nature communications

Article

https://doi.org/10.1038/s41467-025-60023-w

The latent cis-regulatory potential of mobile DNA in *Escherichia coli*

Received: 26 March 2024

Accepted: 8 May 2025

Published online: 21 May 2025

Check for updates

Timothy Fuqua[1,2] & Andreas Wagner[1,2,3]

Transposable elements can alter gene regulation in their host genome, either when they integrate into a genome, or when they accrue mutations after integration. However, the extent to which transposons can alter gene expression, as well as the necessary mutational steps, are not well characterized. Here we study the gene regulatory potential of the prominent IS3 family of transposable elements in *E.coli*. We started with 10 sequences from the ends of 5 IS3 sequences, created 18,537 random mutations in them, and measured their promoter activity using a massively parallel reporter assay. All 10 sequences could evolve de-novo promoter activity from single point mutations. De-novo promoters mostly emerge from existing proto-promoter sequences when mutations create new −10 boxes downstream of preexisting −35 boxes. The ends of IS3s harbor ~1.5 times as many such proto-promoter sequences than the *E. coli* genome. We also estimate that at least 26% of the 706 characterized IS3s already encode promoters. Our study shows that transposable elements can have a high latent cis-regulatory potential. Our observations can help to explain why mobile DNA may persist in prokaryotic genomes. They also underline the potential use of transposable elements as a substrate for evolving new gene expression.

Insertion Sequences (ISs) are simple mobile genetic elements that encode a transposase flanked by inverted repeat sequences that serve as binding sites for the transposase[1]. There are various families of ISs, which differ in the number of open reading frames they encode, and the mechanisms by which they transpose[1,2]. One of the best characterized IS families is IS3, with over 700 characterized members. IS3s are ~1 kilobase (kb) in length and contain 2 ORFs (*orfA* and *orfB*) which regulate transposition[3]. The ORFs fuse via translational frameshifting to create the transposase (*orfAB*)[4,5]. Transposition works through a "copy-paste" mechanism using a DDE-transposase[1].

Canonical prokaryotic promoters encode two motifs called the −35 and −10 boxes[6], which are spaced 15–19 base pairs apart[7], and serve as a binding site for the σ_{70} subunit of the RNA polymerases that transcribes genes[8]. Protein transcription factors (TFs) may bind near such promoter sequences to activate or repress transcription[9]. Additional σ subunits are expressed under different environmental conditions and bind to their own unique motifs[10]. In members of the IS3 family, a native promoter is usually located on the left end to transcribe the ORFs[3]. Additionally, after excision, IS3s form circular DNA intermediary "minicircles" that join the left and right ends[5]. When the right end encodes a −35 box and the left end a −10 box, a strong "junction promoter" can result from this joining[11,12].

In addition to their native promoter, ISs can also contain partial or complete outward-directed promoters[13,14]. An IS with a complete outward-directed promoter can transcribe adjacent genes[15–17]. In addition, an IS can also serendipitously integrate upstream of a genomic −10 box, enabling it to form a "hybrid promoter" with its own −35 box[1,13,18–20]. Furthermore, de-novo promoter activity has been described in ISs upon mutation[21]. Outward-directed promoter activity in ISs has led to many evolutionary innovations[22,23], including the evolution of antibiotic resistance[19,20].

These observations raise the possibility that ISs have a high inherent gene regulatory potential. However, recent studies show that even random DNA sequences can have such a potential. For example,

[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. e-mail: andreas.wagner@ieu.uzh.ch

one study in *E. coli* showed that ~10% of randomly synthesized DNA encodes promoters[8,24–26], and single mutations to non-promoter DNA can be sufficient to create new promoters[8,24,27]. It is thus not clear whether the regulatory potential of IS families like IS3 is especially high, because promoters in IS3s were not identified systematically, and are only described in few IS3s[2].

To evaluate the regulatory potential of IS3s, we created over 18,000 random mutations to 10 ends of 5 IS3 sequences, and measured their promoter activity using a massively parallel reporter assay. De-novo promoters emerge from IS3s in or around preexisting −10 and −35 boxes, primarily when single mutations create new −10 boxes downstream of preexisting −35 boxes. We find that the ends of IS3s are enriched with such proto-promoter sequences, encoding ~1.5× as many compared to the *E. coli* genome. We also estimate that at least 26% of the 706 characterized IS3s already encode complete outward-directed promoters. Overall, our study shows that IS3s do indeed have a high latent cis-regulatory potential, which could help to explain why mobile DNA persists in prokaryotic genomes.

Results

Promoters emerge from the ends of IS3s at different probabilities

To choose an appropriate size of DNA fragments for mutagenesis, we first performed a preliminary experiment that inserted a functional promoter at varying positions across an IS3 backbone, and measured fluorescence from an adjacent reporter gene. We found that promoters need to occur near the end of the IS, within 150 bp from the reporter gene to drive detectable expression (Supplementary Fig. S1). This motivates our choice to study de-novo promoters emerging from the ends of IS3s.

We amplified five 150 bp sequences from the right (R) and left (L) ends of three *E.coli* IS3s. We call these sequences 1L, 1R, 2R, 3L, and 3R (1:IS150, 2:IS2, 3:IS3) (Supplementary Fig. S3a) (Methods). We chose these particular IS3s because they are well-characterized in *E. coli* and because they did not have detectable promoter activity. We pooled these *parent sequences* and created from them a deep mutational scanning library of *daughter sequences* via error-prone polymerase chain reaction (Fig. 1a). We cloned the resulting library into the pMR1 plasmid to measure promoter activity on both DNA strands (top: GFP, bottom: RFP)[28] (Fig. 1b) and transformed the library into *E. coli*. We then carried out Sort-Seq[29–34] (Fig. 1c), separating bacteria with a cell sorter (BD Biosciences, FACSAria III) into eight fluorescence bins according to whether each daughter sequence drives no, low, medium, or high fluorescence for GFP and RFP. We sequenced the plasmid inserts from each subpopulation, and from the number of reads in each bin, calculated fluorescence scores (arbitrary units, arb. units) for the top and bottom strands, where 1.0 arb. unit corresponds to no expression and 4.0 to strong expression (Methods).

We identified 18,537 unique daughter sequences, with a mean of ~3707 daughters per parent (1L = 3154, 1R = 6335, 2R = 4110, 3L = 3013, 3R = 1925), and a median of 2.0 point mutations per daughter (±0.60) (Supplementary Fig. S3). Every parent and genetic orientation (top/bottom strand, 10/10) gave rise to daughter sequences with promoter activity, with 3173/18,537 (~17%) daughters exhibiting promoter activity on either the top or bottom DNA strand (fluorescence score ≥1.5 arb. units). Of these, 2408/18,537 (~13%) expressed GFP, 765/18,537 (~4%) expressed RFP (Supplementary Fig. S3), and 113/18,537 (~1%) expressed both GFP and RFP. For each parent sequence, we calculated the probability P_{new} that mutation creates a new promoter (Fig. 1d, e), which varies 11.5-fold among parent sequences (2 R: $P_{new} = 0.02$, 3 R: $P_{new} = 0.23$). Thus, relative to each other, some IS3 ends are biased towards evolving promoters, while others are biased against it (Fig. 1d, e).

The frequency of de novo promoters and their strength increases with further mutations

We next asked how the strength of de novo promoters relates to the number of mutations per daughter sequence. We categorized the daughter sequences by the number of mutations they harbor (1, 2, 3, and 4+), and determined promoter strengths for all daughter sequences in each category (none, weak, medium, and high) (Fig. 1f). The frequency of daughter sequences encoding promoters increases with the number of mutations. Specifically, ~15% of all single mutants are promoters in either orientation (233/1549), along with ~16% of double mutants (2343/14,663), ~19% of triple mutants (328/1710), and ~25% of mutants with 4 or more mutations (156/615). For all daughters with single mutations, 85.0% have no expression on either strand (1316/1549), 14.5% convey weak expression (224/1549), ~0.4% convey moderate expression (7/1549), and ~0.1% convey high expression (2/1549). (See Supplementary Fig. S4 for genotype-phenotype maps). Conversely, for daughters with 4 or more mutations, ~75% have no expression on either strand (459/615), ~18% weak expression (111/615), ~5% moderate expression (28/615) and ~3% high expression (17/615). Thus, the frequency of promoter emergence and the frequency of strong de novo promoters increases with number of mutations.

Antagonistic dinucleotide interactions in promoter emergence

Because ~15% of single mutants, but only ~16% of double mutants encode promoters (see Fig. 1f), we hypothesized that individual mutations often interact non-additively (epistatically). To test this hypothesis, we identified all double mutants in our dataset, and retained only those whose constituent single mutations were also present in the dataset. The resulting 13,856 double mutants allowed us to identify dinucleotide interactions (Fig. 1g). For each dinucleotide interaction, we plot the fluorescence changes of the double mutant (observed values) against the sum of the individual fluorescence changes (expected values) as a scatterplot (Fig. 1h). An observed value identical to an expected value indicates additivity (no epistasis). An observed value greater than (less than) the expected value indicates synergistic (antagonistic) epistasis. The central tendency of the distributions lies below the 1:1 additive line (+0.41 arb. units expected, 0.0 arb. units observed), meaning that dinucleotide interactions are primarily antagonistic (Fig. 1h). More specifically, most single mutations that increase promoter activity cancel out when combined (Fig. 1h). Trinucleotide and tetranucleotide interactions are also mostly antagonistic (Supplementary Fig. S5a, b).

To categorize mutant interactions, we rounded the fluorescence changes to the nearest 0.5 arb. units for each dinucleotide interaction, and found that ~47% of double mutants exhibit no change in fluorescence (mut 1 = 0.0 arb. units, mut 2 = 0.0 arb. units, muts 1 + 2 = 0.0 arb. units), ~42% are antagonistic, ~7% are synergistic, and ~4% are additive (Fig. 1i). Rounding to the nearest 0.25 arb. units increased the antagonistic category from ~42% to ~89% and lowered the no change category from ~47% to ~2% (Supplementary Fig. S5c). We also asked whether membership in each of these four epistatic categories is influenced by the distance between the mutated nucleotides, but did not observe a significant effect of distance on membership (ANOVA, $p = 0.584$) (Supplementary Fig. S5d). In sum, most individual mutations that increase fluorescence interact antagonistically, i.e., combining them reduces their individual effects on fluorescence. This epistasis is independent of the distance between the individual mutations.

Mutual information reveals emergence hotspots

To find out where promoters arise within the mutagenized IS3 parent sequences, we calculated, for each position i in each parent sequence, the mutual information I_i between the identity of the nucleotide in each daughter sequence at this position (A,T,C or G), and the level of gene expression (fluorescence) associated with the nucleotide (1.0–4.0 arbitrary units, arb. units)[30–32] (Fig. 2a).

https://doi.org/10.1038/s41467-025-60023-w

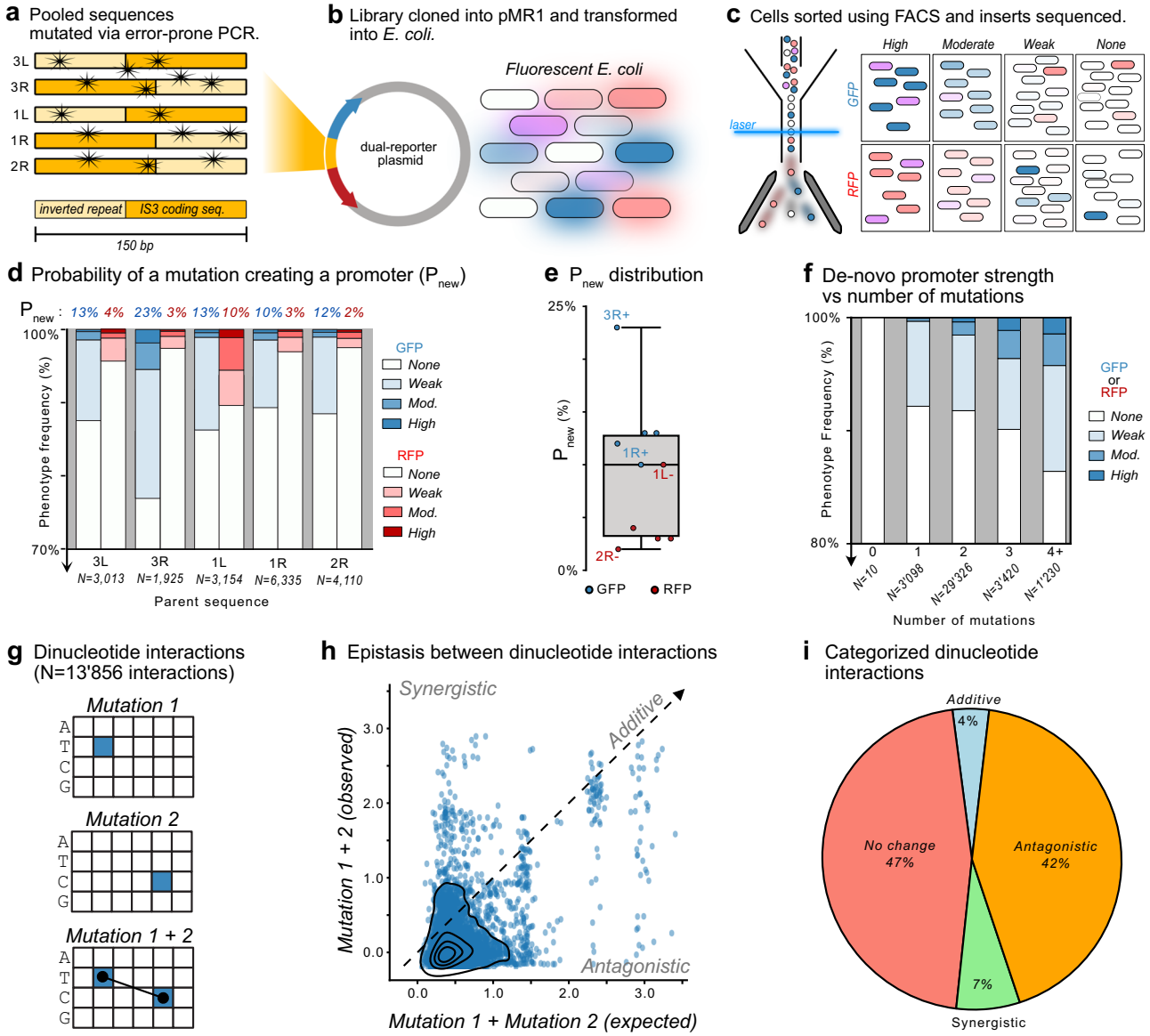

Fig. 1 | Promoters emerge with different probabilities for each parent sequence through epistatic interactions. a We pooled parental IS fragments for error-prone PCR, **b** cloned the mutagenized library into the dual reporter plasmid pMR1, and transformed the resulting plasmid library into *E. coli*. 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 subjected them to Illumina sequencing. **d** The top of the figure shows for each parent sequence (x-axis), the probability of a mutation creating an active promoter de novo (P_{new}) expressed as a percentage. 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). For the boxplot, the box corresponds to the interquartile range (IQR), the line in the box to the median, and the whiskers to 1.5 × IQR. P_{new} values are derived from the mean fluorescence scores of three technical replicates (see Methods). **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** Schematic illustration of dinucleotide interactions based on double mutants where we also have mutational data on the individual mutants. The cartoon shows an example of two hypothetical individual mutations (upper and middle) and a double mutant (lower panel) at a given position (x-axis) and nucleotide (y-axis), where the blue square corresponds to the post-mutation nucleotide. There are 13,856 such interactions in our data. **h** For each dinucleotide interaction, we plot the fluorescence changes in arbitrary units (arb. units) for the sum of the individual mutations on the x-axis as "expected" values vs. the fluorescence change of their corresponding double mutant as an "observed" value. The dotted line indicates identity between expected and observed values and corresponds to additive interactions. Interactions above the dotted line are synergistic, where the observed expression is greater than expected from the sum, and those below are antagonistic, where the observed expression is less than expected. The black striations are kernel density estimates of the scatterplot (seaborn.kdeplot)[66], and illustrate the central tendency of the data, which lies at −0.00 arb. units observed, and +0.41 arb. units expected. **i** We rounded the fluorescence changes of each dinucleotide interaction to the nearest 0.5 arb. units and classified the interactions as follows. If mut 1 = 0.0 arb. units, mut 2 = 0.0 arb. units, and mut 1 + 2 = 0.0 arb. units, the interaction is classified as *no change*. If mut1 + mut2 = mut1 + 2, the interaction is classified as *additive*. If mut1 + mut2 > mut1 + 2, the interaction is classified as *synergistic*. If mut1 + mut2 < mut1 + 2, the interaction is classified as *antagonistic*. The percentages of the pie chart correspond to the ratio of each category for the 13,856 dinucleotide interactions. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-025-60023-w

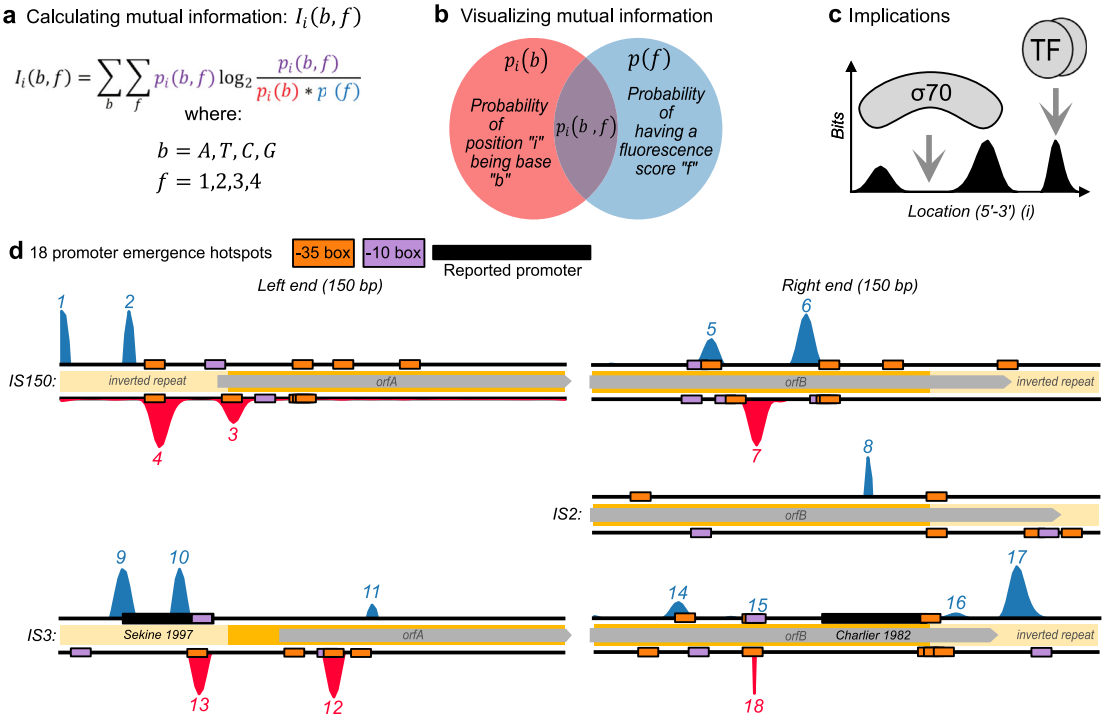

Fig. 2 | Mutual information reveals hotspots for de novo promoter emergence. **a** We calculated the mutual information $I_i(b,f)$ for every position *(i)* in each parent DNA sequence, and for every possible base *(b = A, T, C, G)* and fluorescence value (f = 1, 2, 3, 4) with the equation shown here (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[30,31]. **d** Mutual information for the five parent sequences in both top and bottom orientations. Mutual information reveals 18 regions (numbered histograms) in the parent sequences where daughter sequences are preferentially 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. Black rectangles correspond to previously characterized promoters[3,16]. Light yellow areas correspond to an ISs inverted repeats, and gray arrows to open reading frames (orf) A and B. Note: to illustrate the location of the hotspots within the parent sequences, the y-axis scale differs among parent sequence. See Supplementary Fig. S6 for a figure with identical y-axis scales. Source data are provided as a Source Data file.

The essence of this 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 $p_i(b,f)$ corresponds to the area where the circles overlap (Fig. 2b). The higher the mutual information at position *i*, the more that position contributes to fluorescence changes, revealing where mutations create sites for transcription factor and σ factor binding (Fig. 2c) (Methods).

We identified 18 regions of high mutual information which we call emergence *hotspots*[27]. Each parent sequence harbors 0-4 hotspots (see Table 1, Fig. 2d, and Supplementary Fig. S6). Based on previous works studying promoter emergence and evolution[8,27], we hypothesized that these hotspots correspond to 1) preexisting −10 and −35 boxes, 2) regions where new −10 and −35 boxes form, and 3) repressing sequences.

Promoter activity occurs in regions with preexisting −10 and −35 boxes

To test whether the hotspots correspond to preexisting −10 and −35 boxes, we overlayed the mutual information with position-weight matrix (PWM) predictions for −10 and −35 boxes. PWMs are computational models based on protein-DNA binding experiments. When given a query sequence, they return a score predicting how strongly a focal protein binds to the query sequence[35] (see methods). PWM analysis shows that the majority of hotspots (10/18) overlap with existing −10 or −35 boxes (see Table 1 and Fig. 2d). Additionally, hotspots #9 and #10 overlap with a previously characterized promoter in 3L(+) (Fig. 2d).

This overlap suggests that some mutations which create promoters coincide with increasing −10 and −35 box PWM scores (see Supplementary Fig. S4). To validate this association, we calculated the changes in PWM scores of −10 and −35 boxes before and after each single point mutation (Supplementary Fig. S7). We found that −10 box PWM 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., $p = 0.014$). In contrast −35 box scores are no 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%). Together, these findings suggest that promoters emerge when mutations create or modify −10 boxes, but not necessarily -35 boxes.

Promoter activity emerges when mutations create new −10 and −35 boxes

We next hypothesized that the hotspots correspond to regions where new −10 and −35 boxes form. To this end, we computationally searched for regions in each parent sequence where mutations in a daughter gain a −10 or a −35 box, and tested whether the gain significantly associates with increased fluorescence (Methods). We found that ~50% (9/18) of the hotspots can be explained by mutations creating new −10 boxes, and one additional hotspot can be explained by gaining a −35 box (See Fig. 3a–c and Supplementary Fig. S9. See also

Table 1 | Hotspot summary

Hotspot	Parent	Strand	Overlap	Mechanism	Figure
1	1L	+	None	Unknown	2d
2	1L	+	None	Gain −35 box	S9l
3	1L	−	−35	Gain −10 box	3b
4	1L	−	−35	Gain −10 box	3c
5	1R	+	−35|−10	Gain −10 box	S9b
6	1R	+	−35	Gain −10 box | Gain σ32 −10 box	S9c, S10b
7	1R	−	−35	Gain −10 box	S9h
8	2R	+	None	Unknown	2d
9	3L	+	Promoter	Mutations flanking −35 box	2d
10	3L	+	Promoter	Mutations flanking −10 box	2d
11	3L	+	None	Gain σ28 −35 box	S10c
12	3L	−	−10 or −35	Gain −10 box	S9e
13	3L	−	−35	Gain −10 box	S9f
14	3R	+	−35	Gain −10 box | −35 box	S9j, S8a
15	3R	+	−35|−10	−10 box	S8a
16	3R	+	None	Gain σ28 −35 box | Pausing site	S10d, S8
17	3R	+	None	Pausing site	S8
18	3R	−	−35	Gain −10 box	S9n

Supplementary Fig. S10, where we repeated this analysis using PWMs for the other sigma factors.

We highlight two examples of how gaining a new -10 box can create promoter activity (hotspots #3 and #4) in 1L(−) (Fig. 3a). First, at hotspot #3, 3064 mutant daughter sequences do not encode a −10 box (consensus TGACAT) while 90 daughter sequences do (consensus TaACAT). The newly created −10 box is located 15–17 bp downstream of three overlapping −35 boxes. Gaining this −10 box is significantly associated with a doubling of gene expression (~106% fluorescence increase from 1.10 arb. units to 2.27 arb. units, MWU test, $q = 5.72 \times 10^{-93}$) (Fig. 3b). Second, at hotspot #4, 3005 daughter sequences do not encode a −10 box (consensus: TTTCAT), while 149 daughter sequences do (consensus TaTCAT). This −10 box is located 17 bp downstream of a −35 box. Gaining the −10 box almost triples gene expression (~196% increase in fluorescence from 1.10 arb. units to 3.26 arb. units, MWU test, $q = 1.92 \times 10^{-46}$) (Fig. 3c).

We then asked how the strength of these de-novo promoters associates with their PWM scores and the spacing between the −10 and −35 boxes. From the ten hotspots associated with gaining new −10 boxes or a 35 box (Fig. 3a, b and Supplementary Fig. S9), we also searched for a pre-existing box based on PWM predictions and their proximity. For 7 of 10 hotspots, an existing −35 or −10 box occurs within 25 bp from the new −10 box, while 3 of 10 hotspots did not have a −35 box within 25 bp of the newly created box. In addition, in the latter 3 hotspots, the new −10 boxes do not have an upstream TGn motif (Fig. 3d), which can compensate for promoters without a −35 box[36]. The PWM scores of the −10 and −35 boxes for these promoters moderately correlate with their corresponding median fluorescence values (Pearson correlation, $r = 0.624$, $p = 7.48 \times 10^{-3}$) (Fig. 3d). We did not find any association between the distances between boxes and the respective fluorescence values (Pearson correlation, $r = 0.228$, $p = 0.628$) (Fig. 3e). In sum, these analyses suggest that PWM scores are a better predictor of de-novo promoter strength than spacer lengths.

A polymerase pausing site in 3R(+)

We hypothesized that promoter activity also emerges when mutations destroy repressor sequences. In particular within 3R(+), because we found a −35 box (TTTAAT, hotspot #14) and a −10 box spaced 15 bp apart (TACAAT, hotspot #15), which is indicative of an outward-directed promoter (Supplementary Fig. S8a). Additionally, a previous study described another outward-directed promoter in 3R(+) (TTGGTG and CAATTT)[16], but this promoter does not overlap with any mutual information hotspot, nor does it strongly resemble a canonical promoter.

Downstream of these putative promoters are hotspots #16 and #17 (Supplementary Fig. S8a). We hypothesized that the DNA within these hotspots represses the promoters. To find out, we created a reporter construct without the last 50 bp of 3R(+), which excludes hotspots #16 and #17. Gene expression is indeed ~1.9× higher in this shorter construct (1212 arb. units vs 638 arb. units, two-tailed *t*-test, $p = 4.3 \times 10^{-188}$), showing that the hotspot DNA represses transcription.

To find a candidate repressor protein, we tested the wild-type 3R(+) construct in a variety of genetic backgrounds that do not express the transcription factors FIS, IHF, and HNS, because these transcription factors can regulate ISs and their promoters[23,37–39]. We also tested 3R(+) in a genetic background without any Insertion Sequences[40], because some studies found that IS3 OrfA proteins can repress promoters on ISs[3,41,42]. However, none of these experiments yielded a change in gene expression (Supplementary Fig. S8b).

Although repressor binding is thus not a likely cause for the transcriptional repression we observed, the repression could also be caused by a polymerase pausing site. Indeed a sequence overlapping with hotspot #16 and #17 resembles a consensus pausing site sequence[43] (Supplementary Fig. S8c). In addition, mutations in key nucleotides of this sequence increase fluorescence (Supplementary Fig. S8d, e). Moreover, a pausing site can terminate transcription if it occurs downstream of a GC-rich Rho-utilization site (rut)[44,45], which we also find in 3R(+) (Supplementary Fig. S8a). Such terminator sequences have also been described on the end regions of other insertion sequences[46,47]. Based on this information, we conclude that 3R(+) probably gains promoter activity when mutations destroy a polymerase pausing site.

Gaining −10 boxes can create bidirectional promoters with unequal strengths

Hotspot #3 increases RFP expression by ~106% when gaining a −10 box on the corresponding (bottom) strand (see Fig. 3b). Surprisingly, this gain also increases GFP expression on the opposite strand by ~14% (1.13→1.29 arb. units, MWU test $q = 3.77 \times 10^{-5}$) (Fig. 3f). Similarly, gaining a −10 box at hotspot #4 increases RFP expression by ~196%, and GFP expression from the opposite strand by ~16% (1.13→1.32 arb. units, MWU test, $q = 6.89 \times 10^{-13}$) (Fig. 3g). We asked whether other IS3 parent sequences also gain promoter activity on both strands when daughter sequences gain −10 boxes. This is the case for 3 of 9 parents in which gained −10 boxes lead to increased expression. In these sequences expression also increases on the opposite strand, albeit more weakly (Supplementary Fig. S9). Specifically, gaining a −10 box increases fluorescence on the same strand as the −10 box by 106%, 196%, 144%, and it increases fluorescence on the opposite by 14%, 16%, and 18%, respectively.

Bidirectional promoters can have a *symmetrical −10 box*, i.e., two overlapping −10 boxes on opposite strands (consensus: TATWATA)[48]. We asked if symmetrical −10 boxes could help to explain our findings about bidirectional promoters. Indeed, for 3 of the 9 new −10 boxes associated with de novo bidirectional activity, mutations indeed created motifs that either resemble −10 boxes (0.00 bits <PWM score <3.98 bits), or that are −10 boxes (PWM score ≥3.98 bits) on the opposing strand. For example, mutations in parent 1L (see Fig. 3a, b, f)

https://doi.org/10.1038/s41467-025-60023-w

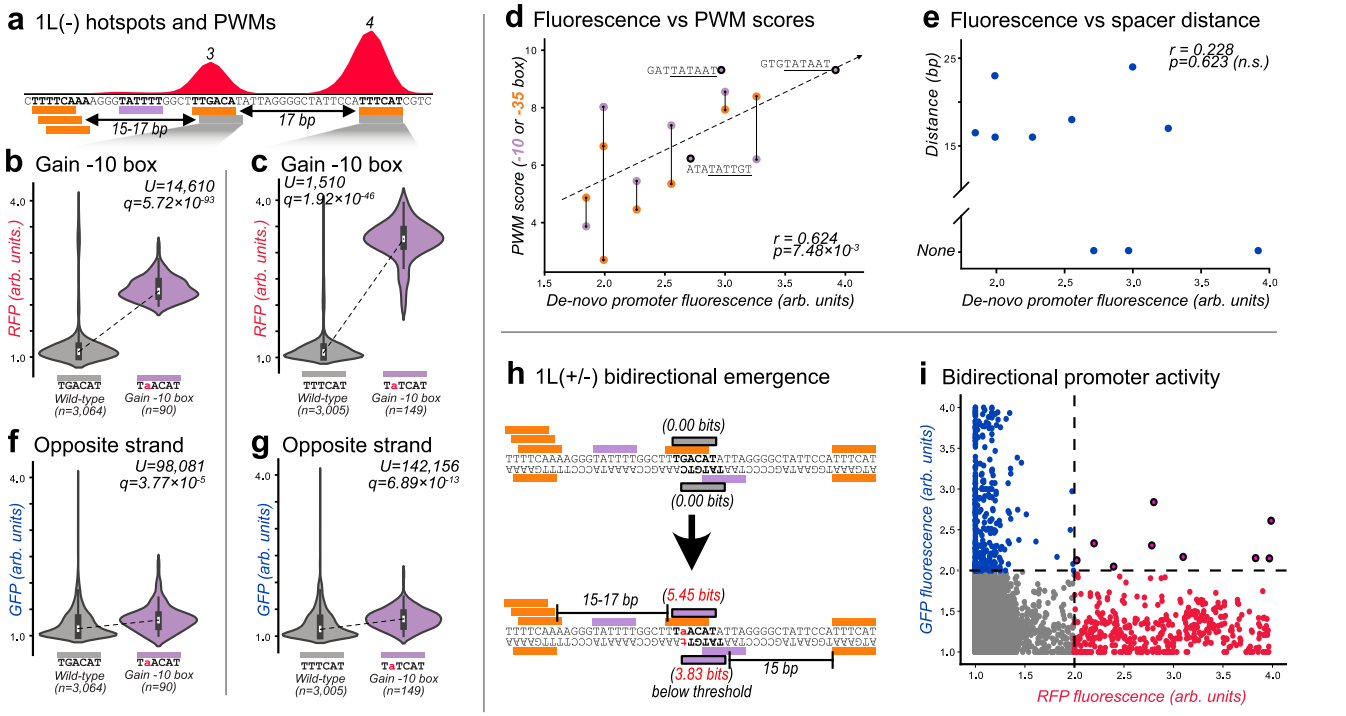

Fig. 3 | Unidirectional and bidirectional promoter emergence. a Parent 1L(−). Top: Mutual information between fluorescence scores and nucleotide identity at every position of the parent sequence. The numbers above each peak correspond to their "hotspot" identifiers (see Table 1). Bottom: Predicted −10 boxes (magenta rectangles) and −35 boxes (orange rectangles). Gray rectangles correspond to regions of interest. Bold sequences overlap with either a −10 or −35 box. Arrows indicate the distance between the regions of interest to the respective upstream −35 boxes. **b** RFP fluorescence scores for daughters without or with a −10 box in the region of interest shown in panel (**a**). The most frequent genotype is written below for each group. We test the null hypothesis that gaining a −10 box in this region of interest does not increase fluorescence (two-tailed Mann–Whitney U [MWU] test), and correct the *p*-values using a Benjamini–Hochberg correction as corresponding *q*-values (see methods, two-tailed MWU test, $q = 5.72 \times 10^{-93}$). The area of each violin plot corresponds to the kernel density estimate of each distribution. Within each violin plot is a boxplot where the white rectangle corresponds to the median, the box to the interquartile range (IQR), and the whiskers to 1.5 × IQR. **c** Analogous to (**b**) but for the other region of interest shown in (**a**) (two-tailed MWU test, $q = 1.92 \times 10^{-46}$). (**d**) Position-weight matrix (PWM) scores of the −10 and −35 boxes for ten de-novo promoters (y-axis) plotted against their respective median fluorescence scores (x-axis, arbitrary units [arb. units]). We test the null hypothesis that there is no correlation between the PWM scores and the fluorescence scores using (two-tailed Pearson's r = 0.624, $p = 7.48 \times 10^{-3}$). **e** Analogous to (**d**) but for the distance between the −10 and −35 boxes (two-tailed Pearson's r = 0.228, $p = 0.623$, n.s. non-significant). **f** Analogous to (**b**) but for the GFP fluorescence scores on the opposite strand (two-tailed MWU test, $q = 3.77 \times 10^{-5}$). **g** Analogous to (**c**) but for the GFP fluorescence scores on the opposite strand (two-tailed MWU test, $q = 6.89 \times 10^{-13}$). **h** PWM scores of the region of interest shown in (**b**) and (**f**) before (top) and after (bottom) a point mutation creates a bidirectional promoter. Orange rectangles correspond to −35 boxes, magenta rectangles to −10 boxes, and gray boxes with bold letters to the region of interest. **i** Red (x-axis) vs green (y-axis) fluorescence scores for all parents and daughter sequences. Dotted lines separate fluorescence levels below and above 2.0 arbitrary units (arb. units). We additionally colored the daughter sequences depending on their fluorescence. Red: RFP ≥ 2.0. Blue: GFP ≥ 2.0. Magenta with black outlines: RFP and GFP ≥ 2.0. Source data are provided as a Source Data file.

create a −10 box on one strand (underlined, 5′-TGACATA-3′ → 5′-TaACATA-3′), which simultaneously creates a −10-like-box on the opposite strand (underlined, 3′-TATGTCA-5′ → 3′-TATGTtA-5′) (Fig. 3h). This symmetrical −10 box lies 15–17 bp from −35 boxes on the appropriate strands. Symmetrical −10 boxes, however, are still insufficient to explain our findings. The reason is that 5 of 6 hotspots that do not create bidirectional promoters when gaining a −10 box also create symmetrical −10 boxes (see Supplementary Fig. S12). To summarize, symmetrical −10 boxes may contribute to bidirectionality, but are not sufficient to explain the origin of bidirectional promoters.

We then asked more generally how often mutations create bidirectional promoters (Fig. 3i). 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. Emergent bidirectional promoters with expression ≥2.0 arb. units on both strands are rare ($n = 9$, 3L: 2 sequences, 1L: 3, 2R: 3, 3R: 1, 1R: 0).

Thus far, we have shown that promoters emerge from either preexisting −10 and −35 boxes or by gaining new boxes upstream or downstream of preexisting boxes. When mutations create a −10 box, this can frequently create bidirectional promoter activity. Promoters also emerge when mutations create motifs for other sigma factors (see Supplementary Fig. S10). In addition, we found that hotspots #16 and #17 likely correspond to mutations that destroy a polymerase pausing site (see Supplementary Fig. S8). See Table 1 for a summary of the hotspots.

IS3s are enriched on their ends with DNA bearing promoter signatures

Previous reports have identified four IS3s with outward-directed promoters[13], but it is unclear how frequent outward-directed promoters are among the more than 700 currently characterized IS3 sequences[2]. To find out, we estimated promoter locations in IS3 sequences using two different approaches.

First, we used PWMs to identify −10 and −35 boxes in 706 IS3s (Fig. 4a). If we found a −10 and −35 box spaced 15–19 bp apart, we classified them as a *promoter signature* (Methods), which hints at but does not prove promoter activity. We found 2026 such signatures, 1263 of which (~62%) occurred on the bottom strand of DNA. To visualize the distribution of signatures, we partitioned the total length of each IS3 into 10 equidistant bins (median length

https://doi.org/10.1038/s41467-025-60023-w

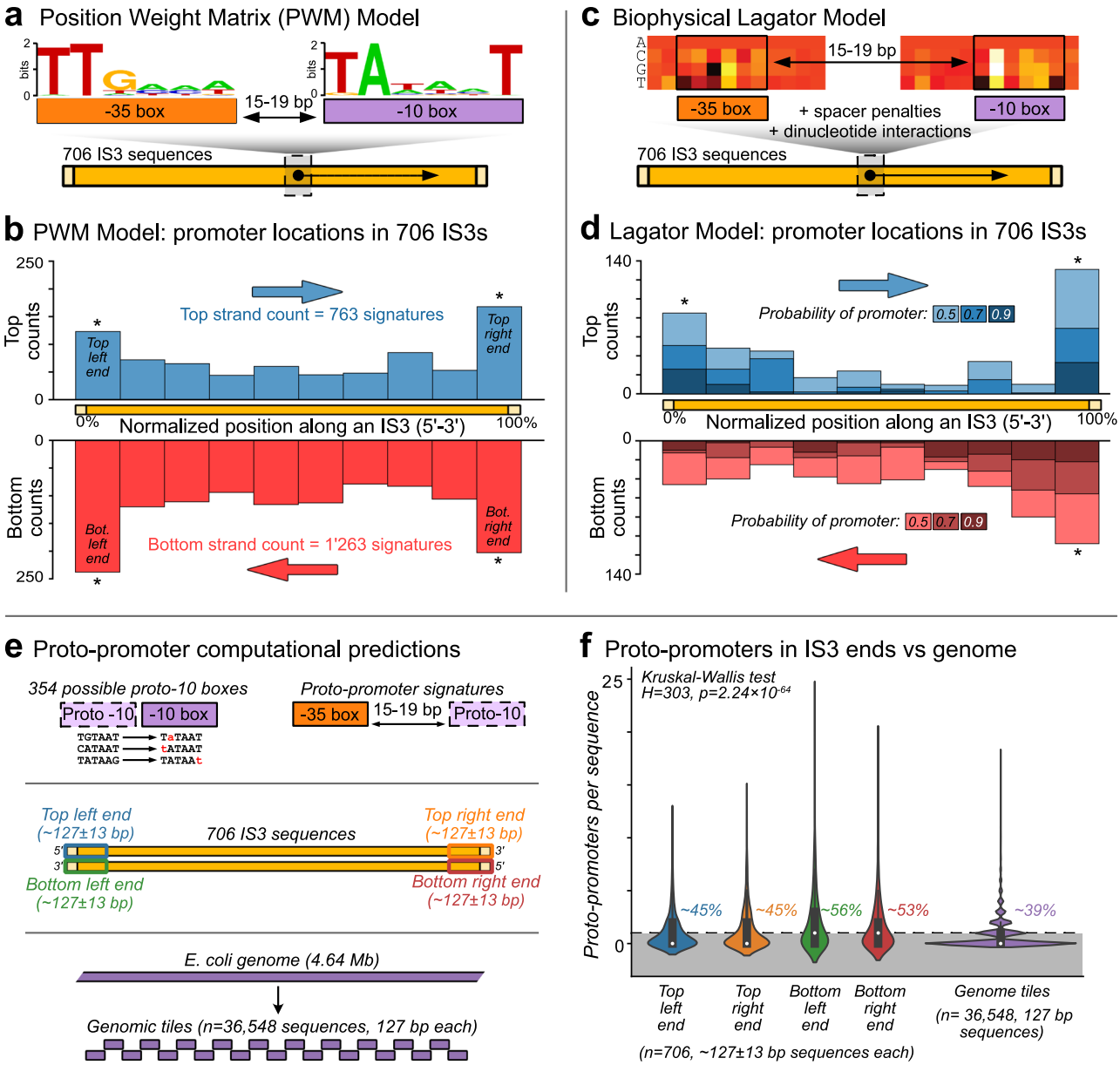

Fig. 4 | IS3s preferentially harbor promoter signatures close to their ends. **a** Sequence logos derived from position weight matrices (PWMs) depicting the likelihood of a base being bound at each position by a protein, in this case, the σ_{70} factor. The taller the letter at each position, the more 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. To computationally identify promoter signatures in IS3s, we searched for −35 and −10 boxes using PWMs spaced 15-19 base pairs (bp) apart in both the top and the bottom strands of 706 IS3s. **b** We plotted the number of identified promoter signatures as histograms with a fixed bin width of 10% of IS3 length (10 bins in total). The top and bottom histograms correspond to promoter signatures on the top and bottom strand of IS3s, respectively. **c** The biophysical Lagator model uses energy matrices to identify −10 and −35 boxes (top left and right). The binding energies of σ_{70} to DNA sequences are represented as a heatmap, in which darker colors correspond to stronger binding (top left and right). The model additionally penalizes spacer distances which deviate from the canonical 17 bp, and accounts for dinucleotide interactions between and within boxes. **d** We divided each of the 706 IS3 sequences into 10 bins with a fixed bin width (10% of IS length), and predicted the number of promoters in each bin and sequence. Because the Lagator model only returns the probability of a query sequence being a promoter, we plot the number of promoters in each bin based on whether the predicted probability is greater than or equal to 0.5, 0.7, and 0.9, for both the top strand (blue histogram, top) and the bottom strand (red histogram, bottom). **e** Top left: proto −10 boxes are 6-mer DNA sequences that a −10 box PWM does not predict to be a −10 box, but that are one point mutation away from becoming one. Top right: proto-promoters are sequences that contain a proto −10 box with a −35 box 15–19 bp upstream. Middle: we look for proto-promoters in the first and final 10% of 706 IS3s on both the top and bottom strands. We refer to these areas as the top left (blue), top right (orange), bottom left (green), and bottom right (red) ends. The median length of an IS3 is 1274 bp, with a standard deviation of 128 bp. Because the ends comprise 10% of each sequence, each end contains a sequence of 127 ± 13 bp. Bottom: we computationally split the *E. coli* genome into 36,548 non-overlapping sequences that are each 127 bp long to match the length of the ends. We call each 127 bp sequence a genomic tile, and count the proto-promoters in them as well. **f** Number of proto-promoters per sequence in IS3 ends and genomic tiles. We test the null hypothesis that the central tendency of each distribution is the same, using a Kruskal-Wallis test ($H = 303$, $p = 2.24 \times 10^{-64}$). Within each violin plot is a boxplot where the white circle corresponds to the median, the box to the interquartile range (IQR), and the whiskers to 1.5 × IQR. Source data are provided as a Source Data file.

127 ± 13 bp), and counted the signatures in each bin across all 706 IS3s (Fig. 4b). On both strands, signatures are significantly non-uniformly distributed across the IS3s. The reason is that the distal-most 10 percent of IS3s contain more signatures on average (Kolmogorov–Smirnov (KS) test, top $p = 3.80 \times 10^{-5}$, bottom $p = 1.82 \times 10^{-5}$, Methods), with ~36% of all promoter signatures occurring on the top right, top left, bottom right, or bottom left ends (732/2032).

Second, we identified putative promoters with a biophysical model we call the *Lagator model*[8] (Fig. 4c). Briefly, the Lagator Model uses binding energy matrices for −10 and −35 boxes that account for sequence composition outside of the core −10 and −35 box hexamers. The Lagator Model also incorporates dinucleotide interactions and penalizes sequences if their boxes are not canonically spaced 15–19 bp apart. Using the Lagator Model, we scanned 706 IS3s for promoter signatures, predicting the probability that each sequence and position encodes a promoter (see methods) (Fig. 4d). While the Lagator Model predicts fewer promoters overall than PWMs, it still finds an enrichment of promoter signatures on the ends of IS3s (K.S. test at 50% probability, top $p = 4.40 \times 10^{-20}$, bottom $p = 6.53 \times 10^{-21}$, Methods).

We also validated our model predictions experimentally. To this end, we randomly selected five IS3s with and five IS3s without PWM-predicted outward-directed promoter signatures within 120 bp from their ends. We synthesized these terminal 120 bp, cloned them into our reporter plasmid, transformed it into *E.* coli (Methods), and measured fluorescence with a flow cytometer (BD Biosciences, FACSAria III). All tested IS3s (5/5) predicted to have outward-directed promoters drove higher fluorescence than the controls (Supplementary Fig. S11). Two of five IS3s not predicted to have outward-directed promoters did so as well (Supplementary Fig. S11). Extrapolating from this low false positive rate, we estimate that ~20% of putative promoter signatures (406/2032) correspond to outward-directed promoters. These signatures occur on ~26% (181) of our 706 IS3s. In other words, we predict that at least a quarter of IS3s encode outward-directed promoters.

The ends of IS3 sequences have more proto-promoter signatures on average than the genome

Approximately 50% of new promoters emerge from the ends of 10 IS3 sequences when point mutations create −10 boxes downstream of existing −35 boxes (Fig. 3 and Supplementary Fig. S9). We asked whether we could predict such regions where promoters could arise through single mutations computationally. Specifically, we generated a set of all 6-mer DNA sequences that 1) were not classified as −10 boxes but 2) could become a −10 box with a single point mutation. We refer to these 354 sequences as *proto −10 boxes*. Next, we identified instances where a −35 box was located 15–19 bp upstream of a proto −10 box, defining these sequences as *proto-promoters* (Fig. 4e).

Once having identified all proto-promoters, we asked if the 10% ends of IS3s contain more proto-promoters than the *E. coli* genome. To find out, we counted the number of proto-promoters on the top right, top left, bottom right, and bottom left 10% ends of IS3s (~127 ± 13 bp each). In addition, we split the *E. coli* genome into 127 bp-long, non-overlapping sequences offset by one base pair ($n = 36{,}548$), which we call *genomic tiles*, and counted the number of proto-promoters in them as well. We then plotted the distribution of proto-promoters per sequence for each sequence category (Fig. 4f), and found a significant difference between the groups (Kruskal–Wallis test, $H = 303$, $p = 2.24 \times 10^{-64}$). Specifically, ~45% of IS3s have at least one proto-promoter on their top left end, ~45% on their top right end, ~56% on their bottom left end, and ~53% on their bottom right end. Among the genomic tiles, ~39% have at least one proto-promoter. Thus, the ends of IS3s harbor 1.15–1.44× more proto-promoters than the *E. coli* genome.

Discussion

We created a mutagenesis library from the ends of five mobile IS3 sequences to explore how new promoters emerge in mobile DNA. We find that such non-regulatory sequences can evolve promoters through as little as one point mutation. Additional mutations can further increase the strengths of such de novo promoters. Most promoter-creating mutations occur in 18 hotspots. Approximately 50% of all promoters emerge from these hotspots when mutations alter or create new −10 boxes. IS3s harbor 1.15–1.44 times as many proto-promoter sequences compared to the *E. coli* genome. Motivated by this high latent regulatory potential, we also asked how likely it is that any IS3 encodes an outward directed promoter (without further mutation), and estimate that >26% (181 of 706) of IS3s do.

While most mobile DNA insertions are deleterious, an insertion can benefit the host if the mobile DNA drives fortuitous adjacent gene expression[13,14,49]. Thus, mobile DNA that drives such expression may be preferentially preserved in evolution. For example, *E. coli* uses an IS3 as a mobile promoter to increase the expression of various genes during experimental evolution[15,16,50,51]. This has also been observed for other families of mobile DNA in pathogens, where mobile elements integrate at key genomic regions to ectopically express genes critical to conferring antibiotic resistances[52]. These findings are also consistent with the observation that eukaryotes use mobile DNA to evolve new cis-regulatory activity[53–59]. It is also worth noting that a subset of the IS3 sequences we studied are constrained from evolving new promoters. They have a low probability of evolving a promoter de novo and harbor no proto-promoters in their end regions. This lack of cis-regulatory potential in these IS3s may also influence evolution, since these IS3s would prevent the spontaneous emergence of new promoters or even abolish harmful gene expression.

All of the 10 IS3s and orientations we tested can evolve promoters from single mutations. This is important because single mutations provide the easiest route towards new promoters. Specifically, among 1549 IS3 daughter sequences with single point mutations, ~7.8% acquired promoter activity. Additionally, the incidence of strong promoters increases with the number of mutations, such that sequences with four or more mutations are 24 times more likely to encode strong promoters than those with single mutations (1.44% vs 0.06%). We observed both synergistic and antagonistic epistatic interactions between these mutations. Thus, after mutation creates a new promoter, additional mutations can readily enhance that promoter's activity without necessarily occurring in the core promoter sequence. Conversely, additional mutations can also readily remove a new promoter without necessarily occurring in the core. In other words, epistasis can allow selection to fine-tune the activity of new promoters.

A previous study in *E. coli* synthesized 40 randomly generated 103 bp parent sequences and tested their ability to evolve promoters de novo. ~10% (4/40) of the sequences already encoded promoter activity, and for ~60% (23/40), single point mutations sufficed to create promoters[24]. Another study used Sort-Seq and thermodynamic modeling to sample the genotypic space of 115 bp-long DNA sequences for promoters. It estimated that ~20% of such random sequences encode promoters, that ~80% can become promoters through single point mutations, and that ~1.5% of all single point mutations create promoters. For genomic DNA, these values were estimated to be lower[8]. By comparison, IS3s are 1.3–2.6 times more likely (26% vs 10–20%) to encode promoters than random sequences. In addition, 1.25–1.67 times as many IS3s (100% vs 60–80%) can evolve promoters from single mutations than random sequences. Moreover, ~15% of all single point mutations in IS3s create promoter activity on either DNA strand (~7.5% per strand), while in random sequences, only ~1.5% of mutations create promoter activity on a single strand[8]. Finally, IS3s contain 1.15–1.44 times as many proto-promoter sequences compared to the *E. coli* genome. Collectively, these findings suggest that IS3s may be a

better substrate to create new promoters than both random and genomic DNA. This could potentially explain why IS3s persist in host genomes.

Mutual information has been used to map existing promoter architectures[27,30]. Here, we show that it can also map *future* cis-regulatory architectures. Specifically, mutual information helped us to identify 18 promoter emergence hotspots. We found likely explanations for 16 of these 18 hotspots. In 9 of the 18 hotspots, mutations create −10 boxes that increase expression, sometimes bidirectionally (3/9), and 1/18 create a −35 box (see Supplementary Fig. S9). The new promoters frequently emerge from hotspots when the new −10 box emerges downstream of a preexisting −35 box in regions we refer to as proto-promoters. 3/10 of the mapped de-novo promoters do not have a −35 box. This is consistent with estimates that ~20% of *E. coli* promoters do not encode −35 boxes[60]. Additionally, 12/18 hotspots overlap or are directly adjacent to preexisting −10 or −35 boxes, or previously described promoters[3,16]. Three hotspots also correlate with regions where mutations create motifs for other sigma factors (see Supplementary Fig. S10), and two hotspots in 3R(+) likely correspond to a polymerase pausing site (see Supplementary Fig. S8). The remaining 2 of 18 (~11%) hotspots have eluded characterization, but may overlap with IS3-specific transcription factor binding sites[13,41,61,62], or with other unknown motifs awaiting discovery.

De novo promoters can be bidirectional, but almost exclusively drive stronger expression on one strand over the other. Only ~0.05% of daughter sequences (9/18,537) harbor bidirectional promoters with more than 2 arb. units of fluorescence from both strands. Given that ~19% of genomic *E. coli* promoters are bidirectional[48], our finding suggests that such strong genomic bidirectional promoters are difficult to create de novo and that most bidirectional promoters exist due to positive selection for their bidirectionality.

Mobile DNA can be used by its host to evolve novel gene regulation[56–59]. In eukaryotes, many mobile DNA sequences have gene regulatory activity during development[63], and many young, species-specific regulatory elements (enhancers) originate from mobile DNA[53,56]. In prokaryotes, this has been observed in the evolution of antibiotic resistance[50–52,64]. The kind of mobile DNA we study here is well-suited for such co-option, because at least 26% of IS3s may already drive the expression of nearby genes. Those that do not can acquire this ability through as little as one point mutation. This latent potential for new gene regulation raises intriguing questions about its evolutionary origins.

Methods

Bacterial strains

We tested the reporter constructs in *E. coli* DH5α electrocompetent cells (Takara, Japan, product #9027). In the experiments of Supplementary Fig. S8b, we tested 3R(+) constructs in the following backgrounds. The Δfis, Δihf, and Δhns are from the KEIO collection[38] (Δfis: *JW3229*, ΔihfA: JW1702, Δhns: *JW1225*). The ΔIS strain is derived from ref. 40 (Scarab Genomics, Clean Genome® E-6265-05, USA).

DNA sequences

We acquired all IS3 sequences from the ISfinder database[2] (https://isfinder.biotoul.fr). See Supplementary Data 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, 1 uL 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 s, and extending for 30 s at 72 °C. See Supplementary Data 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 H_20 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[28] 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 2 mm 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 h (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 Supplementary Figs. S2c 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[65]. We prepared all plots using the python libraries seaborn[66] and matplotlib[67]. 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, 1 uL 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. The supplemented $MnCl_2$ provides the mutations. See Supplementary Data 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 s, 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 (Supplementary Fig. 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 Supplementary Fig. 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 Supplementary Figs. S2c 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 1 mL of SOC medium (Sigma, USA, product #CMR002K) without antibiotics to the sorted cultures, and let the cells recover for 2 h 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 Supplementary Fig. 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). Supplementary Table S1 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 Supplementary Data 1 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[68]. 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 (Supplementary Data 2). 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 Eq. (1):

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

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, Supplementary Table S2 shows the number of read counts of a specific sequence in each bin of replicate r1, which yields a final green fluorescence score of $F_{r1} = (1 \times 49) + (2 \times 4) + (3 \times 3) + (4 \times 0)/(49 + 4 + 3 + 0) = 1.179$ arbitrary units (arb. units) 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, Supplementary Data 2 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 Supplementary Fig. 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, 18,537 unique daughter sequences remained for further analysis, with a mean of 3707 daughter sequences per parent sequence (3 L = 3013, 3 R = 1925, 1 L = 3154, 1 R = 6335, 2 R = 4110). See also Supplementary Fig. S3 for pertinent data.

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[69]. We converted the list of −10 and −35 sequences into a PWM using the Biopython.motifs package[70]. 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 the likelihood is 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[35]. 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[70]. 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 15–19 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 (Supplementary Fig. S7), and a list of daughter sequences with single mutations that did not create promoter activity. 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 Supplementary Fig. S7, 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.

Identifying promoter signatures with the Lagator model

For a given query sequence, we first computationally extracted the first 36 bps of the sequence. We then computationally added 40 G's upstream and downstream of this extracted sequence. We calculated the probability that the resulting 116 bp sequence has promoter activity (P_{on}) using the "Extended Model" from ref. 8, which we refer to as the Lagator Model. We then repeated this procedure, sliding a 36 bp long window in +1 bp increments across the query to extract 36 bp subsequences, and recalculated P_{on} for each subsequence. This computation results in an array of promoter probabilities, which when plotted along the length of the original sequence, reveals peaks, i.e., regions where the Lagator model predicts a promoter's location. We

identified these peaks using find_peaks from scipy.signal, which returns the position of peaks above a height of either 0.5, 0.7, and 0.9, numbers that refer to the respective probabilities that a promoter is present in Fig. 4d. We performed this sliding window analysis for each of the 706 IS3s. To plot the locations of the promoter signatures as a histogram, we normalized the positions of the identified peaks to the length of the respective sequences from which the peaks are derived (see Fig. 4d).

Association between the gain and loss of −10/−35 boxes and fluorescence changes

For the analyses of Fig. 3 and Supplementary Fig. S9, 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[35] (−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 two-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 = 144). 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)[71]. 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. Supplementary Data 3 provides all of the identified significant changes, along with a list of the *p*-values and corrected *q*-values.

For hotspot #12 (Supplementary Fig. S9e), we hypothesized that mutations that increase the PWM score of an existing −10 box are associated with increased promoter activity. We tested this hypothesis by tabulating the fluorescence scores of (i) daughters with a −10 box PWM score less than or equal to the wild-type score, and (ii) daughters with a −10 box PWM score greater than the wild-type score. We then tested the null hypothesis that the two categories of daughters had the same fluorescence scores using a two-sided Mann–Whitney U test with the mannwhitneyu function from scipy.stats. Because this analysis was done post-hoc, we do not report a *q*-value.

For hotspot #7 (Supplementary Fig. S9h), we hypothesized that mutations create a weak −10 box that was below our threshold limit of 3.98 bits. We tested this hypothesis by tabulating the fluorescence scores of (i) daughters with a PWM score of 0.00 bits at the given position, and (ii) daughters with a PWM score greater than 0.00 bits (but less than 3.98 bits). We then tested the null hypothesis that the two groups of daughters had the same fluorescence scores using a Mann–Whitney U test. Because this analysis was done post-hoc, we do not report a *q*-value.

For hotspots #2, #14, and #18 (Supplementary Fig. S9j, l), we hypothesized that mutations create a −10 box and a −35 box, respectively, but our analysis could not detect this because the boxes were created in fewer than 10 daughter sequences. We grouped daughters with and without the box of interest, but did not have the necessary sample sizes to carry out a Mann–Whitney U test. For this reason, we do not provide a *p* or *q*-value.

Testing additional sigma factor binding motifs

We acquired position weight matrices (PWMs) for additional sigma (σ) factors. Specifically, we acquired the σ_{32} −35 and −10 box PWMs from ref. 72, the σ_H −35 and −10 box PWMs from ref. 73, the σ_{28} −35 and −10 boxes from ref. 74, and the σ_{54} from ref. 75. Logos were drawn using Logomaker[76]. We then repeated the analysis described in the subsection: *association between the gain and loss of −10/−35 boxes and fluorescence changes*.

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 ($1 \le i \le 150$), and the fluorescence score f for daughter sequences of a given parent. To calculate the mutual information for each parent sequence, we used Eq. (2) as previously described in ref. 31:

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

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, and having a fluorescence score of 1, 2, 3, or 4 arb. units

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 $0.5 \times 0.5 = 0.25$. 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 Eq. (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 Eq. (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 Eq. (2), we used a previously described correction for finite sample sizes[31], which renders the final mutual information we computed equal to Eq. (3):

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

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 in Source Data.

Kolmogorov–Smirnov tests

We used a Kolmogorov–Smirnov (KS) test to compare 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. We then compared the actual distributions of top or bottom promoter signatures with their respective null distributions using the kstest function from scipy.stats.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sort-seq data generated in this study have been deposited in the Sequence Read Archive (SRA) database under accession code PRJNA1021969: [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1021969]. The processed data are available at [https://github.com/tfuqua95/promoter_emergence_mobile_DNA]. The data generated in this study are also provided in the **Supplementary Information**. The **Source data** are provided as a Source Data file. File name: Supplementary Data 1. Description: An Excel spreadsheet with a list of primers and DNA sequences to recreate all of the constructs tested in this manuscript. File name: Supplementary Data 2. Description: A large csv file with each unique daughter sequence, its respective parent sequence, and the GFP and RFP fluorescence scores from the sort-seq experiment. File name: Supplementary Data 3. Description: Contains a large 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. Supplementary Data files are also available on the Github repository: https://github.com/tfuqua95/promoter_emergence_mobile_DNA Source data are provided with this paper.

Code availability

Python scripts, an anaconda environment with relevant packages and their versions, and supplementary data files can be found on Github: https://github.com/tfuqua95/promoter_emergence_mobile_DNA.

References

1. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).
2. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
3. Sekine, Y., Izumi, K., Mizuno, T. & Ohtsubo, E. Inhibition of transpositional recombination by OrfA and OrfB proteins encoded by insertion sequence IS3. *Genes Cells* **2**, 547–557 (1997).
4. Chandler, M. & Fayet, O. Translational frameshifting in the control of transposition in bacteria. *Mol. Microbiol.* **7**, 497–503 (1993).
5. Sekine, Y., Eisaki, N. & Ohtsubo, E. Translational control in production of transposase and in transposition of insertion sequence IS3. *J. Mol. Biol.* **235**, 1406–1420 (1994).
6. Pribnow, D. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl Acad. Sci. USA* **72**, 784–788 (1975).
7. van Hijum, S. A. F. T., Medema, M. H. & Kuipers, O. P. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol. Mol. Biol. Rev.* **73**, 481–509 (2009).
8. Lagator, M. et al. Predicting bacterial promoter function and evolution from random sequences. *eLife* **11**, e64543 (2022).
9. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
10. Mauri, M. & Klumpp, S. A model for sigma factor competition in bacterial cells. *PLOS Computational Biol.* **10**, e1003845 (2014).
11. Ton-Hoang, B., Bétermier, M., Polard, P. & Chandler, M. Assembly of a strong promoter following IS911 circularization and the role of circles in transposition. *EMBO J.* **16**, 3357–3371 (1997).
12. Lewis, L. A. et al. The left end of IS2: a compromise between transpositional activity and an essential promoter function that regulates the transposition pathway. *J. Bacteriol.* **186**, 858–865 (2004).
13. Vandecraen, J., Chandler, M., Aertsen, A. & Van Houdt, R. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit. Rev. Microbiol.* **43**, 709–730 (2017).
14. 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).
15. uz-Zaman, M. H., D'Alton, S., Barrick, J. E. & Ochman, H. Promoter recruitment drives the emergence of proto-genes in a long-term evolution experiment with Escherichia coli. *PLOS Biol.* **22**, e3002418 (2024).
16. Charlier, D., Piette, J. & Glansdorff, N. IS3 can function as a mobile promoter in E. coli. *Nucleic Acids Res.* **10**, 5935–5948 (1982).
17. Safi, H. et al. IS6110 functions as a mobile, monocyte-activated promoter in Mycobacterium tuberculosis. *Mol. Microbiol.* **52**, 999–1012 (2004).
18. Prentki, P., Teter, B., Chandler, M. & Galas, D. J. Functional promoters created by the insertion of transposable element IS1. *J. Mol. Biol.* **191**, 383–393 (1986).

19. Simpson, A. E., Skurray, R. A. & Firth, N. An IS257-derived hybrid promoter directs transcription of a tetA(K) tetracycline resistance gene in the Staphylococcus aureus chromosomal mec region. *J. Bacteriol.* **182**, 3345–3352 (2000).
20. Maki, H. & Murakami, K. Formation of potent hybrid promoters of the mutant llm gene by IS256 transposition in methicillin-resistant Staphylococcus aureus. *J. Bacteriol.* **179**, 6944–6948 (1997).
21. Hinton, D. M. & Musso, R. E. Specific in vitro transcription of the insertion sequence IS2. *J. Mol. Biol.* **169**, 53–81 (1983).
22. Consuegra, J. et al. Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. *Nat. Commun.* **12**, 980 (2021).
23. Kopkowski, P. W., Zhang, Z. & Saier, M. H. The effect of DNA-binding proteins on insertion sequence element transposition upstream of the bgl operon in Escherichia coli. *Front. Microbiol.* **15**, 1388522 (2024).
24. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1530 (2018).
25. Wolf, L., Silander, O. K. & van Nimwegen, E. Expression noise facilitates the evolution of gene regulation. *eLife* **4**, e05856 (2015).
26. Warman, E. A., Singh, S. S., Gubieda, A. G. & Grainger, D. C. A non-canonical promoter element drives spurious transcription of horizontally acquired bacterial genes. *Nucleic Acids Res.* **48**, 4891–4901 (2020).
27. Fuqua, T., Sun, Y. & Wagner, A. The emergence and evolution of gene expression in genome regions replete with regulatory motifs. *eLife* **13**, RP98654 (2024).
28. 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).
29. 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).
30. Ireland, W. T. et al. Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *eLife* **9**, e55308 (2020).
31. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl Acad. Sci.* **107**, 9158–9163 (2010).
32. Belliveau, N. M. et al. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc. Natl Acad. Sci.* **115**, E4796–E4805 (2018).
33. 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 Biol.* **15**, e1006226 (2019).
34. 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).
35. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
36. Mitchell, J. E., Zheng, D., Busby, S. J. W. & Minchin, S. D. Identification and analysis of 'extended –10' promoters in Escherichia coli. *Nucleic Acids Res.* **31**, 4689–4695 (2003).
37. Zerbib, D. et al. Functional organization of the ends of IS1: specific binding site for an IS 1-encoded protein. *Mol. Microbiol.* **4**, 1477–1486 (1990).
38. Baba, T. et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
39. Weinreich, M. D. & Reznikoff, W. S. Fis plays a role in Tn5 and IS50 transposition. *J. Bacteriol.* **174**, 4530–4537 (1992).
40. Pósfai, G. et al. Emergent properties of reduced-genome Escherichia coli. *Science* **312**, 1044–1046 (2006).
41. 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).
42. Machida, C. & Machida, Y. Regulation of IS1 transposition by the *insA* gene product. *J. Mol. Biol.* **208**, 567–574 (1989).
43. Larson, M. H. et al. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344**, 1042–1047 (2014).
44. Di Salvo, M., Puccio, S., Peano, C., Lacour, S. & Alifano, P. Rho-TermPredict: an algorithm for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases. *BMC Bioinforma.* **20**, 117 (2019).
45. Ray-Soni, A., Bellecourt, M. J. & Landick, R. Mechanisms of bacterial transcription termination: all good things must end. *Annu. Rev. Biochem.* **85**, 319–347 (2016).
46. Yarulin, V. R. & Gorlenko, Z. M. Effect of mutations in the RNA polymerase gene and that of the transcription termination factor rho on expression of the Escherichia coli galactose operon with an IS2 polar insertion. *Mol. Gen. Genet.* **201**, 344–346 (1985).
47. Hübner, P., Iida, S. & Arber, W. A transcriptional terminator sequence in the prokaryotic transposable element IS1. *Mol. Gen. Genet.* **206**, 485–490 (1987).
48. 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).
49. Glansdorff, N., Charlier, D. & Zafarullah, M. Activation of gene expression by IS2 and IS3. *Cold Spring Harb. Symp. Quant. Biol.* **45**, 153–156 (1981).
50. 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).
51. 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).
52. 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. *Int. J. Antimicrobial agents* **41**, 122–125 (2013).
53. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLOS Genet.* **9**, e1003504 (2013).
54. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
55. 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).
56. Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
57. Bejerano, G. et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
58. 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).
59. 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).
60. 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).
61. 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).

62. Hu, S. T. et al. Functional analysis of the 14 kDa protein of insertion sequence 2. *J. Mol. Biol.* **236**, 503–513 (1994).
63. Todd, C. D., Deniz, Ö., Taylor, D. & Branco, M. R. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *eLife* **8**, e44344 (2019).
64. 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. *J. Bacteriol.* **188**, 6506–6514 (2006).
65. 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).
66. Waskom, M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
67. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
68. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
69. Tierrafría, V. H. et al. RegulonDB 11.0: comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microb. Genomics* **8**, 000833 (2022).
70. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
71. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
72. Gunesekere, I. C. et al. Comparison of the RpoH-dependent regulon and general stress response in Neisseria gonorrhoeae. *J. Bacteriol.* **188**, 4769–4776 (2006).
73. Raman, S. et al. The alternative sigma factor SigH regulates major components of oxidative and heat stress responses in Mycobacterium tuberculosis. *J. Bacteriol.* **183**, 6119–6125 (2001).
74. Yu, H. H. Y., Di Russo, E. G., Rounds, M. A. & Tan, M. Mutational analysis of the promoter recognized by Chlamydia and Escherichia coli sigma(28) RNA polymerase. *J. Bacteriol.* **188**, 5524–5531 (2006).
75. Samuels, D. J. et al. Use of a promiscuous, constitutively-active bacterial enhancer-binding protein to define the σ^{54} (RpoN) regulon of Salmonella Typhimurium LT2. *BMC Genomics* **14**, 602 (2013).
76. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

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). T.F. is supported by a postdoctoral fellowship from the European Molecular Biology Organization (ALTF 963-2021) and a University of Zurich Postdoc Grant (FK-23-120). We thank all members of the Wagner group, Philipp Schätzle and Mario Wickert from the UZH cytometry facility, and Baxter for enforcing a work-life balance.

Author contributions

Conceptualization: T.F. and A.W., Methodology: T.F., Investigation T.F. and A.W., Visualization: T.F., Supervision: A.W., Writing: T.F. and A.W.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-60023-w.

Correspondence and requests for materials should be addressed to Andreas Wagner.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.