Unravel the start element and promoter architecture across the domain Bacteria

Authors: Syue-Ting Kuo[1]†, Joshua Kevin Chang[1]†, Clara Chang[2]†‡, Wei-Yi Shen[1], Christine Hsu[1]§, Sheng-Wen Lai[1], Hsin-Hung David Chou[1,2]*

Affiliations:

[1]Department of Life Science, National Taiwan University; Taipei, 106319, Taiwan.

[2]Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University; Taipei, 106319, Taiwan.

†These authors contributed equally to this work.

‡Present address: Department of Biomedical Engineering, University of California Davis; Davis, 95616, USA.

§Present address: Department of Molecular and Cell Biology, University of California San Diego; San Diego, 92122, USA.

*Corresponding author. Email: chouhh@ntu.edu.tw

Short Title: Promoter architecture across the domain Bacteria

Abstract: Promoters comprise multiple elements to modulate transcription. Despite decades of research, significant sequence and length variation of promoter elements renders elucidation of promoter function and the evolutionary diversity of transcriptional regulation challenging. Combining massively parallel assays with biophysical modeling, we mapped each promoter element across 49 phylogenetically diverse bacterial genomes (GC%: 27.8-72.1). We discovered a conserved promoter element, 'start,' that dictates transcription start site selection. We uncovered broad sequence conservation in the underappreciated spacer element, affecting transcription by up to 600-fold. We showed the discriminator element conserved in major bacterial clades Terrabacteria but variable in Gracilicutes. High sequence variation in Gracilicutes' discriminator element resulted from diversifying evolution, enabling coordination of global gene expression with cellular growth rate through differential regulation of individual promoters. Identifying sequence and functional similarity between bacterial promoter elements and their archaeal and eukaryotic counterparts strongly suggests a shared evolutionary origin of promoter architecture.

Introduction:

Genomes use just four nucleobases to encode myriads of genes and regulatory sequences. Deciphering this language of life is crucial — not only for understanding how organisms survive today but also for tracing the evolutionary history on Earth and projecting their future

trajectories. Identifying genes' coding sequences is straightforward because they consist of codons, adhere to the reading frame, and show evident homology across species. In contrast, regulatory sequences, particularly core promoters, comprise multiple elements varying in length and sequence composition, making them difficult to recognize, especially in non-model organisms[3].

In bacteria, core promoters contain UP, -35, spacer, extended -10 (Ex), -10, and discriminator (Dis) elements in the 5′-to-3′ order (Fig. 1a)[1]. Transcription begins with RNA polymerase (RNAP) binding to the -35 element via the interchangeable σ subunit, followed by unwinding of DNA at the -10 element to synthesize RNA from the downstream transcription start site (TSS). The -35 and -10 elements are both hexamers showing consensus sequences TTGACA and TATAAT, respectively. The spacer element, ranging from 15 to 19 bp, exhibits minimal sequence features and limited RNAP contact, with 17 bp as the preferred length for transcription[4,5]. The Ex element refers to a TG motif (or TRTG, commonly seen in Firmicutes; R = purine), appearing 2 bp upstream of the -10 element within the spacer element[6]. The UP element, bound by the C-terminal domain of the RNAP α subunit, displays alternating A- and T-tracts and lacks a defined length[7]. The Dis element, lying downstream of the -10 element, interacts with RNAP σ, β, and β′ subunits and shows high sequence variation[8,9]. Promoter evolution is shaped by multiple factors, including the corresponding σ subunit, genomic GC content, selection for gene expression, interaction with transcription factors, and functional complementation among promoter elements[1,10,11]. Even after half a century of research, the high variability among bacterial promoters continues to challenge bioinformatics prediction, despite being more compact and architecturally better defined than eukaryotic promoters[12–14]. The stagnant progress has prompted researchers to call for a revision of the bacterial promoter concept[1].

Changes in gene expression are critical to adaptive evolution[15], but limited knowledge of the regulatory syntax has hindered our endeavor to unravel the underlying mechanisms. Advancing this field requires precisely segmenting and aligning promoter elements for characterization. Since the -35 and -10 elements are relatively conserved and separate variable elements in bacterial promoters, we propose building an accurate -35 and -10 prediction model as the solution (Fig. 1a). We consider this approach feasible because residues on the primary σ subunit (σ^{70}) that contact the -35 and -10 elements are conserved across the domain Bacteria, although the whole σ^{70} subunit has diverged significantly (Extended Data Fig. 1). To achieve that, we quantified transcription of all possible -35 and -10 sequence variants in a synthetic promoter library (4^{12} = 16,777,216; PL library) using sort-seq, which employed green fluorescent protein (GFP) expression to report promoter strength in *Escherichia coli* (Extended Data Fig. 2a-d)[16]. From this dataset, we built a biophysical model that inferred promoter strength based on the influence of the -35 and -10 sequences and the spacer length on RNAP-promoter binding energy ($-\Delta\varepsilon_{RNAP}$, Fig. 1a and Extended Data Fig. 2e-g)[5,17]. We then applied this model to predict promoter architecture in 49 phylogenetically distinct bacteria (genome GC%: 27.8-72.1; Fig. 2a; Supplementary Table 1).

Results:

Model predicts promoter architecture across diverse bacteria

Under the constraint of a 15- to 19-bp spacer length, our model scanned for potential -35 and -10 elements that maximized $-\Delta\varepsilon_{RNAP}$ within the promoter region preceding each experimentally

mapped TSS (Fig. 1a). To evaluate the model's power in discerning each species' promoter structure, we applied it to random genomic sequences and promoter-shuffled sequences, which retained base frequencies but not order. Differences between the $-\Delta\varepsilon_{RNAP}$ distributions of these sequences and native promoters were quantified by Kolmogorov-Smirnov statistics. Overall, the model distinguished native promoters from random genomic sequences (D_g) better than from promoter-shuffled sequences (D_s, Fig. 1b and Extended Data Fig. 3a). A cutoff of D_g, $D_s < 0.2$ was applied to exclude *Phytoplasma sp.* (Ph.sp), *Pseudomonas putida* (Ps.pu), and *Shigella flexneri* (Sh.fl) from further analysis (marked red in Figs. 1, 2, and 4). Poor model prediction for these species likely stemmed from the quality and quantity of TSS mapping (Supplementary Discussion).

Trained with empirical data from *E. coli*, model-predicted -35, -10, and spacer elements displayed the canonical sequence and length characteristics across species (Extended Data Figs. 4b and 3b). Yet three properties, unique to native promoters and unsought by the model, confirmed its reliability and revealed the broad conservation of bacterial promoter architecture: (1) the occurrence of the TG and TRTG motifs aligned with the Ex location (Fig. 1c and Extended Data Fig. 3c); (2) TSS was centered 7 bp downstream of model-predicted -10 elements (Fig. 1d and Extended Data Fig. 3d), consistent with empirical findings in *E. coli*[18]; (3) $-\Delta\varepsilon_{RNAP}$ increased when the model predicted -10 elements to be 7 bp upstream of TSS (Fig. 1e and Extended Data Fig. 3d). The distance between TSS and model-predicted -10 elements appeared more variable (standard deviations > 3.6 bp) in *Nanosynbacter lyticus* (Na.ly), *Mycoplasma pneumoniae* (My.pn), and *Lachnoclostridium phytofermentans* (La.ph) (marked blue in Figs. 1, 2, and 4), which may reflect these species' characteristics or methodological differences in the TSS-mapping accuracy (Supplementary Table 1).

Identifying a broadly conserved promoter element — start

Model prediction of the -35 and -10 elements allowed us to separately align each promoter element to reveal sequence enrichment, defined as the log-odds ratios of base frequencies relative to genomic backgrounds (Fig. 2b, Extended Data Fig. 4, and Supplementary Data 1). We chose this measure over the information content because it enhanced signal contrast by offsetting genomic composition bias. Intriguingly, aligning TSS unveiled a conserved pyrimidine-purine-pyrimidine (YRY) motif centered on TSS. Among all 64 triplet bases, YRY was more frequent than the others in most bacteria examined (Fig. 2a). Taking the genomic composition and purine bias of bacterial TSS into account[19], YRY enrichment remained statistically significant (one-tailed binominal test: $P < 0.05$). Besides YRY, we noted the second abundant pyrimidine-purine-purine (YRR) motif overrepresented in clade Terrabacteria, particularly *Thermotoga maritima* (Th.ma), indicating lineage-specific divergence. In contrast, three species, *Geobacter sulfurreducens* (Ge.su), *Agrobacterium tumefaciens* (Ag.tu), and *Neisseria gonorrhoeae* (Ne.go), were enriched for the TTT motif and lacked TSS purine bias. Given their co-localization, could the 3-bp motifs influence TSS selection?

To address this, we analyzed data from another study that employed saturation mutagenesis and RNA-seq to assess how sequence variation 4 to 10 bp downstream of the -10 element affected TSS positioning by *E. coli* RNAP (Fig. 3a)[18]. By grouping all TSS based on their distance from the -10 element, we confirmed YRY enrichment and co-localization with TSS (Fig. 3b). Notably, co-localization occurred at 93.1% and 79.0% when the YRY motif was centered 7 or 8 bp downstream of the -10 element, respectively (Fig. 3c). The YRR motif had a similar but less

129 prominent effect on TSS selection (86.2% and 67.4%, respectively). Promoter strength increased
130 significantly when TSS occurred 7 or 8 bp downstream of the -10 element, increasing further
131 when YRY or YRR motifs co-localized (two-tailed Mann-Whitney U test: $P < 10^{-190}$; Fig. 3d).
132 This indicates that positive selection may have driven the centralized TSS distribution and
133 prevalence of YRY and YRR motifs across diverse bacteria (Figs. 1d and 2a). In view of its
134 functional significance and broad conservation, we designated the 3-bp region centered on TSS
135 as a new promoter element, 'start.'

136

137 **Variable promoter elements show broad conservation**

138 Beyond the start element, we inspected enriched features in the other promoter elements. The -35
139 and -10 elements in most species were enriched for the consensus sequences (Fig. 2b and
140 Extended Data Fig. 4b). In contrast, Alphaproteobacteria, including *Novosphingobium*
141 *aromaticivorans* (No.ar), showed reduced -10 sequence conservation. The -10 element of
142 *Bacteroides thetaiotaomicron* (Ba.th) diverged even further, exhibiting a TATCTT hexamer
143 followed by a TGC trimer extending into the Dis element. Both findings are consistent with
144 previous reports[20,21]. The -35 element of *B. thetaiotaomicron* also evolved distinctly (TTGCTA),
145 reflecting the significant divergence of its primary σ subunit (Extended Data Fig. 1)[22].
146 Recapitulating these deviations underscores the robustness of our model predictions.

147 In most species, the UP element exhibited alternating A- and T-tracts, typically spanning 40 bp
148 but extending up to 100 bp in Gammaproteobacteria (e.g., *E. coli* (Es.co); Fig. 2b and Extended
149 Data Fig. 4a). Additionally, the 3′ end of the UP element was often enriched for cytosine,
150 suggesting its contact with R554 and R586 of σ^{70} is functionally important (Extended Data Fig.
151 1a)[23]. Regarding the spacer element, our model predicted its length to range mostly 16-18 bp,
152 with 17 bp predominant across species (Extended Data Fig. 3b). Despite considerable sequence
153 variation, the spacer element can be subdivided into four distinct regions: from 5′ to 3′, an AT-
154 tract (mainly thymine), a low-signal region with slight GC enrichment (e.g., *N. aromaticivorans*
155 and *Rhodobacter sphaeroides* (Rh.sp)), another AT-tract, and a G-rich region including the Ex
156 element (Fig. 2b and Extended Data Fig. 4c). Similar to the start element, the Dis element and
157 initial transcribed region (ITR) of mRNA (defined as the regions directly upstream and
158 downstream of the start element, respectively, to avoid overlap) displayed lineage specificity.
159 The Dis element was AG- or AT-rich in Terrabacteria but highly variable in Gracilicutes (Fig. 2b
160 and Extended Data Fig. 4d). ITR was AG-rich in Firmicutes (e.g., *Bacillus subtilis* (Ba.su)) and
161 *T. maritima*, AT-rich in Actinobacteriota (e.g., *Streptomyces coelicolor* (St.co)), and frequently
162 contained a 5′ T-tract in Gracilicutes (e.g., *Helicobacter pylori* (He.py); Fig. 2b and Extended
163 Data Fig. 4e).

164 To systematically assess the conservation of promoter elements and ITR within and between
165 genomes across species, we applied two quantitative measures. Intra-genome conservation was
166 estimated using Kullback-Leibler (KL) divergence, which gauged the information contrast
167 between a subject and its genomic background. As expected, the -10 and -35 elements ranked as
168 the top two conserved elements across species (Fig. 4a). Notably, the third place went to the start
169 element, reaffirming its functional significance. Inter-genome conservation was quantified as the
170 similarity of a subject's sequence enrichment between two species, defined by the Spearman's
171 correlation of enrichment scores (Fig. 4b and Extended Data Fig. 5). Consistent with
172 aforementioned observations, the -35 and -10 elements were highly conserved across species
173 except *B. thetaiotaomicron*. The UP, spacer, and Ex elements were moderately conserved except

174 *Synechocystis* sp. (Sy.sp), *R. sphaeroides, Leptospira interrogans* (Le.in), *Mycoplasma*
175 *pneumoniae* (My.pn), and *B. thetaiotaomicron*. On the other hand, the downstream promoter
176 elements, Dis and start, and ITR showed more incidences of clade-specific variation, with the
177 former being more conserved within Terrabacteria than Gracilicutes. Moreover, the promoter
178 elements and ITR of Firmicutes and Actinobacteriota were more conserved within the lineage
179 than the other bacterial clades. Interestingly, promoter-contacting residues on RNAP remain
180 broadly conserved, while σ^{70}, β, and β′ subunits show clade-specific divergence at the whole-
181 protein level (Extended Data Figs. 1a,b and 6a,b). This suggests that RNAP, beyond the
182 promoter-contacting sites, may contribute to clade-distinctive traits in promoter elements and
183 ITR.

184

185 **Positive selection shapes broad spacer sequence conservation**

186 To assess the importance of sequence conservation in the spacer element, we performed sort-seq
187 to map its sequence-function landscapes in *E. coli*. We synthesized saturation mutagenesis
188 libraries, SL16, SL17, and SL18, for 16-, 17-, and 18-bp spacer elements, respectively, paired
189 with distinct -35 and -10 elements to reduce context dependence (Fig. 5a). From 3,855,981,
190 6,984,696, and 12,242,605 detected variants, we retained reliable measurements, divided each
191 library into five groups based on promoter strength, and computed each group's sequence
192 enrichment relative to the whole library (Fig. 5b and Extended Data Fig. 7). Despite differences
193 in spacer length and promoter context, the three libraries revealed parallel characteristics from
194 low to high promoter strength (Groups 1 to 5 in Fig. 5c): (1) the spacer element contained four
195 distinct regions, mirroring those observed across species (Fig. 2b); and (2) enriched sequences in
196 Group 5 resembled those in natural spacer elements, except for an apparent lack of the Ex
197 element's TG motif. Notably, the TG motif frequency increased steeply with promoter strength
198 (Fig. 5d), indicating its positive contribution to transcription. Underrepresentation of the TG
199 motif was therefore an artifact of enrichment analysis, masked by the positive enrichment of
200 guanine.

201 To verify the function of enriched sequences, we constructed three promoter sets for SL16,
202 SL17, and SL18, each comprising eight combinations based on three two-state characters:
203 enriched spacer sequence (Group 1 or 5), -35 and -10 element pair (library context or consensus
204 sequence), and TG motif (absence or presence). Relative to the basal promoter (-35/-10: library
205 context, spacer: Group 1, TG: absence), switching each character's state increased promoter
206 strength across all three sets (Fig. 5e). Remarkably, altering the spacer state elevated expression
207 by up to 600-fold. Nevertheless, simultaneously switching all characters to their expression-
208 promoting states conferred only moderate rather than the highest expression level. Confirmed by
209 both GFP and mRNA expression (Extended Data Fig. 8), these findings indicate that positive
210 selection shapes the broad spacer sequence conservation in Nature and support a previously
211 postulated convex relationship between RNAP binding affinity and transcription initiation[24].

212

213 **Diversifying evolution of the Dis element driven by growth rate regulation**

214 To probe the function of variable Dis elements in Gracilicutes, we performed saturation
215 mutagenesis and sort-seq to characterize the 8-bp region directly downstream of the -10 element
216 in *E. coli* (Fig. 6a). We classified variants in this DL library, where promoter strength varied up
217 to 240-fold, into five groups to examine their sequence-function relationship (Fig. 6c). From low

218 to high promoter strength (Groups 1 to 5), enriched features shifted from GC-rich to C-poor
219 sequences, with the 5` 3-bp region showing stronger enrichment (Fig. 6d). The GC-rich pattern
220 mirrored the 8-bp Dis element in the ribosomal promoter *rrnB* P1 ($P1_{rrnB}$; Fig. 6b)[25].
221 Paradoxically, the exact sequence was among the lowest-expression variants in the DL library
222 (24th in 58,065; Fig. 6c), while $P1_{rrnB}$ was among the strongest promoters in *E. coli*[26].

223 Why would a strong promoter bear a strength-compromising component? A prior work
224 demonstrates that the GC-rich Dis element made $P1_{rrnB}$ transcription scale positively with growth
225 rate[25]. This growth rate-dependent regulation results from increasing production of the stringent
226 response signal ppGpp under slow growth, which allosterically regulates RNAP to reduce $P1_{rrnB}$
227 transcription[27–29]. A study further reports changes in 35.4% of the *E. coli* transcriptome due to
228 exogenous ppGpp overproduction (Fig. 6e) — genes involved in translation and nucleotide
229 metabolism reduced transcription significantly, whereas those in amino acid metabolism and
230 signal response strongly increased transcription[8]. Unexpectedly, Dis sequences in native
231 promoters showing ppGpp-induced decreased to increased transcription (Groups 1 to 5 in Fig.
232 6f) resembled the low- to high-expression variants in the DL library (Groups 1 to 5 in Fig. 6d),
233 although all DL variants were assayed under the same promoter background and growth
234 condition. Examining the base composition of DL variants and native Dis elements across groups
235 further highlighted their similar design principles (Fig. 6g,h).

236 Based on above findings, we hypothesized that high Dis variation within Gracilicutes arose in
237 part from diversifying evolution, enabling cells to fine-tune individual promoters to meet global
238 cellular demands across growth rates. If so, we would expect: (1) growth rate dependency
239 conferred by Dis elements to be modular; (2) ppGpp-contacting residues on RNAP to be
240 conserved; and (3) genes subjected to growth rate-dependent regulation to show sequence
241 conservation in their Dis elements. To test modularity, we selected eight DL variants from low-
242 to high-expression and examined their effects on promoter strength in the $P1_{rrnB}$ and DL library
243 promoter contexts (P_{DL}) in *E. coli* across growth rates (0.3-1.7 hr^{-1}; Fig. 6i). Despite the distinct
244 expression ranges of P_{DL} and $P1_{rrnB}$, seven of eight DL variants exerted similar effects on
245 expression and growth rate dependency in both promoter backgrounds. Moreover, low- and high-
246 expression DL variants consistently led to positive and negative correlations, respectively,
247 between promoter strength and growth rate. Next, we examined ppGpp-contacting residues on
248 RNAP and found higher conservation in Gracilicutes than Terrabacteria, except for
249 Campylobacterota (Ca in Extended Data Fig. 6c).

250 Finally, we examined Dis sequences associated with genes in translation, nucleotide metabolism,
251 amino acid metabolism, and signal response, based on the functional categorization of the KEGG
252 Orthology database[33]. Compared with genome-wide Dis sequences, those of the four gene
253 categories in Gracilicutes shifted from being overall GC-rich to diverse and cytosine-reduced
254 (Fig. 6j), echoing *E. coli* Dis design principles (Fig. 6g,h). Dis sequences in Terrabacteria were
255 AG- or AT-rich regardless of gene categories (Fig. 6k and Extended Data Fig. 9a-e). Examining
256 the similarity of Dis sequences across gene categories, defined by the Spearman's correlation of
257 enrichment scores, further revealed the higher conservation of translation genes' Dis sequences
258 and the contrasting Dis features between translation and signal response genes in Gracilicutes
259 (Fig. 6l and Extended Data Fig. 9f,g). Such gene-specific divergence, in contrast, was not
260 apparent among Dis sequences in Terrabacteria. Collectively, these results support the
261 significance of growth rate-dependent regulation in shaping the diverse Dis sequence
262 composition in Gracilicutes.

263

Discussion:

Conventionally, promoters and other regulatory elements are defined by specific motifs. Although intuitive, the motif concept is limited by the architectural and sequence variability of regulatory elements. To address this, we took an integrative approach that combined massively parallel assays, biophysical modeling, comparative genomics, and experimental validation to decipher the regulatory syntax across diverse bacteria. Notably, revealing broad sequence conservation in 'poorly conserved' promoter elements will facilitate promoter engineering beyond model organisms, support development of an automated promoter annotation pipeline for genomes, and open new avenues for probing the underlying molecular basis and selective pressures. In light of evolution, uncovering the distinctive start element preference, the gene-specific sequence enrichment in the Dis element, and its role in growth rate-dependent regulation supports the ancient divergence of the two major bacterial clades, Gracilicutes and Terrabacteria[30]. Extending to the cross-domain context, identification of sequence and functional similarities between bacterial promoter elements, start (YRY) and -10 (TATAAT), and their archaeal and eukaryotic counterparts, initiator (YR) and TATA box (TATAAAA), strongly suggests a common evolutionary origin of promoter architecture across the three domains of life, despite significant divergence of the transcription machinery[2,31].

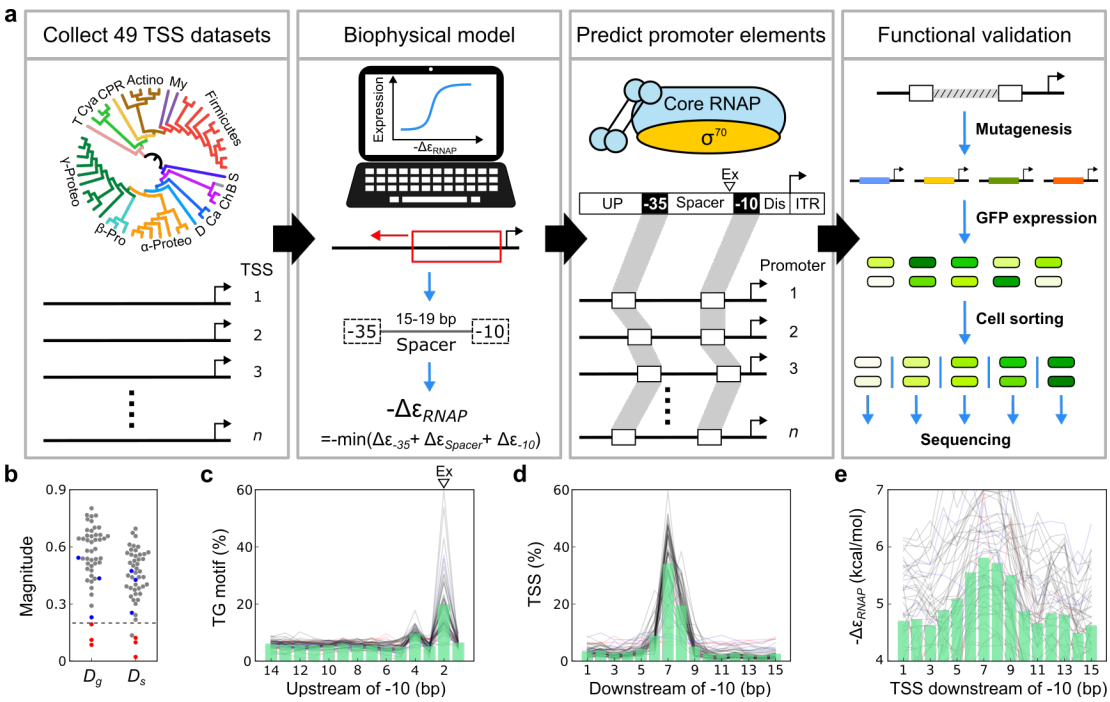

Fig. 1 | Applying model prediction and experiments to dissect bacterial promoters. **a**, Study overview. **b**, Differences between the distribution of model-predicted RNAP binding energy ($-\Delta\varepsilon_{RNAP}$) for native promoters and those for random genomic (D_g) and promoter-shuffled sequences (D_s) are measured as Kolmogorov-Smirnov statistics for each species. The dashed line indicates a 0.2 magnitude cutoff. **c**, Sliding window analysis shows the frequency and the 3′ end location of the TG motif relative to the -10 element. **d**, Distribution of TSS relative to the -10 element. **e**, Correspondence between $-\Delta\varepsilon_{RNAP}$ and TSS relative to the -10 element. In **b-e**, species having D_g, $D_s < 0.2$ or greater distance variation between TSS and the -10 element (standard deviations > 3.6 bp) are marked in red and blue, respectively. In **c-e**, individual species are indicated by curves, while histograms show the mean distributions of the 49 species.

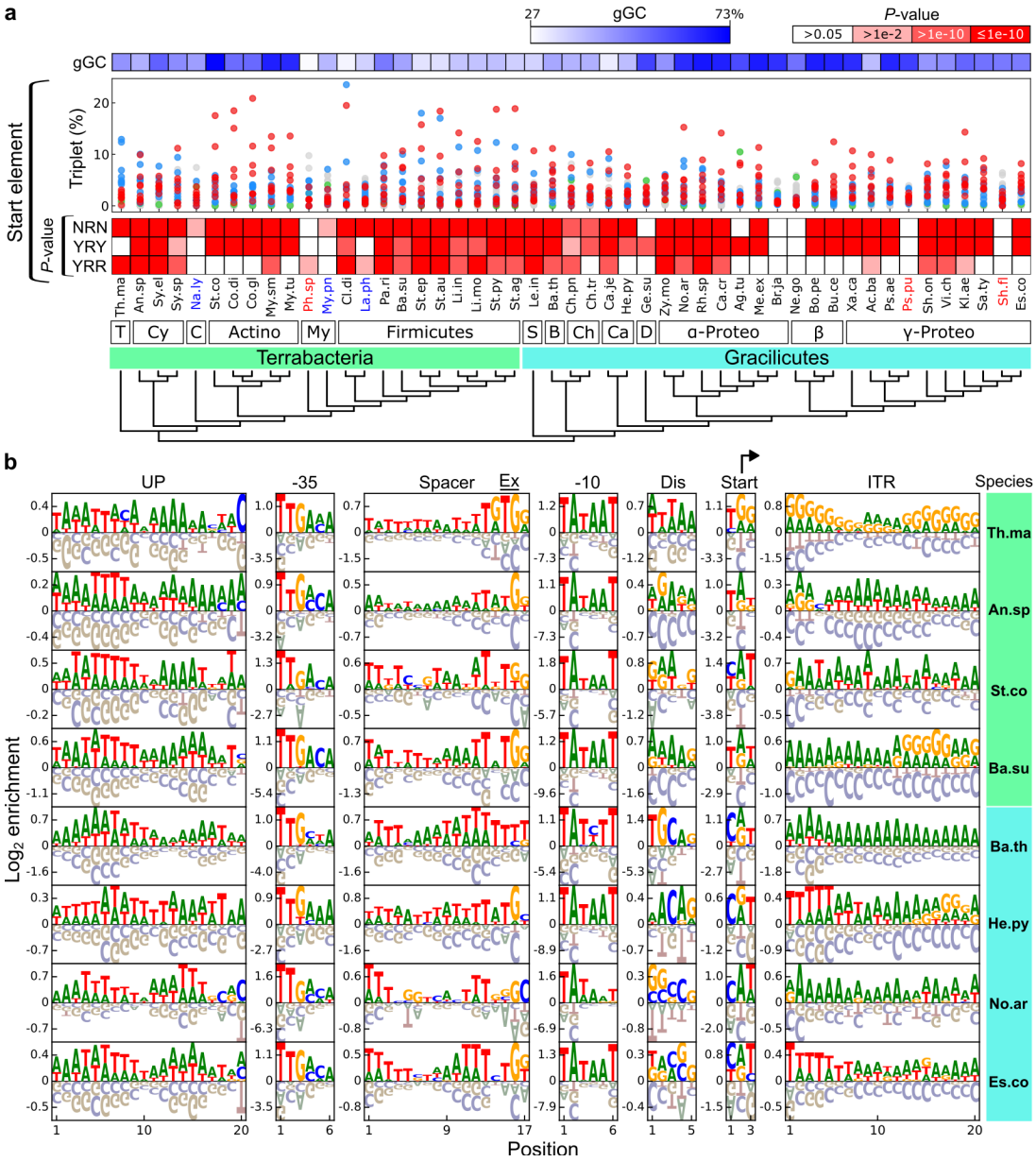

Fig. 2 | Characteristics of TSS, promoter elements, ITR, and GC content.
a, Top to bottom: genomic GC content (gGC), proportions of all 64 triplets in the start element, statistical significances (*P*-value) of the TSS purine bias (i.e., NRN motif) and YRY and YRR motif frequencies conditioned upon TSS being purine, and the phylogenetic relationship of 49 bacteria[30]. Color bars in the top right indicate heatmap scales. Triplets belonging to YRY, YRR, or TTT motifs are indicated by red, blue, and green dots, respectively. Species with D_g, $D_s < 0.2$ or greater -10-to-TSS distance variation are marked in red and blue, respectively. T: Thermotogota, Cy: Cyanobacteria, C: Candidate Phyla Radiation, Actino: Actinobacteriota, My: Mycoplasmatota, S: Spirochaetota, B: Bacteroidota, Ch: Chlamydiota, Ca: Campylobacterota, D: Desulfobacterota, α-Proteo: Alphaproteobacteria, β: Betaproteobacteria, γ-Proteo: Gammaproteobacteria. **b**, Sequence enrichment in selected species. TSS is marked by a bended arrow. Graphic representation of positive and negative enrichment is rescaled for uniform visualization. Full enrichment analysis is shown in Supplementary Fig. 4. In **a**,**b**, species are shown as abbreviations (i.e., first two letters of genus and species; Supplementary Table 1).

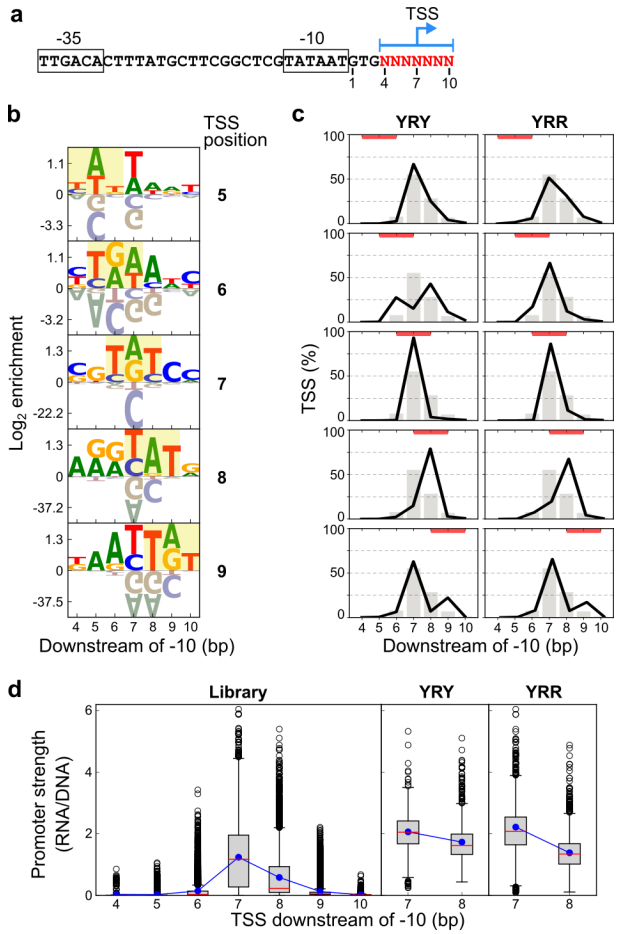

Fig. 3 | The start element influences TSS selection and promoter strength in *E. coli*. **a,** Promoter bearing a 7-bp randomized region (red) in a prior TSS mapping study[18]. **b,** Sequence enrichment in the 7-bp region for TSS occurring downstream of the -10 element. The 3-bp region centered on TSS is highlighted. Graphic representation for positive and negative enrichment is rescaled for uniformity. **c,** Influence of the location of YRY and YRR motifs (red bar) on TSS selection (black curve). The TSS distribution of the whole library (N = 16,212) is shown as the gray histogram. **d,** Library distribution of promoter strength with respect to the TSS location. Promoter strength is calculated as RNA reads divided by DNA reads (i.e., transcripts expressed by each promoter variant). The library subset, variants having YRY or YRR motifs co-localized with TSS = 7 or 8, are shown for comparison. Boxes, whiskers, blue dots, red lines, and empty circles indicate the interquartile range, 1.5-fold interquartile range, mean, median, and outliers, respectively, of the distribution.

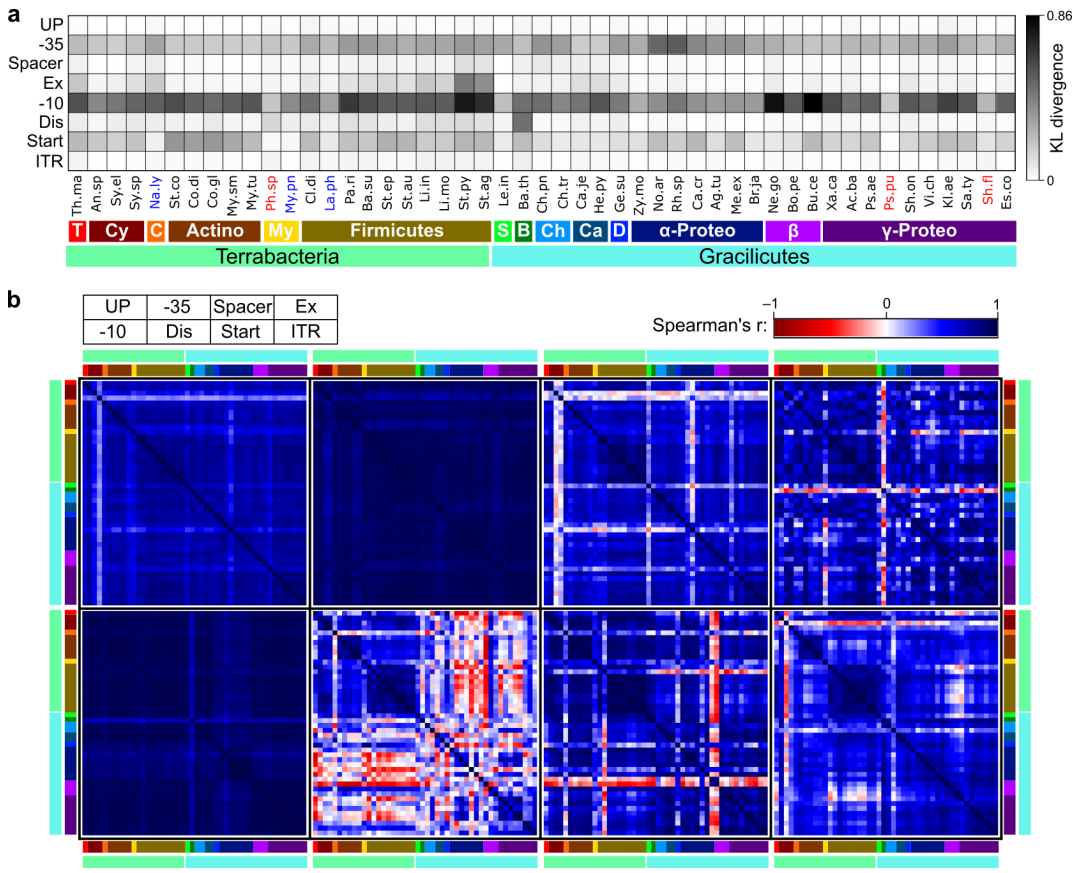

Fig. 4 | Conservation of promoter elements and ITR within and between genomes. **a**, Intra-genome conservation quantified by KL divergence per bp. **b**, Inter-genome conservation measured as the Spearman's correlation of enriched sequence features. Species having D_g, D_s < 0.2 (i.e., Sh.fl, Ps,pu, Ph.sp) are excluded. Lineages are distinguished by colors shown in **a**. In **a**-**b**, analyses consider the dominant length type of Dis elements (5-bp) and spacer elements (17-bp, excluding the 2-bp Ex region), the 40-bp UP element region upstream of the -35 element, and the 40-bp ITR region downstream of the start element. The abbreviations of species and lineage names follow Fig. 2 and Supplementary Table 1.

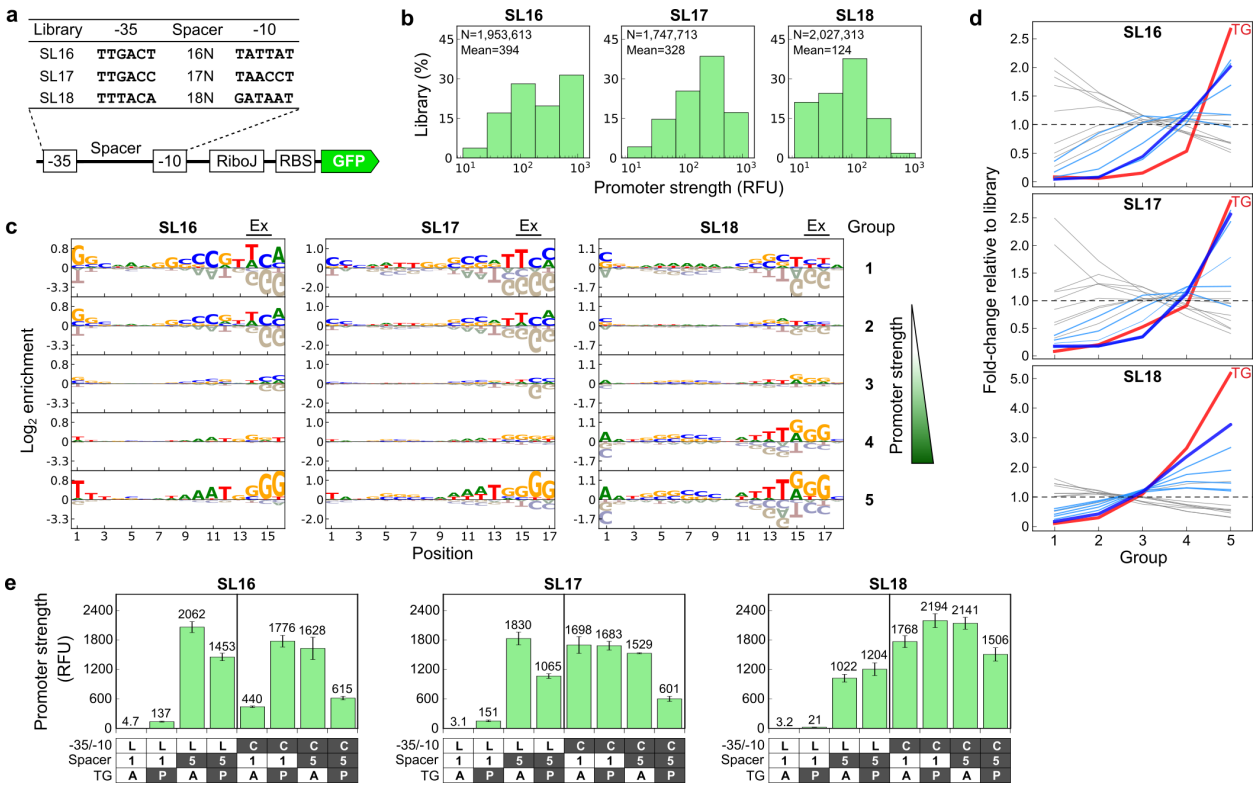

Fig. 5 | Determining the spacer sequence-function relationship.
a, Promoter backgrounds of the SL16, SL17, and SL18 libraries. RiboJ[32], self-cleaving ribozyme; RBS, ribosome binding site; GFP, green fluorescent protein. **b**, Library distribution of promoter strength in *E. coli*. RFU, relative fluorescence unit. **c**, Enriched sequences in the spacer element with respect to promoter strength. Spacer variants are grouped based on the binning in **b**. Graphic representation for positive and negative enrichment is rescaled for uniformity. **d**, Fold-changes in the frequencies of the 16 2-bp motifs in the Ex element relative to the library (dashed line) across groups. The trends of 'TG,' 'GG,' and the remaining G-bearing motifs are marked in red, deep blue, and light blue, respectively. **e,** Impacts of enriched spacer features on promoter strength across promoter contexts. Three promoter sets for SL16, SL17, and SL18 comprise eight combinations based on three two-state characters: enriched spacer sequence (Group 1 or 5, denoted as 1 and 5, respectively), -35 and -10 element pair (library context or consensus sequence, denoted as L and C, respectively), and TG motif (absence or presence, denoted as A and P, respectively). Results are shown as the means and standard deviations of three independent measurements.

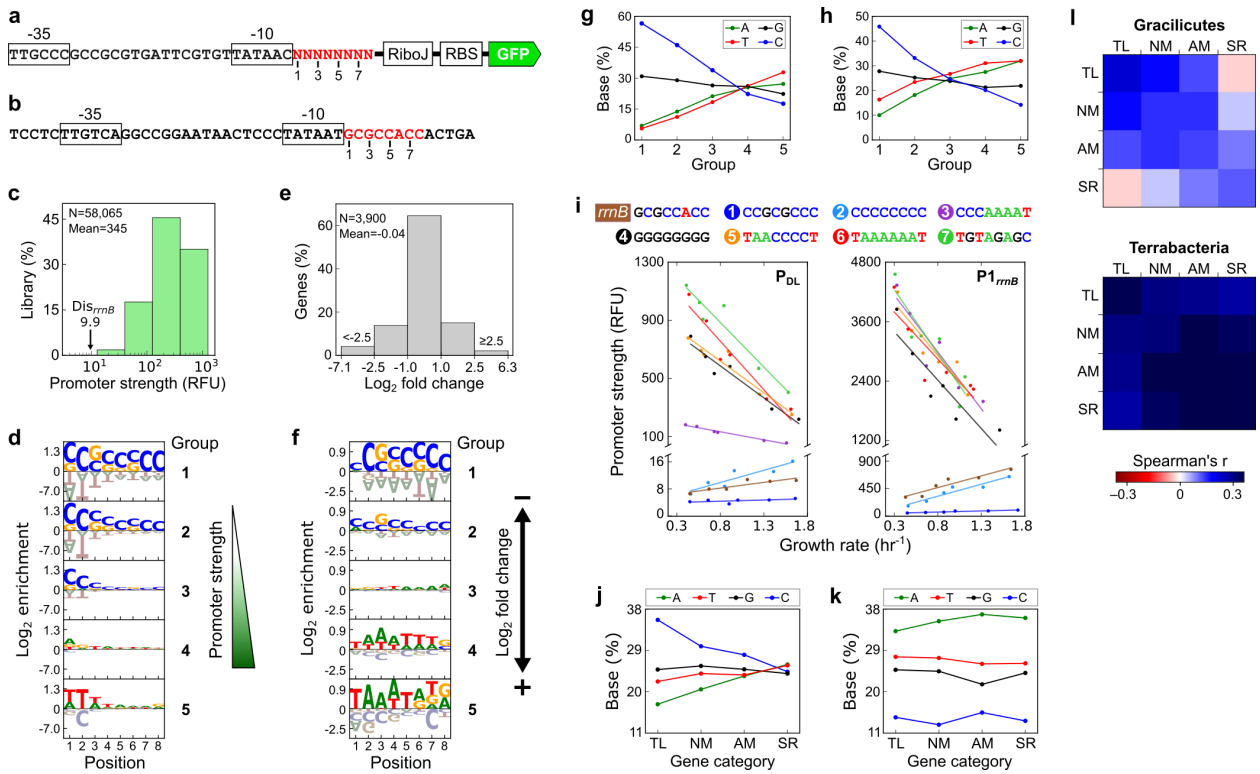

Fig. 6 | Determining the Dis sequence-function relationship.
a,**b**, DL library promoter background (**a**) and P1$_{rrnB}$ promoter (**b**), with the 8-bp Dis marked red. **c**, DL library distribution of promoter strength in *E. coli*. RFU, relative fluorescence unit. Dis$_{rrnB}$, variant bearing the P1$_{rrnB}$ Dis in **b**. **d**,**g**, Sequence enrichment (**d**) and base composition (**g**) of DL variants with respect to promoter strength. Variants are divided into five groups based on the binning in (**c**). **e**, Distribution of the mRNA fold-changes of *E. coli* genes in response to ppGpp elevation[8]. **f**,**h**, Sequence enrichment (**f**) and base composition (**h**) of *E. coli* native Dis with respect to ppGpp-induced transcriptional changes[8]. Native Dis is grouped based on the binning in (**e**). **i**, Promoter strength of DL variants (top) in the library (P$_{DL}$) or P1$_{rrnB}$ promoter contexts across growth rates in *E. coil*. Dots and lines represent data and linear regression curves, respectively. All slopes are statistically significant ($R^2 > 0.75$, $P < 0.03$), except DL$_1$ variant in the P$_{DL}$ context. **j**,**k**, Base composition of Dis sequences associated with genes involved in translation (TL), nucleotide metabolism (NM), amino acid metabolism (AM), and signal response (SR) in Gracilicutes (**j**) and Terrabacteria (**k**). **l**, Similarity (Spearman's correlation) of Dis sequence enrichment across TL, NM, AM, and SR genes. Shown here are the mean similarities within and between gene categories across species, based on Extended Data Fig. 9g. Species are excluded if they have D_g, $D_s < 0.2$ or < 10 Dis sequences in a given category. In **d**,**f**, graphic representation for positive and negative enrichment is rescaled for uniformity.

Methods:

Development of a biophysical model

We constructed the PL library containing all possible -35 and -10 sequence variants in the *E. coli* P_{purR} promoter context (Extended Data Fig. 2a). We quantified the promoter strength of each variant by sort-seq and utilized high-reproducibility data for model training (Extended Data Fig. 2b,c). The modeling framework was based on statistical thermodynamics initially proposed by Phillips and Kinney (Extended Data Fig. 2e)[34,35]. It assumes promoters exist in either the unbound or RNAP-bound state, with promoter strength proportional to the RNAP-bound probability. The state's relative probability is determined by the Boltzmann weight $B = e^{-\beta\varepsilon}$, where -ε denotes the free energy, $\beta = 1/RT$, gas constant R = 1.98e-3 kcal mol^{-1} K^{-1}, and T = 310 K. The unbound-state free energy is set to zero as the ground condition. The free energy of the RNAP-bound state ($-\Delta\varepsilon_{RNAP}$) is the sum of the free energies of the -35 element, spacer element, -10 element, and the P_{purR} promoter context ($-\Delta\varepsilon_{-35}$, $-\Delta\varepsilon_{Spacer}$, $-\Delta\varepsilon_{-10}$, $-\Delta\varepsilon_{0}$, respectively). The model assumes each base at each position in the -35 and -10 elements contributes to free energies additively. It further assumes $-\Delta\varepsilon_{Spacer}$ varies with the spacer length regardless of sequence[36]. During model fitting, the model scanned through each PL variant sequence, identified all potential -35 and -10 element pairs with spacer length as 16-18 bp, and selected the configuration that maximized $-\Delta\varepsilon_{RNAP}$ as the promoter. The model assumes the transcription rate as r_{max} in the bound state, whereas there exists basal measurement noise r_{min} in the unbound state, with $r_{min} << r_{max}$. We estimated model parameters by minimizing the mean squared error using the Python SciPy and scikit-learn packages[37]. The PL library promoter strength distribution was divided into ten equal-sized bins, from which 1,000 promoter variants per bin were randomly sampled with replacement as the training datasets (Extended Data Fig. 2d)[38]. We set r_{max} = 1000 due to insufficient high-expression PL variants for parameter estimation. To prevent parameter degeneracy in the ample solution space, the free energy sum of four bases at each position and the free energy of the 17-bp spacer ($-\Delta\varepsilon_{Spacer(17)}$) were set to zero. Model fitting estimated $-\Delta\varepsilon_{0}$ = 7.3, r_{min} = 3.13, the free energies of the 16-bp ($-\Delta\varepsilon_{Spacer(16)}$) and 18-bp ($-\Delta\varepsilon_{Spacer(18)}$) spacer elements, and that of each base in the -35 and -10 elements (Extended Data Fig. 2g). To expand the model prediction range for spacer length, we adopted $-\Delta\varepsilon_{Spacer(15)}$ and $-\Delta\varepsilon_{Spacer(19)}$ estimated by another study[36]. The two parameters were rescaled to adjust for the numerical difference, reflected by the fold difference of $-\Delta\varepsilon_{Spacer(16)}$, $-\Delta\varepsilon_{Spacer(17)}$, and $-\Delta\varepsilon_{Spacer(18)}$ between the two studies.

Model prediction and sequence analysis of promoter elements

We applied the biophysical model to predict promoter elements across the domain Bacteria. We collected the TSS information of 49 diverse species from the literature (Supplementary Table 1). Among them, 47 TSS datasets were obtained by dRNA-seq or other direct detection methods, while two (*N. lyticus* and *M. pneumoniae*) were indirectly inferred based on the abrupt increase of the RNA-seq read coverage upstream of genes. Only primary TSS, the dominant TSS in a TSS cluster upstream of genes[39], was chosen for model prediction. The model scanned the 46-bp region upstream of each TSS to predict the -35 and -10 elements based on the criteria that their hexamer composition and the spacer length (15-19 bp) maximized $-\Delta\varepsilon_{RNAP}$. To verify the model's ability to identify true promoters, we created two negative control datasets with the sample size identical to the TSS dataset: one containing 46-bp random genomic sequences and the other containing promoter-shuffled sequences generated by permutating the 46-bp sequence upstream of each TSS. Model prediction of the -35 and -10 elements and identification of the

start element (i.e., the 3-bp region centered upon TSS) defined ITR and the remaining promoter elements for each TSS (Supplementary Data 1). Sequence features in each promoter element and ITR were revealed by Log_2 enrichment and KL divergence analysis. Let $p_i(b)$ and $q_i(b)$ denote the frequencies of base $b \in \{A, T, G, C\}$ at position i in a data subset and the whole dataset, respectively, where $\Sigma_b p_i(b) = \Sigma_b q_i(b) = 1$. The Log_2 enrichment $E_i(b)$ was computed as $E_i(b) = \log_2 (p_i(b)/q_i(b))$. The KL divergence D_i of position i was calculated as $D_i = \Sigma_b p_i(b) \cdot \ln (p_i(b)/q_i(b))$.

Sequence and phylogenetic analysis of RNAP subunits

We applied Biopython 1.81 to collect all coding sequences (CDS) of the 49 species from their genome files (Supplementary Table 1)[37]. We then used RpoA, RpoB, RpoC, and RpoD of *E. coli* as the query sequences for BLAST, setting E-value =1e-5 as the cutoff to identify the homologous proteins of α, β, and β` subunits and $\sigma^{70}$, respectively (Extended Data Figs. 1 and 6). BLAST with *E. coli* RpoA identified exactly one α subunit homolog per species. BLAST with *E. coli* RpoB identified one β subunit homolog per species, but the corresponding CDS in *H. pylori* additionally encoded the β` subunit at the C-terminal. Based on multiple sequence alignment, we divided this CDS into N- and C-terminal segments homologous to the β and β` subunits, respectively. We used the MUSCLE package in Jalview 2.11.3.3 with default settings to generate multiple sequence alignments, both here and subsequently[40,41]. BLAST with *E. coli* RpoC found two CDS in each of *Anabaena* sp., *Synechococcus elongatus*, and *Synechocystis* sp. as the cyanobacterial β` subunit has evolved into two smaller γ and β` subunits[42]. We identified their protein residues homologous to the full-length β` subunit by performing multiple sequence alignments with the β` subunits of the remaining 46 bacteria. Then, the homologous segments were concatenated to create a chimeric full-length β` subunit sequence for each cyanobacterium. BLAST with *E. coli* RpoD found 164 CDS showing significant sequence similarity in 49 bacteria. To distinguish the σ^{70} subunit from other paralogs in the same protein family, we constructed a maximum likelihood (ML) tree using the smart model selection (SMS) and fast likelihood-based method of PhyML 3.0[43,44]. Results revealed a 56-taxa clade that comprised single CDS from 40 species (e.g., *E. coli* RpoD) and 2-4 CDS from species belonging to Actinobacteriota (*S. coelicolor*, *Corynebacterium diphtheriae*, *Corynebacterium glutamicum*, *Mycolicibacterium smegmatis*, *Mycobacterium tuberculosis*) and Firmicutes (*L. phytofermentans*, *Clostridioides difficile*). As two species, *Phytoplasma* sp. and *Burkholderia cenocepacia*, had no CDS identified by BLAST due to missing annotation in their genome files, we separately retrieved the corresponding CDS (BAD04713.1 and CAR54775.1) from the NCBI GenBank and merged them with the 56-taxa clade, making the σ^{70} subunit collection. The σ^{70} subunit collection was subjected to multiple sequence alignment, followed by ML tree construction using SMS and the standard bootstrap analysis (n=1,000) of PhyML 3.0 (Extended Data Fig. 1a,c)[43,44]. The ML tree was visualized by FigTree 1.4.4[45]. The pairwise sequence similarity of CDS was calculated using the PairwiseAligner function in the Bio.Align package of Biopython 1.81 using BLOSUM62 as the scoring matrix (Extended Data Fig. 1b). Since the RNAP ω subunit is too short and variable to be reliably identified by BLAST, we retrieved its CDS primarily through finding annotations that contain keywords 'omega' or 'rpoZ' in genome files. We obtained single CDS of the ω subunit in 46 of 49 bacteria. We used *E. coli* RNAP ω subunit (RpoZ) as the BLAST query sequence but still could not find CDS with E-value ≤ 1e-5 in *N. lyticus*, *Chlamydia pneumoniae*, and *Chlamydia trachomatis*. The three species may lack the ω subunit since it is not essential. Multiple sequence alignment of the ω subunit was performed to examine its sequence conservation.

458
459 **Strain and growth medium**
460 *E. coli* K-12 MG1655 was the host strain for experiments throughout this study. LB medium
461 comprised 10 g NaCl (BioShop), 10 g bacterial tryptone (BD Bacto), and 5 g yeast extract (BD
462 Bacto) in one liter of deionized water. LBK medium, LB medium supplemented with filter-
463 sterilized kanamycin sulfate (final concentration as 50 mg/L, Thermo Fisher Scientific), was
464 used for growing cells for plasmid extraction. LBKG medium, LBK medium supplemented with
465 filter-sterilized glucose (final concentration as 10 g/L, Bionovas) to enhance growth and GFP
466 expression, was used in growth assays or fluorescence-activated cell sorting (FACS)
467 experiments. One liter of M9 minimal medium comprised 5.98 g Na_2HPO_4 (AppliChem), 3 g
468 KH_2PO_4 (AppliChem), 0.5 g NaCl, 0.8 g NH_4Cl (AppliChem), 976.7 mL of deionized water, and
469 the following components that were filter-sterilized separately and added immediately before
470 use: 1 mL of 0.1 M $CaCl_2$ (Hayashi Chemical), 2 mL of 1 M $MgSO_4$ (Sigma-Aldrich), 0.2 mL of
471 185 mM $FeCl_3$ (AppliChem), 0.3 mL of 1 mM thiamine hydrochloride (WS Simpson), 9.8 mL of
472 M9 trace element solution, 1 mL of kanamycin sulfate (50 mg/mL), and the intended amount of
473 carbon source and amino acid supplements. M9 trace element solution comprised 0.18 g
474 $ZnSO_4 \cdot 7H_2O$ (J.T.Baker), 0.12 g $CuCl_2 \cdot 2H_2O$ (Sigma-Aldrich), 0.12 g $MnSO_4 \cdot H_2O$ (J.T.Baker),
475 and 0.18 g $CoCl_2 \cdot 6H_2O$ (Sigma-Aldrich) in 980 mL of deionized water. The final concentrations
476 of glucose and glycerol (J.T.Baker) were 5 g/L and 2.2 g/L, respectively. The final concentrations
477 of casamino acids (Bioshop) were 2 or 10 g/L (denoted as 1×CA and 5×CA, respectively). Six
478 kinds of growth medium were applied to grow *E. coli*, resulting in slow to fast cellular growth:
479 M9-glycerol, M9-glucose, M9-glucose-1×CA, M9-glucose-5×CA, LBK, LBKG. SOB medium,
480 consisting of 20 g bacterial tryptone, 5 g yeast extract, 0.584 g NaCl, 0.186 g KCl (J. T. Baker),
481 2.033 g $MgCl_2 \cdot 6H_2O$ (J. T. Baker), and 2.465 g $MgSO_4 \cdot 7H_2O$ (J. T. Baker) in one liter of
482 deionized water, was used in preparation of electrocompetent cells. SOC medium, SOB medium
483 supplemented with filter-sterilized glucose (180 g/L), was applied to restore cells after
484 electroporation. To achieve a higher transformation efficiency in promoter library construction,
485 we used Recovery Medium (LGC Biosearch Technologies) to revive electrocompetent cells
486 instead of SOC medium. Solid medium was made by supplementing one liter of liquid medium
487 with 20 g agar (BioShop).
488
489 **Plasmid and library construction**
490 All primers and oligonucleotides were synthesized by Integrated DNA Technologies and listed in
491 Supplementary Table 2. Most plasmids were constructed by inverse PCR, while some were made
492 by restriction digestion and ligation. Plasmid features and construction procedures are delineated
493 in Supplementary Table 3. Primers for inverse PCR were 5' phosphorylated using T4
494 polynucleotide kinase (NEB). Inverse PCR was conducted using Phusion DNA polymerase
495 (Thermo Fisher Scientific). DpnI restriction enzyme (NEB) was employed to remove template
496 plasmids. T4 DNA Ligase (NEB) was utilized to circularize plasmids. After purification by the
497 Zymo DNA Clean & Concentrator kit, the ligation product was transformed into *E. coli* using the
498 BIO-RAD MicroPulser electroporator. Electrocompetent cells were prepared following
499 established protocols[16]. The promoter backbone and the ribosome binding site on the template
500 plasmids for constructing the PL, SP16, SP17, SP18, and DL libraries were adjusted to prevent
501 growth inhibition by excessive GFP expression while providing a wide dynamic range of
502 fluorescence signals for FACS experiments[46]. Primers for library construction were pre-
503 phosphorylated and HPLC-purified. PL, SL16, SL17, SL18, and DL libraries were constructed
504 by inverse PCR of template plasmids, followed by ligation, and transformation into *E. coli*

505 (Supplementary Table 3). Q5 DNA polymerase (NEB) and two inverse primers, each containing
506 randomized nucleotides, were applied to amplify template plasmids. Following DpnI treatment,
507 T4 DNA Ligase was utilized to circularize plasmids. After DNA purification, ligation products
508 were mixed with 180 μL of freshly made electrocompetent cells and divided into six equal
509 aliquots for electroporation. All transformed cells were pooled and revived in 6 mL of Recovery
510 Medium at 37°C and 225 rpm for 1 hour, transferred to 30 mL of LBK medium, and incubated at
511 37°C and 225 rpm overnight. The transformation efficiency for library construction was $\geq 5\times10^7$
512 colony forming units. Cells in the overnight culture were collected by centrifugation,
513 resuspended in 1 mL of PBS containing 25% glycerol (v/v), aliquoted into 20 portions, and
514 stored at -80°C as the inoculum for sort-seq experiments.
515
516 **Measurement of cell growth and GFP expression**
517 Cellular GFP fluorescence was quantified by a BD FACSJazz cell sorter. Each experiment began
518 with inoculating 1 μL of *E. coli* frozen stocks into 2 mL of growth medium incubated at 37°C
519 and 225 rpm overnight. The next day, 5 μL of the pre-culture was transferred to 2 mL of growth
520 medium, incubated and monitored in a Tecan Spark 10M microplate reader. When OD_{600} reached
521 0.55-0.65, 100 μL of culture was mixed thoroughly with 900 μL of pre-chilled phosphate-
522 buffered saline (PBS) and stored on ice. For each strain, the cell sorter quantified GFP
523 fluorescence (excitation/emission: 488 nm/513 ± 8.5 nm) of 50,000 to 200,000 cells in three
524 replicates.
525
526 **Sort-seq experiment**
527 One aliquot of the promoter library frozen stock was revived in 20 mL LBKG medium.
528 Meanwhile, spike-in variants, promoters with known sequences and expression levels (
529 Supplementary Table 4), were revived in 1 mL LBKG medium. Promoter libraries and spike-in
530 variants were incubated overnight at 37°C and 225 rpm. Subsequently, 50 μL and 2.5 μL of the
531 library and spike-in precultures were transferred to 20 mL and 1 mL LBKG medium,
532 respectively. Both were grown at 37°C and 225 rpm until their OD_{600} reached 0.55-0.65. All
533 spike-in variants were evenly pooled and added to the library in a 1:2500 ratio (v/v). This
534 mixture was diluted ten times with PBS and stored on ice. FACS experiments were performed by
535 a BD FACSJazz cell sorter, and cells were maintained at 4°C throughout. The photomultiplier
536 tube of the cell sorter was calibrated based on spike-in variants pWS19, LP02v14, and LP03v4.
537 The fluorescence distribution of each library was divided into 7-8 ranks, numbered in ascending
538 order according to GFP fluorescence, denoted as the relative fluorescence unit (RFU;
539 Supplementary Table 5). A total of 1.5×10^7 cells were collected, with the number of cells
540 collected for each rank proportional to its relative abundance in the library. After FACS, cells in
541 each rank were immediately grown in 10 mL LBK medium at 37°C and 225 rpm overnight.
542 Subsequently, the plasmids of each rank were extracted using the Qiagen Plasmid Miniprep Kit.
543 To inspect the FACS purity of each rank, 5 μL of overnight cultures were inoculated into 2 mL
544 LBKG medium incubated at 37°C and 225 rpm. The fluorescence distribution was examined by
545 the cell sorter when its OD_{600} reached 0.55-0.65 (Extended Data Figs. 2b and 7a). Promoter
546 sequences in each sorting rank were determined by the Illumina NovaSeq 6000 system.
547 Sequencing library preparation followed the Illumina two-step PCR procedure. A 0.3-kb
548 fragment, including the promoter, 5′ untranslated region, and partial *gfp* gene, was amplified by
549 amplicon PCR using KAPA HiFi DNA polymerase (Roche), along with equal-volume mixed
550 forward primers ILp11, ILp11a, ILp11b, and ILp11c and the reverse primer ILp12. Amplicon
551 PCR products were purified by AMPure XP Beads (Beckman) and then subjected to index PCR

using KAPA HiFi DNA polymerase and a unique pair of index primers from the Nextera XT Index Kit (Illumina). Index PCR products were purified by AMPure XP Beads, whose DNA concentrations were measured by a Qubit fluorometer (Thermo Fisher Scientific). Index PCR products were pooled based on the proportion of each rank in the library. The multiplexed sample was subjected to 150-bp paired-end sequencing. We set FastQC quality score Q ≥ 25 to filter out poor sequencing results.

Using sort-seq data to estimate promoter strength

The occurrence of each promoter variant was counted in each sort-seq rank. Let $F_v(n)$ denote the frequency of promoter variant v in rank n, where $\Sigma_n F_v(n) = 1$. The rank mean R_v of variant v was calculated as $R_v = \Sigma_n n \cdot F_v(n)$. The rank means of spike-in variants were computed in the same manner. Let E_v denote RFU on the Log_{10} scale and $E_v = a \cdot R_v + b$. Parameters a and b were obtained by performing a linear regression of R_v and E_v of the spike-in variants (Supplementary Table 6). The rank mean of a promoter variant was then converted into its promoter strength using E_v. We set a minimal read count threshold for promoter variants (PL: 20, SL16: 8, SL17: 23, SL18: 31, DL: 9) such that the Pearson's correlation coefficients of two sort-seq experimental replicates were ≥ 0.9 (Extended Data Fig. 7b,c). To select data with high reproducibility for model training, we further required variants in the PL library to be detected in at least two sort-seq ranks. (Extnded Data Fig. 2c). We combined data from two sets of sort-seq replicates, averaging promoter strength for variants present in both replicates, for subsequent analysis.

RNA extraction and quantitative PCR

E. coli was grown in 10 mL of LBKG medium at 37°C and 225 rpm overnight. When OD_{600} reached 0.55-0.65, cells were harvested by adding 1/10 the volume of a growth-stopping solution (5% Tris-EDTA saturated phenol and 95% ethanol), followed by centrifugation at 9,000 × g for 10 minutes at 4°C. Total RNA was extracted using the QIAGEN RNeasy Mini Kit, followed by removal of genomic DNA with the Turbo DNA-free Kit (Thermo Fisher Scientific). cDNA was synthesized by the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific). For each variant, mRNA extraction and cDNA synthesis were performed twice independently. The primer pairs for detecting the *gfp* and *gapA* transcripts were YLEp5/YLEp6 and YLEp7/YLEp8, respectively. Real-Time PCR of each cDNA sample was performed in two replicates with the iQ SYBR Green Supermix (Bio-Rad) on a CFX Connect Real-Time PCR System (Bio-Rad). Only measurements with the standard deviation of cycle threshold values ≤ 0.2 were retained for analysis. The *gapA* housekeeping gene served as the reference for data normalization. mRNA levels were calculated based on an established method[46,47].

Data availability. Data supporting the results of this study are available at Figshare. The raw sequence reads of the PL, SL16, SL17, SL18, and DL libraries are deposited under National Center for Biotechnology Information BioProject ID PRJNA1049603, PRJNA1049606, and PRJNA1138934.

Code availability. The code of the promoter prediction model 'Promoter Scanner' is available at GitHub (https://github.com/AntonySTKuo/Promoter-scanner/tree/main).

Acknowledgments. We thank the National Taiwan University College of Life Science Technology Commons for instrument support, Uwe Sauer, Chris Marx, Gene-Wei Li, Sheng-Hong Chen, Jun-Yi Leu, and Sen-Lin Tang for insightful discussions. This research is supported by National Taiwan University grants (104R104034, 111L7842, 112L7834, 113L4000-1) and Taiwan National Science and Technology Council grants (110-2311-B-002-008, 111-2311-B-002-020, 112-2311-B-002-011, 113-2311-B-002-008).

Author contributions. H.H.D.C. conceived the idea, acquired funding, supervised the project, and wrote the manuscript with input from all authors. S.T.K., J.K.C., C.C. and H.H.D.C. developed experimental and analytical methods. S.T.K., J.K.C., C.C., W.Y.S., C.H., S.W.L. and H.H.D.C. led the investigation. H.H.D.C. and S.T.K analyzed and visualized the data.

Competing interests. Authors declare that they have no competing interests.

Correspondence and requests for materials should be addressed to Hsin-Hung David Chou.

612 **References:**

613

614 1. Mejía-Almonte, C. *et al.* Redefining fundamental concepts of transcription initiation in
615 bacteria. *Nat. Rev. Genet.* **21**, 699–714 (2020).
616 2. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription
617 initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).
618 3. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of
619 regulatory elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).
620 4. Singh, S. S., Typas, A., Hengge, R. & Grainger, D. C. *Escherichia coli* σ^{70} senses sequence
621 and conformation of the promoter spacer region. *Nucleic Acids Res.* **39**, 5109–5118 (2011).
622 5. Lagator, M. *et al.* Predicting bacterial promoter function and evolution from random
623 sequences. *eLife* **11**, e64543 (2022).
624 6. Voskuil, M. The -16 region of *Bacillus subtilis* and other gram-positive bacterial promoters.
625 *Nucleic Acids Res.* **26**, 3584–3590 (1998).
626 7. Estrem, S. T., Gaal, T., Ross, W. & Gourse, R. L. Identification of an UP element consensus
627 sequence for bacterial promoters. *Proc. Natl. Acad. Sci.* **95**, 9761–9766 (1998).
628 8. Sanchez-Vazquez, P., Dewey, C. N., Kitten, N., Ross, W. & Gourse, R. L. Genome-wide
629 effects on *Escherichia coli* transcription from ppGpp binding to its two sites on RNA
630 polymerase. *Proc. Natl. Acad. Sci.* **116**, 8310–8319 (2019).
631 9. Zhang, Y. *et al.* Structural basis of transcription initiation. *Science* **338**, 1076–1080 (2012).
632 10. Hook-Barnard, I. G. & Hinton, D. M. Transcription initiation by mix and match elements:
633 flexibility for polymerase binding to bacterial promoters. *Gene Regul. Syst. Biol.* **1**,
634 117762500700100 (2007).

635 11. Lee, D. J., Minchin, S. D. & Busby, S. J. W. Activating transcription in bacteria. *Annu. Rev.*
636 *Microbiol.* **66**, 125–152 (2012).
637 12. Pribnow, D. Nucleotide sequence of an RNA polymerase binding site at an early T7
638 promoter. *Proc. Natl. Acad. Sci.* **72**, 784–788 (1975).
639 13. Chevez-Guardado, R. & Peña-Castillo, L. Promotech: a general tool for bacterial promoter
640 recognition. *Genome Biol.* **22**, 318 (2021).
641 14. Cassiano, M. H. A. & Silva-Rocha, R. Benchmarking bacterial promoter prediction tools:
642 potentialities and limitations. *mSystems* **5**, e00439-20 (2020).
643 15. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and
644 evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2012).
645 16. Kuo, S.-T. *et al.* Global fitness landscapes of the Shine-Dalgarno sequence. *Genome Res.*
646 (2020) doi:10.1101/gr.260182.119.
647 17. Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.*
648 **15**, 116–124 (2005).
649 18. Vvedenskaya, I. O. *et al.* Massively Systematic Transcript End Readout, "MASTER":
650 Transcription Start Site Selection, Transcriptional Slippage, and Transcript Yields. *Mol. Cell*
651 **60**, 953–965 (2015).
652 19. Kim, D. *et al.* Comparative analysis of regulatory elements between *Escherichia coli* and
653 *Klebsiella pneumoniae* by genome-wide transcription start site profiling. *PLOS Genet.* **8**,
654 e1002867 (2012).
655 20. Myers, K. S., Noguera, D. R. & Donohue, T. J. Promoter architecture differences among
656 *Alphaproteobacteria* and other bacterial taxa. *mSystems* **6**, e00526-21 (2021).
657 21. Bayley, D. P., Rocha, E. R. & Smith, C. J. Analysis of *cepA* and other *Bacteroides fragilis*
658 genes reveals a unique promoter structure. *FEMS Microbiol. Lett.* **193**, 149–154 (2000).

659 22. Vingadassalom, D. *et al.* An unusual primary sigma factor in the Bacteroidetes phylum. *Mol.*
660 *Microbiol.* **56**, 888–902 (2005).
661 23. Campbell, E. A. *et al.* Structure of the bacterial RNA polymerase promoter specificity σ
662 subunit. *Mol. Cell* **9**, 527–539 (2002).
663 24. Einav, T. & Phillips, R. How the avidity of polymerase binding to the –35/–10 promoter sites
664 affects gene expression. *Proc. Natl. Acad. Sci.* **116**, 13340–13345 (2019).
665 25. Haugen, S. P. *et al.* rRNA promoter regulation by nonoptimal binding of σ region 1.2: an
666 additional recognition element for RNA polymerase. *Cell* **125**, 1069–1082 (2006).
667 26. Schneider, D. A., Ross, W. & Gourse, R. L. Control of rRNA expression in *Escherichia coli*.
668 *Curr. Opin. Microbiol.* **6**, 151–156 (2003).
669 27. Wu, C. *et al.* Cellular perception of growth rate and the mechanistic origin of bacterial
670 growth law. *Proc. Natl. Acad. Sci.* **119**, e2201585119 (2022).
671 28. Ross, W., Vrentas, C. E., Sanchez-Vazquez, P., Gaal, T. & Gourse, R. L. The magic spot: a
672 ppGpp binding site on *E. coli* RNA polymerase responsible for regulation of transcription
673 initiation. *Mol. Cell* **50**, 420–429 (2013).
674 29. Irving, S. E., Choudhury, N. R. & Corrigan, R. M. The stringent response and physiological
675 roles of (pp)pGpp in bacteria. *Nat. Rev. Microbiol.* **19**, 256–271 (2021).
676 30. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**,
677 eabe0511 (2021).
678 31. Decker, K. B. & Hinton, D. M. Transcription regulation at the core: similarities among
679 bacterial, archaeal, and eukaryotic RNA polymerases. *Annu. Rev. Microbiol.* **67**, 113–139
680 (2013).
681 32. Lou, C., Stanton, B., Chen, Y.-J., Munsky, B. & Voigt, C. A. Ribozyme-based insulator parts
682 buffer synthetic circuits from genetic context. *Nat. Biotechnol.* **30**, 1137–1142 (2012).

683 33. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for
684 taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592
685 (2023).
686 34. Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.*
687 **15**, 116–124 (2005).
688 35. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to
689 characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl.*
690 *Acad. Sci.* **107**, 9158–9163 (2010).
691 36. Lagator, M. *et al.* Predicting bacterial promoter function and evolution from random
692 sequences. *eLife* **11**, e64543 (2022).
693 37. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular
694 biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
695 38. Botev, Z. & Ridder, A. Variance reduction. in *Wiley statsRef: statistics reference online* 1–6
696 (John Wiley & Sons, Ltd, 2017).
697 39. Thomason, M. K. *et al.* Global transcriptional start site mapping using differential RNA
698 sequencing reveals novel antisense RNAs in *Escherichia coli*. *J. Bacteriol.* **197**, 18–28
699 (2015).
700 40. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview
701 Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*
702 **25**, 1189–1191 (2009).
703 41. Troshin, P. V., Procter, J. B. & Barton, G. J. Java bioinformatics analysis web services for
704 multiple sequence alignment—JABAWS:MSA. *Bioinformatics* **27**, 2001–2002 (2011).
705 42. Shen, L. *et al.* An SI3-σ arch stabilizes cyanobacteria transcription initiation complex. *Proc.*
706 *Natl. Acad. Sci.* **120**, e2219290120 (2023).

707 43. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood
708 Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
709 44. Lefort, V., Longueville, J.-E. & Gascuel, O. SMS: Smart Model Selection in PhyML. *Mol.*
710 *Biol. Evol.* **34**, 2422–2424 (2017).
711 45. Rambaut, A. Figtree ver 1.4.4. Institute of Evolutionary Biology, University of Edinburgh
712 (2018).
713 46. Kuo, S.-T. *et al.* Global fitness landscapes of the Shine-Dalgarno sequence. *Genome Res.* **30**,
714 711–723 (2020).
715 47. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time
716 quantitative PCR and the 2−ΔΔCT method. *Methods* **25**, 402–408 (2001).

717
718