits), we added the fluorescence scores to a list of motif "positives" , and otherwise to a list of motif "negatives". If each list contained more than 10 fluorescence scores, we tested the null hypothesis that the two lists had the same fluorescence scores, using a one-sided Mann-Whitney U test with the `mannwhitneyu` function from `scipy.stats`.

We repeated these procedures for all positions of the sliding window within the parent sequence, from the beginning (position 1) to the end (position 150-6=154). We performed this analysis on all five parent sequences, both on the top and bottom strands, for -10 and -35 box motifs, and for both green and red fluorescence scores. Because we thus performed multiple hypothesis tests, we corrected all of our p-values into q-values using a Benjamini-Hochberg correction (false discovery rate = 0.05)[60]. We classified a region as significantly associated with a gain in promoter activity when the test rejected the null hypothesis at $q<0.05$.

To focus our analysis on mutations with large effects sizes, we only report fluorescence gains greater than 10% that also partially overlap with the emergence hotspots in the manuscript. **Table S6** provides all of the identified significant changes, along with a list of the p-values and corrected q-values.

Mutual information

Mutual information is a measure of dependence between two variables. We calculated the mutual information I_i between the nucleotide identity *b* at position *i* of daughter sequences of a given parent

(*1≤i≤150*), and the fluorescence score *f* for daughter sequences of a given parent. To calculate the mutual information for each parent sequence, we used equation (2) as previously described in Kinney et al. 2010[40]:

$$I_i(b,f) = \sum_b \sum_f p_i(b,f) \log_2 \frac{p_i(b,f)}{p_i(b) \times p(f)} \quad (\mathbf{2})$$

In this equation, the variable b represents all possible nucleotides ($b = A, T, C, G$). The variable f represents fluorescence scores rounded to the nearest integer ($f = 1,2,3,4$) (see "Processing sequencing results" for calculation of these scores); $p_i(b)$ corresponds to the probability (relative frequency) of each sequence variant encoding an A, T, C, or G at position (i); $p(f)$ corresponds to the probability (relative frequency) of fluorescence scores being equal to 1, 2, 3, or 4; and $p_i(b,f)$ is the corresponding joint probability, i.e., the probability of position *i* encoding an A, T, C, or G, <u>and</u> having a fluorescence score of 1, 2, 3, or 4.

The concept of mutual information is best illustrated with two simple examples. For the first, we calculate the mutual information for two consecutive and fair coin flips. Here, *b* equals the possible states of the first coin flip (heads or tails), and *f* equals the possible states of the second coin flip (heads or tails). For the event of first flipping heads and then tails, the joint probability $p_i(b,f)$ equals the probability of first flipping heads (0.5) and then tails (0.5), which is 0.5 x 0.5 = 0.25. The individual probabilities $p_i(b)$ and $p_i(f)$ correspond to the probabilities of getting heads on the first toss (0.5) and tails on the second toss (0.5), respectively. For this state (heads flip and then tails flip), and all the other possible states, the right hand side of equation (2) will equal 0, because $\log_2(1) = 0$, and thus the sum of these values also 0.

This example thus yields a mutual information of zero, because the outcome of the first and second coin flip are *independent* of each other.

Now let us assume that for whatever reason, the outcome of the first coin flip somehow influences the outcome of the second coin flip, rendering it more likely to be heads if the first flip yielded heads. In this example, the individual probabilities remain the same, with $p_i(f) = 0.5$ and $p_i(b)$ = 0.5, but the joint probabilities differ. Upon completing the calculation in Equation 2, the total mutual information will be greater than 0. The reason is that the two variables are no longer independent. The stronger this statistical dependency is, the greater is the absolute value of the mutual information.

In the context of our experiment, we calculate the mutual information between the identity of different bases at position *i* of a DNA sequence $p_i(b)$ and fluorescence scores $p_i(f)$. Positions with low mutual information correspond to promoter activity similar to the background, indicating that base identity and fluorescence are independent of each other. In contrast, for positions with high mutual information, some underlying sequence architecture causes fluorescence to be dependent on base identity. Large mutual information indicates that this dependency is strong, for example because position *i* is part of a promoter or a transcription factor binding sites.

Correcting mutual information calculations for small sample size

Small datasets can skew the mutual information calculation, just as they affect other procedures in statistics. To account for the finite number of mutant daughter sequences that we used to calculate

mutual information in Equation 2, we used a previously described correction for finite sample sizes[40], which renders the final mutual information we computed equal to equation 3:

$$I_i(b,f) = \sum_b \sum_f p_i(b,f) \log_2 \frac{p_i(b,f)}{p_i(b) \times p(f)} - \frac{(n_b - 1)(n_f - 1) \log_2 e}{2N} + O(N^{-2}) \qquad (\mathbf{3})$$

Here n_b is the number of bases (4) and n_f the number of fluorescence bins (4). *N* is the total number of mutant daughter sequences tested. The value O(N^{-2}) indicates a term that is of the order of N^{-2}. The correction term is dependent on the degrees of freedom of all possible states $(n_b - 1)(n_f - 1)$ and the size of the library itself. The larger the library, the smaller the correction term.

To visualize mutual information "hotspots," we additionally smoothened mutual information as a function of position, using a Gaussian filter implemented in the python `scipy` package `ndimage` (parameter alpha=2). We report the mutual information values (smoothened and not smoothened) in **Supplemental Table 2**.

Uncertainty of mutual information

To calculate the uncertainty of mutual information resulting from uneven sampling, potential amplification biases, sequencing errors, and FACS sorting mistakes, we indicate on our mutual information plots the magnitude of a previously described uncertainty term δI[40] in equation (4):

$$\delta I(b,f) = \frac{1}{\sqrt{2}} \sqrt{var\left(I_{naive}^{50\%}(b,f)\right)} \qquad (\mathbf{4})$$

In this equation, $I_{naive}^{50\%}$ is the mutual information computed from a random sampling of half the data, calculated as described in equation 3. To calculate $\delta I(b,f)$, we calculated $I_{naive}^{50\%}$ 30 times, and from these data, calculated the variance *var* between the 30 mutual information values. We then calculated the standard deviation by taking the square root of the variance, and divided it by the square root of 2 to get $\delta I(b,f)$. We also smoothened δI with the Gaussian filter described in the previous section.

Acknowledgements

This work was supported by the European Research Council (Grant Agreement No. 739874), the Swiss National Science Foundation (grants 31003A_172887 and 310030_208174). TF is supported by a postdoctoral fellowship from the European Molecular Biology Organization (ALTF 963-2021).

We would additionally like to thank all members of the Wagner group for discussions – both scientific and not, Philipp Schätzle and Mario Wickert from the UZH cytometry facility for their training and support, and Baxter for enforcing something resembling a work-life balance.

Author contributions

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

Competing interests

The authors declare no competing interests.

Corresponding author

Please make all correspondence to: andreas.wagner “at” ieu.uzh.ch

SUPPLEMENTAL FIGURES

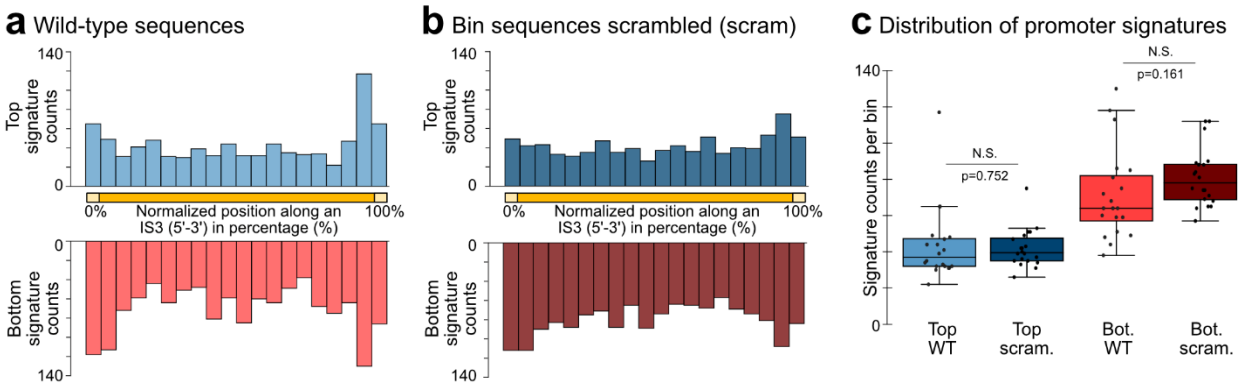

Figure S1. Scrambling IS3 sequences reveals similar promoter signature counts. (a) We plotted the identified promoter signatures as histograms with a fixed bin width of 5% (20 bins total). The top and bottom histograms correspond to promoter signatures on the top and bottom strand of the IS3s, respectively (see figure 1c). **(b)** Analogous to a, except the sequences in each bin were randomly scrambled, maintaining AT-GC content. **(c)** The number of promoter signatures in each bin. We use a two-tailed t-test to validate the null-hypothesis that the means of the wild-type distributions and scrambled distributions do not differ. The center of each box plot is the median and the boundaries indicate quartiles 1 and 3. Whiskers extend to ±1 standard deviation of the mean. N.S. = not significant.

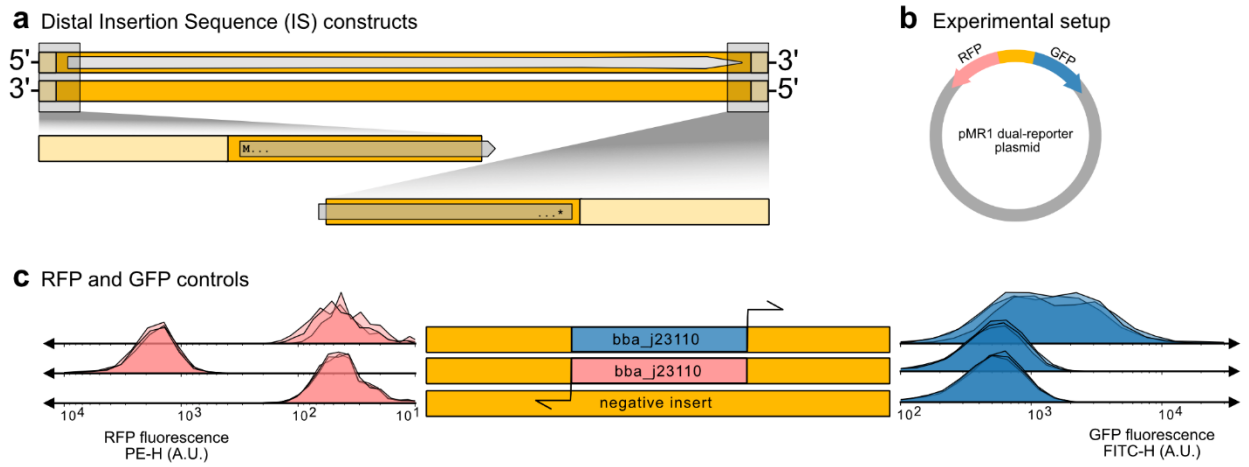

Figure S2. Control constructs for fluorescence readouts. (a) We tested the distal ends of IS3s for their ability to drive outward-directed promoter activity. We cloned 120 bps from the ends of each IS3, including the inverted repeat (light yellow) and either the beginning or the end of the Transposase coding sequence. **(b)** To test for promoter activity, we cloned the sequences into the pMR1 dual-reporter plasmid between an RFP and GFP coding sequence to simultaneously measure promoter activity in both genetic orientations**.** The top strand drives GFP expression and the bottom strand RFP expression. **(c)** We compared the promoter activity of the IS3 fragments, as quantified by fluorescence output, to three different controls. For GFP expression, we compared their fluorescence output to a positive control, in which GFP expression is driven by the bba_j23110 promoter (a moderate constitutive promoter) oriented towards the GFP coding sequence. For RFP expression we compared it to fluorescence driven by the bba_j23110 promoter oriented towards the RFP coding sequence. As negative controls, we used the fluorescence readout of pMR1 not encoding an insert, as well as the fluorescence readout of bba_j23110 but oriented in the opposite direction from that in the positive controls.

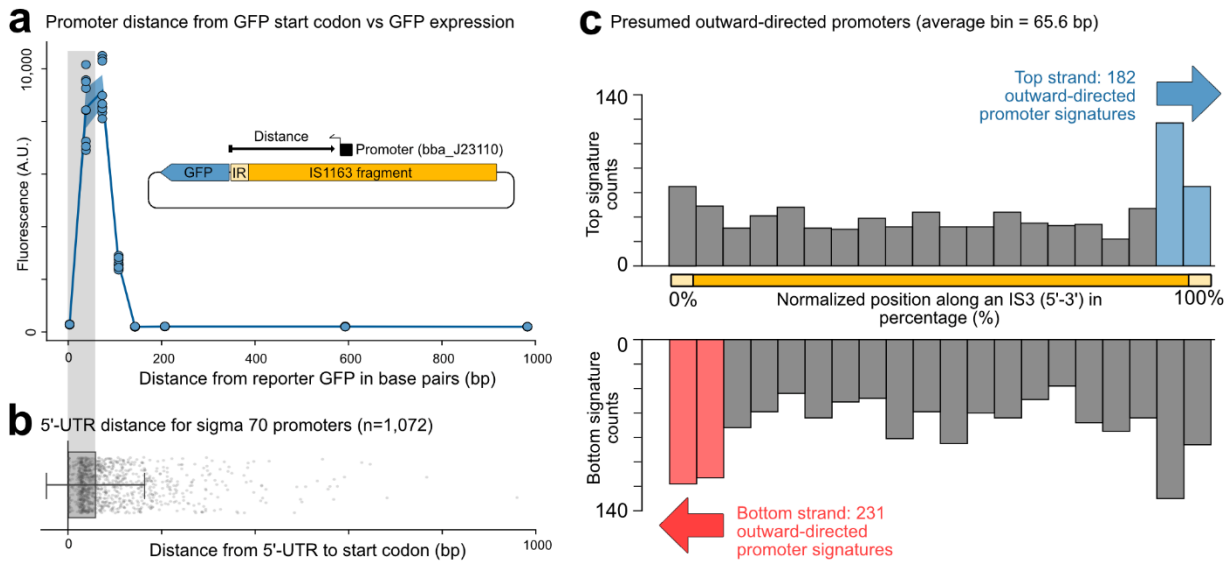

Figure S3. Empirically deriving the maximum construct size for the experiments. (a) To determine the maximum size of IS fragments allowing the detection of promoter activity within them, we cloned an IS3 fragment (IS1163) downstream of the GFP coding sequence in pMR1. We then placed a moderate-strength constitutive promoter (bba_j23110) at increasing distances from the GFP coding sequence, and measured GFP fluorescence using a plate reader. Circles represent individual measurements, the solid line the mean fluorescence value. The shaded blue area surrounding the mean represents ±1 standard deviation. **(b)** We plot the distance from the 5' untranslated region (UTR) to the start codon for sigma 70 promoters from the Regulon DB database[58]. Single points represent individual 5'-UTR lengths. The height of the grey horizontal bar indicates mean 5'-UTR length, and whiskers extend to ±1 standard deviation of the mean. **(c)** We plotted the identified promoter signatures as histograms with a fixed bin width of 5% (20 bins total). The top and bottom histograms correspond to promoter signatures on the top and bottom strand of the IS3s, respectively (see figure 1c). Extrapolating from the experiment in panel a, we presume that the average promoter signature must lie within ~150 bp from the end of an IS3 to function as an outward-directed promoter. Based on this distance, we highlight the last two bins (average size 65.6 bp each) as outward-directed promoter signatures in blue (top strand) or red (bottom strand). Other promoter signatures are in gray.

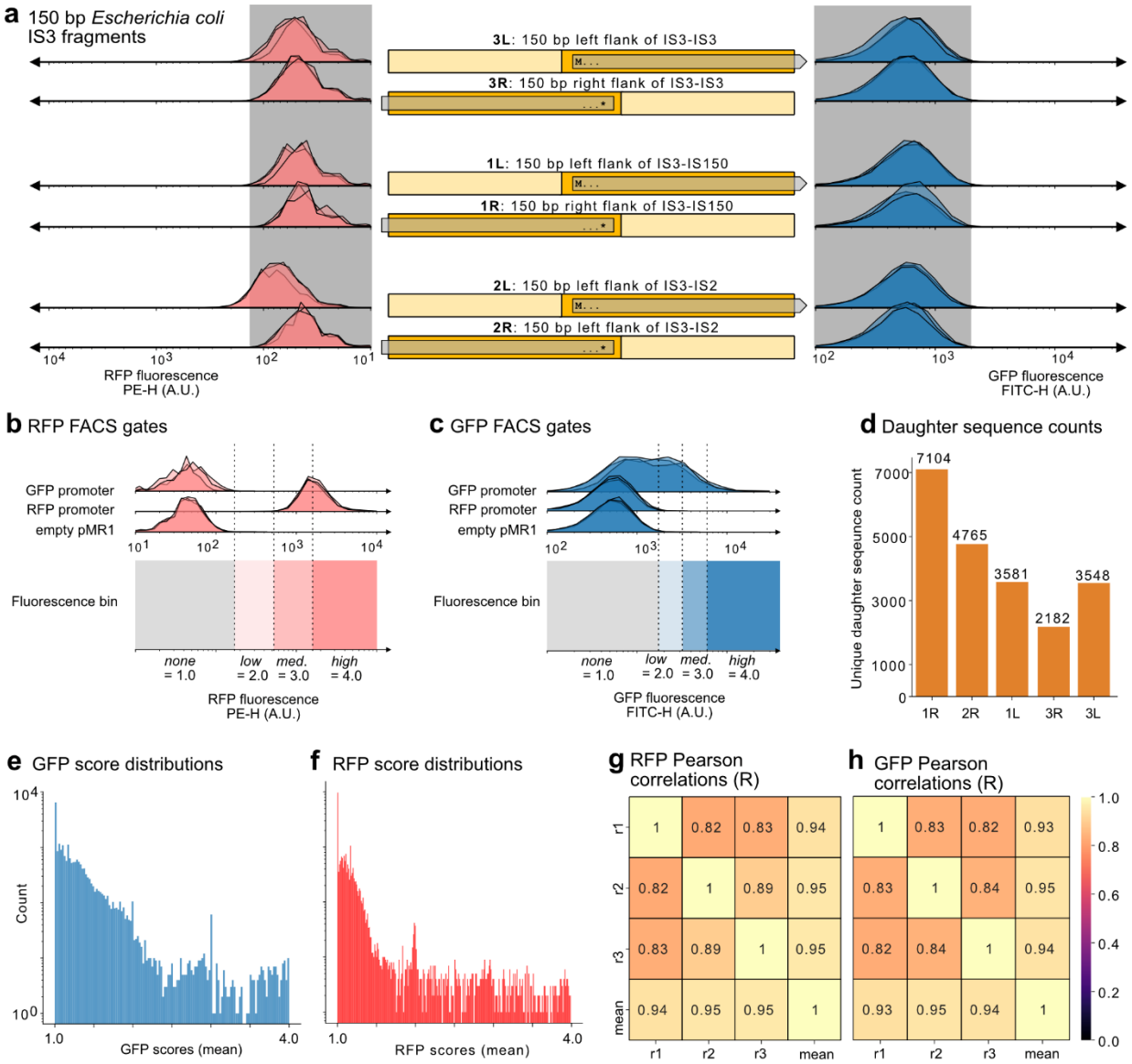

Figure S4. Sort-Seq library. (a) The IS3 parent sequence fragments used to generate the sort-seq library. For each parent sequence, the middle of the panel represents the parent sequence itself, where the bold name is the shorthand name used throughout this work. "L" denotes the left flank of the IS3 and "R" the right flank respectively. Light yellow indicates the inverted repeat sequence. The internal gray arrow represents the coding sequence, where "M…" is the start codon and "…*" the stop codon. We measured the promoter activity of each parent sequence in both genetic orientations using a flow cytometer. Left column of histograms: RFP expression, a proxy for the bottom strand's promoter activity. Right column of histograms: GFP expression, a proxy for the top strand's promoter activity. The gray box refers to the fluorescence of the negative controls (see panels b and c of Figure S1c). We removed sequence 2L from all downstream analysis because it encodes promoter activity on the bottom strand. **(b,c)** We defined fluorescence bin boundaries based on fluorescence measurements from the following three control plasmids. Top: bba j23110 promoter oriented towards GFP. Middle: bba j23110 promoter oriented towards RFP. Bottom: empty pMR1 plasmid. See also Figure S1c. See methods for details. See methods for the binning strategy. **(e,f)** Distribution of fluorescence score (see methods) for GFP **(e)** and RFP **(f).** (**g-h**) Pearson correlation coefficients between the fluorescence scores of the same sequences in different sort-seq replicates (r1, r2, r3), and of the mean fluorescence score (which is the score we used for our analyses) for both GFP (**g**) and RFP (**h**).

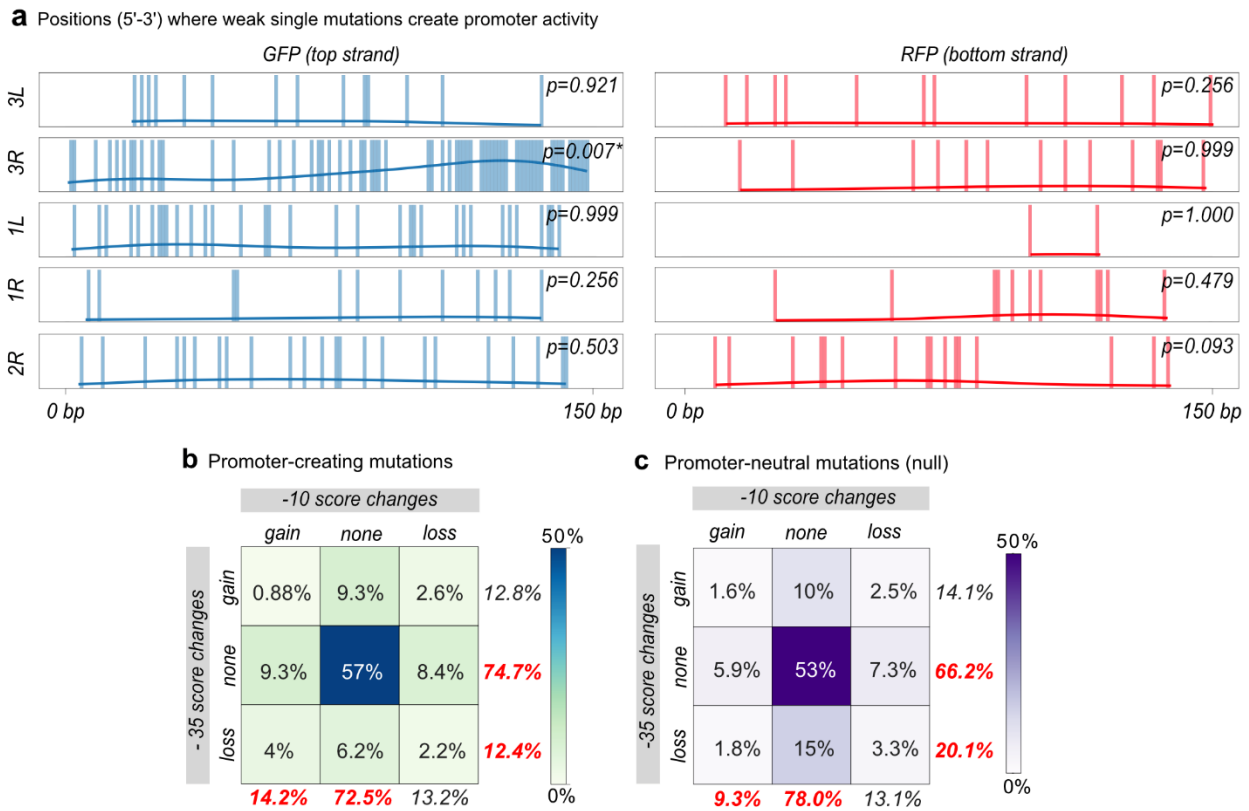

Figure S5. Single mutation biases and contingency tables. (**a**) Bar plots of the locations where single mutations create weak promoter activity along a parental DNA sequence (5'→ 3'). Curves indicate Kernel Density Estimate (KDE) plots[61,62]. P-values are based on Kolmogorov-Smirnov tests comparing the observed distribution to a uniform distribution. Left: GFP (top strand), right: RFP (bottom strand). Parent sequences with a p-value greater than 0.05 are not significantly different from a uniform distribution. We found that apart from parent sequence 3R(+), mutations causing weak promoter activity occur in a nearly uniform distribution across the sequence. (**b,c**) For each point mutation, we calculated how the mutation changed the -10 and -35 position weight matrix (PWM) scores, and classified the changes into either an increase ("gain"), a decrease ("loss"), or no change ("none") in the PWM score for both the -10 and -35 boxes. We plotted these categories in a contingency matrix. We split the matrix into two groups of changes, those that created weak promoter activity **(b)**, and those that did not create promoter activity (promoter-neutral mutations) **(c)**. We used a chi-squared test of the null hypothesis that there is no difference between the two contingency tables in b) and c). This test rejects that null hypothesis at p=0.014 (4 d.f.). We highlight that -10 scores increase in 14.2% of promoter-creating mutations vs 9.3% in promoter-neutral mutations. We also highlight that -35 scores decrease in 12.4% of cases in promoter-creating mutations, but in 20.1% of promoter-neutral mutations. Remarkably, -10 and -35 scores do not change in promoter-creating mutations 72.5% and 74.7% of the time, respectively, compared to 78.0% and 66.2% of the time in promoter-neutral mutations. Taken together, these numbers suggest that gain or loss of -10 and -35 sites, as indicated by their changing PWM scores, is not the primary path towards weak promoter emergence.

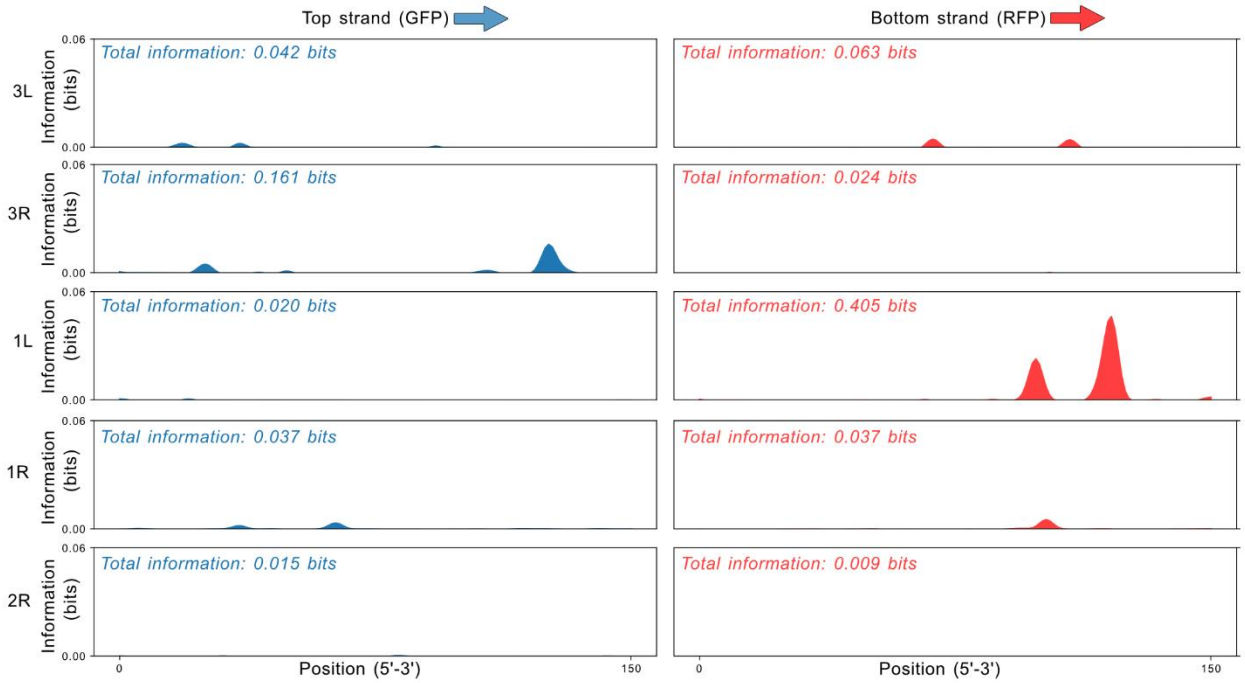

Figure S6. Mutual information for the 10 parent sequences. We calculated the mutual information for each parent sequence (rows) and sequence strand (columns) using Equation 3 (see methods). For each panel the x-axis corresponds to the position (i) in the parent sequence from the 5′ end (base pair 0) to the 3′ end (base pair 150). The y-axes show the amount of mutual information in bits. The heights of the y-axes are fixed to 0.06 bits. Total information equals the sum of mutual information values across all positions.

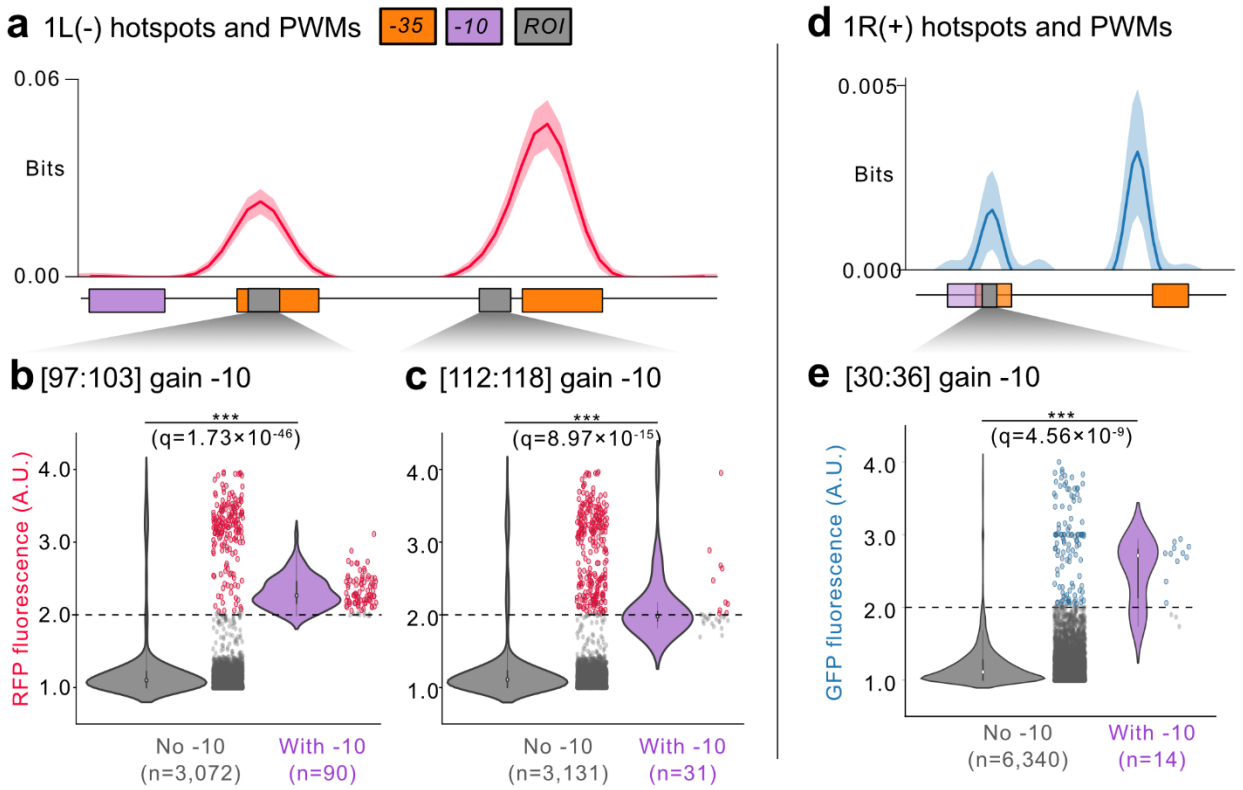

Figure S7. Gaining -10 boxes is associated with increased promoter activity. (a) Promoter emergence hotspots for parent sequence 1L(-) (see also Figure 4a). Solid line: mutual information. Shaded area: ±1 standard deviation (methods). Orange: -35 boxes, magenta: -10 boxes, gray: region of interest (ROI). We compared for sequence 1L(-) mutational data indicating gains of -10 boxes in **b)** region 97:103 (left gray region in panel a). For this panel (and the remaining similar panels in this figure), we plot the fluorescence levels of all daughter sequences, splitting the daughter sequences into two groups, left and right, which correspond to sequences that do not or that do gain a -10 box in the ROI by mutation, respectively. We tested the null hypothesis of indistinguishable fluorescence for the two categories with a Mann-Whitney U test, and corrected all p-values with a Benjamini-Hochberg correction to calculate a q-value. Q-values < 0.05 indicates a significant association between gaining a -10 box and increased fluorescence. We added a dotted line at a fluorescence of 2.0 arbitrary units (a.u.) above which we consider a promoter to have *weak* activity, and colored each data point above this value. **(c)** Analogous to b), but for region 112:118 (right gray region in panel a). **(d)** Analogous to a), but for the promoter emergence hotspots of parent sequence 1R(+). **(e)** Analogous to b) but for region 30:36 of 1R(+) (gray region in panel d).

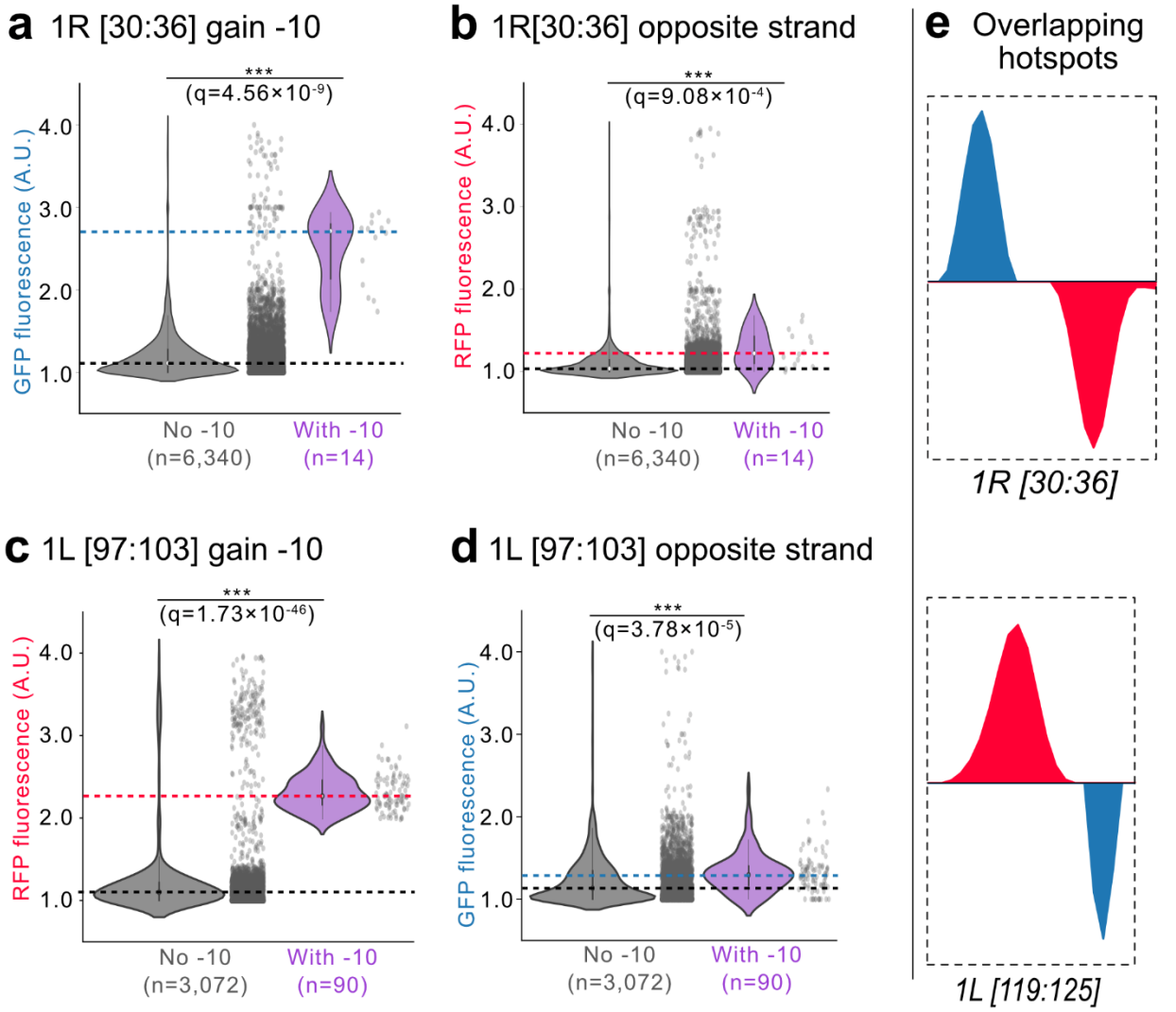

Figure S8. Gaining -10 boxes increases promoter activity on both genetic strands. (a) region of interest (ROI) 30:36 for parent sequence 1R(+). For this panel (and the remaining similar panels in this figure), we plot the fluorescence levels of all daughter sequences, splitting the daughter sequences into two groups, left and right, which correspond to sequences that do not or that do gain a -10 box in the ROI by mutation, respectively. We tested the null hypothesis of indistinguishable fluorescence for the two categories with a Mann-Whitney U test, and corrected all p-values with a Benjamini-Hochberg correction to calculate a q-value. Q-values < 0.05 indicates a significant association between gaining a -10 box and increased fluorescence. We added a dotted line for the median fluorescence values in arbitrary units (a.u.) of each group. Black: without the gained -10 box. Blue or red: with the gained -10 box. Gaining a -10 box on the top strand at region 30:36 of 1R(+) is associated with a GFP fluorescence increase of 144% on the top strand (1.10 → 2.72 a.u.). **(b)** analogous to a) except for fluorescence values from the fluorophore (RFP) on the opposite strand. RFP fluorescence increases by 14% (1.13 → 1.29 a.u.). **(c)** Analogous to a, but for the region 97:103 of 1L(-), where gaining a -10 box is associated with a RFP fluorescence increase of 144% (1.11 → 2.71 a.u.). **(d)** analogous to c) except for fluorescence values from the fluorophore (GFP) on the opposite strand. GFP fluorescence increases by 17% (1.04 a.u. → 1.22 a.u.). **(g)** Promoter emergence hotspots for (top) 1R, (bottom) 1L, where hotspots on opposite strands overlap. Promoter emergence hotspots, where gaining a -10 on one strand is associated with increasing fluorescence in both orientations simultaneously. Promoter emergence hotspots for the top (GFP) strand are in blue, and the bottom (RFP) strand in red. The peaks either lie next to each other or overlap. Note: to illustrate the location of the hotspots within the parent sequences, the y-axis scale differs among parent sequence. See Figure S5 for a figure with identical y-axis scales.

REFERENCES

1. Fuqua, T. *et al.* Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* **587**, 235–239 (2020).
2. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**, 59–69 (2012).
3. Prud'homme, B., Gompel, N. & Carroll, S. B. Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences* **104**, 8605–8612 (2007).
4. Patel, V. & Matange, N. Adaptation and compensation in a bacterial gene regulatory network evolving under antibiotic selection. *eLife* **10**, e70931 (2021).
5. Kvon, E. Z. *et al.* Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633-642.e11 (2016).
6. Loker, R. & Mann, R. S. Divergent expression of paralogous genes by modification of shared enhancer activity through a promoter-proximal silencer. *Current Biology* **32**, 3545-3555.e4 (2022).
7. Goode, D. K., Callaway, H. A., Cerda, G. A., Lewis, K. E. & Elgar, G. Minor change, major difference: divergent functions of highly conserved cis-regulatory elements subsequent to whole genome duplication events. *Development* **138**, 879–884 (2011).
8. Emera, D., Yin, J., Reilly, S. K., Gockley, J. & Noonan, J. P. Origin and evolution of developmental enhancers in the mammalian neocortex. *Proc. Natl. Acad. Sci. U.S.A.* **113**, (2016).
9. Schmitz, J. F. & Bornberg-Bauer, E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Res* **6**, 57 (2017).
10. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat Commun* **9**, 1530 (2018).
11. Gould, S. J. & Vrba, E. S. Exaptation—a Missing Term in the Science of Form. *Paleobiology* **8**, 4–15 (1982).

12. Glassford, W. J. & Rebeiz, M. Assessing constraints on the path of regulatory sequence evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20130026 (2013).
13. Galupa, R. *et al.* Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development. *Developmental Cell* **58**, 51-62.e4 (2023).
14. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat Commun* **9**, 1530 (2018).
15. Lagator, M. *et al.* Predicting bacterial promoter function and evolution from random sequences. *eLife* **11**, e64543.
16. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
17. Villanueva-Cañas, J. L., Horvath, V., Aguilera, L. & González, J. Diverse families of transposable elements affect the transcriptional regulation of stress-response genes in Drosophila melanogaster. *Nucleic Acids Res* **47**, 6842–6857 (2019).
18. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
19. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
20. de Souza, F. S. J., Franchini, L. F. & Rubinstein, M. Exaptation of Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always Strong? *Mol Biol Evol* **30**, 1239–1251 (2013).
21. Santangelo, A. M. *et al.* Ancient Exaptation of a CORE-SINE Retroposon into a Highly Conserved Mammalian Neuronal Enhancer of the Proopiomelanocortin Gene. *PLoS Genet* **3**, e166 (2007).
22. Glansdorff, N., Charlier, D. & Zafarullah, M. Activation of gene expression by IS2 and IS3. *Cold Spring Harbor symposia on quantitative biology* **45 Pt 1**, 153–156 (1981).

23. Szeverényi, I., Hodel, A., Arber, W. & Olasz, F. Vector for IS element entrapment and functional characterization based on turning on expression of distal promoterless genes. *Gene* **174**, 103–110 (1996).

24. Vandecraen, J., Chandler, M., Aertsen, A. & Van Houdt, R. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology* **43**, 709–730 (2017).

25. Jacques, P.-É., Jeyakani, J. & Bourque, G. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. *PLOS Genetics* **9**, e1003504 (2013).

26. Touchon, M. & Rocha, E. P. C. Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* **24**, 969–981 (2007).

27. Rice, P. & Kiyoshi, M. Structure of the bacteriophage Mu transposase core: A common structural motif for DNA transposition and retroviral integration. *Cell* **82**, 209–220 (1995).

28. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**, D32-36 (2006).

29. Treves, D. S., Manning, S. & Adams, J. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. *Mol Biol Evol* **15**, 789–797 (1998).

30. Charlier, D., Piette, J. & Glansdorff, N. IS3 can function as a mobile promoter in E. coli. *Nucleic Acids Res* **10**, 5935–5948 (1982).

31. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* **489**, 513–518 (2012).

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

33. 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).

34. Westmann, C. A., Alves, L. de F., Silva-Rocha, R. & Guazzaroni, M.-E. Mining Novel Constitutive Promoter Elements in Soil Metagenomic Libraries in Escherichia coli. *Front Microbiol* **9**, 1344 (2018).
35. Ireland, W. T. *et al.* Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *eLife* **9**, e55308 (2020).
36. 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).
37. Barnes, S. L., Belliveau, N. M., Ireland, W. T., Kinney, J. B. & Phillips, R. Mapping DNA sequence to transcription factor binding energy in vivo. *PLOS Computational Biology* **15**, e1006226 (2019).
38. Belliveau, N. M. *et al.* Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences* **115**, E4796–E4805 (2018).
39. Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).
40. 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).
41. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).
42. Sóki, J., Eitel, Z., Urbán, E. & Nagy, E. Molecular analysis of the carbapenem and metronidazole resistance mechanisms of Bacteroides strains reported in a Europe-wide antibiotic resistance survey. *International journal of antimicrobial agents* **41**, 122–125 (2013).
43. Pattenden, S. G., Gogol, M. M. & Workman, J. L. Features of Cryptic Promoters and Their Varied Reliance on Bromodomain-Containing Factors. *PLoS One* **5**, e12927 (2010).

44. Crocker, J. *et al.* Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell* **160**, 191–203 (2015).

45. 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).

46. Lei, G.-S., Chen, C.-J., Yuan, H. S., Wang, S.-H. & Hu, S.-T. Inhibition of IS2 transposition by factor for inversion stimulation. *FEMS Microbiol Lett* **275**, 98–105 (2007).

47. Zerbib, D., Polard, P., Escoubas, J. M., Galas, D. & Chandler, M. The regulatory role of the IS1-encoded InsA protein in transposition. *Mol Microbiol* **4**, 471–477 (1990).

48. Hu, S. T. *et al.* Functional analysis of the 14 kDa protein of insertion sequence 2. *J Mol Biol* **236**, 503–513 (1994).

49. Paget, M. S. Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. *Biomolecules* **5**, 1245–1265 (2015).

50. 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).

51. 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).

52. Ito, J. *et al.* Endogenous retroviruses drive KRAB zinc-finger protein family expression for tumor suppression. *Science Advances* **6**, eabc3020 (2020).

53. Aubert, D., Naas, T., Héritier, C., Poirel, L. & Nordmann, P. Functional characterization of IS1999, an IS4 family element involved in mobilization and expression of beta-lactam resistance genes. *Journal of bacteriology* **188**, 6506–6514 (2006).

54. Castillo-Hair, S. M. *et al.* FlowCal: A User-Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from Arbitrary to Calibrated Units. *ACS Synth. Biol.* **5**, 774–780 (2016).

55. Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).

56. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

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

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

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

60. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

61. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* **27**, 832–837 (1956).

62. Parzen, E. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **33**, 1065–1076 (1962).