Computation-guided redesign of promoter specificity of a bacterial RNA polymerase

Xiangyang Liu[1,‡], Anthony T. Meger[2,3,‡], Thomas Gillis[2,‡], and Srivatsan Raman[2,3,4,5,*]

[1]Biophysics Graduate Program, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[2]Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[3]Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[4]Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[5]Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[‡]These authors contributed equally to this work

Contact Information

*Corresponding author: Srivatsan Raman (sraman4@wisc.edu)

ABSTRACT

Synthetic genetic circuits enable precise and dynamic control of gene expression, making them useful tools in metabolic engineering and synthetic biology for regulating cellular responses under transient states, such as metabolic flux and stress. However, the complexity and utility of these systems are restricted by the limited set of available regulatory parts. To engineer circuits with desirable levels of programmable control, these components must be orthogonal from the host expression systems. Bacterial sigma factors and RNA polymerases are ideal targets for synthetic regulation, as they represent the first level of transcriptional control and modulate gene expression in endogenous systems. While attempts have been made to engineer orthogonality through heterologous sigmas, a lack of diversity in consensus promoter sequences and toxicity often limit the utility of these sigmas. In this work, we leverage computational design and high-throughput screening to engineer σ70 bacterial transcription factors capable of targeting any promoter sequence to maximize orthogonality. To demonstrate the power of this approach, we engineered several σ70 variants, each capable of targeting one of five different orthogonal promoter sequences, leading to the expression of a downstream reporter gene. By engineering these sigma variants, we present a generalized workflow capable to produce orthogonal sigma factors that may serve as a vital tool in the design of customizable synthetic genetic circuits.

KEYWORDS: Protein-DNA interactions; Computational design; biological orthogonalization; specificity; high-throughput screening

INTRODUCTION

Dynamic and precise control of gene expression is a fundamental challenge in synthetic biology. With emerging technologies to synthesize chromosomes from scratch, multiple gene regulation codes can be partitioned and programmed into cells.[1,2] Such a system could have orthogonal regulation encoded in different genomic regions while sharing common core machinery like transcription.[3,4] Gene regulation in bacteria is governed at the global and local levels. At a global level, the RNA polymerase (RNAP) is directed to hundreds to thousands of loci simultaneously by a battery of sigma factors to maintain homeostasis and respond to environmental changes.[5] At a local level, individual genes and operons may be subject to the additional control of individual transcription factors. Over the past two decades, the synthetic biology community has developed a large toolkit of transcription factors to build circuits, pathways, switches, and gene expression systems in bacteria.[6-8] In contrast, efforts to redesign RNAP to exert control at a global level remain poorly developed. Furthermore, although some genes or operons can be regulated from synthetic promoters with non-native regulation, large swathes of the bacterial genome, especially genes involved in core metabolic processes, require native regulation as they are often layered in complex networks that cannot be easily disentangled. Orthogonal regulation in these genes can be introduced by modifying RNAP promoter specificities without disrupting native regulation. Thus, redesigning the RNAP is a powerful but underexplored route to engineer orthogonal genetic regulation.

The sigma subunit of the bacterial RNAP governs the interaction of the RNAP with the promoter.[9-11] The bacterial RNAP is a multi-subunit molecular machine composed of alpha, beta, omega, and sigma subunits. The alpha, beta, and omega subunits form the core RNAP enzyme responsible for RNA synthesis using DNA as a template and ribonucleotides as the substrate.[12,13] To initiate transcription, the core RNAP binds to a sigma factor which directs the RNAP to specific promoters.[10,11,14,15] Bacteria express several types of sigma factors that can recognize and bind specific promoters in response to cellular signals and environmental conditions.[15-17] Of the seven types of sigma factors in *E. coli*, sigma-70 is the dominant as it is responsible for expressing housekeeping genes, and the other sigma factors are only transiently expressed during stress.[18-23] Prior strategies to engineer RNAP have used heterologous sigma factors from other hosts or constructed chimeras of native and non-native sigma factors to engineer new RNAP promoter

67 specificities.[24-26] However, heterologous σ factors are often toxic to the host, limiting the extent to which
68 they may be used in synthetic expression systems. Although some σ factors engineered via domain swap
69 display minimal crosstalk with host expression systems, this approach has generated few orthogonal σ
70 factor-promoter pairs, and the sequence variation between promoter target sequences are often minimal.
71 Additionally, the utility of domain-swapped and heterologous σ factors are strain-dependent. Thus, although
72 the sigma factor is an attractive target for orthogonal genetic regulation, existing strategies to engineer them
73 have limitations.
74
75 In this study, we redesigned the *E. coli* sigma-70 to recognize and transcribe from five orthogonal promoters
76 that are transcriptionally inactive to wild-type sigma-70, using computation-guided design and pooled cell-
77 based screens. These five promoters were chosen as targets to engineer sigma specificity based on a
78 randomized screen of the -35 element enriching for variants that do not express a fluorescent reporter. The
79 chosen targets (CCGCCG, GCATCC, TTCATC, CCCCTC, and GGAACC) have hamming distances of
80 three to six for the canonical GGTACC sequence. We used Rosetta to create a customized library of
81 computationally redesigned variants of the domain of *E. coli* sigma-70 that interacts with -35 site for each
82 promoter target. Fluorescence-activated cell sorting followed by deep sequencing of the sigma libraries of
83 the unsorted and sorted populations revealed the sequence determinants of promoter specificity and how
84 they changed for different targets. We found that recognition of new promoters occurs through a
85 combination of highly conserved residues that likely make polar interactions with DNA backbone and target
86 specific adaptations. We identified orthogonal sigma variants for each of the five targets whose activities
87 ranged from 20-70% of wildtype sigma-70 on GGTACC. Our workflow is generalizable and can be applied
88 to alternative sigma factors in *E. coli* and sigma factors from other host organisms. The redesigned sigma
89 factor-promoter pairs constitute an important tool for engineering orthogonal genetic regulation at the
90 genome level.

91 **RESULTS**

92 **The partitioned functional domains of σ70 enable targeted redesign of promoter specificity**

93 As the "housekeeping" regulator of transcription in *E. coli*, σ70 performs three functions. First, σ70 must
94 recognize the consensus -10 (TATAAT) and -35 (TTGACA) promoter elements. The -10 element is

95 recognized by conserved domains 2 and 3, and the -35 element is recognized by the helix-turn-helix motif
96 of domain 4 (domain 1 is unstructured) (**Fig. 1a**). Second, σ70 is responsible for recruiting the core RNA
97 polymerase complex to the promoter. The interface between RNA polymerase and σ70 is extensive (>8000
98 $Å^2$) as it spans the entire length of domains 2-4. The third function of σ70 is to mediate the melting of DNA
99 near the -10 region, which is a critical step for initiating transcription. To operate as a global regulator of
100 cellular transcription, σ70 must perform each of these functions with high efficiency and fidelity.
101
102 To engineer orthogonal sigma variants, we must alter promoter specificity while preserving interactions with
103 the core RNA polymerase complex and those required for DNA melting. Therefore, we aimed to restrict the
104 design to key residues of the helix-turn-helix motif of domain 4 to alter promoter specificity through the -35
105 DNA element. These include residues R584, E585, R586, R588, and Q589 (**Fig. 1a, box**). In the structure
106 of the *E. coli* transcription initiation complex (PDB ID: 4YLN), the positively charged R586 and R588 form
107 polar interactions with the negatively charged phosphate backbone to likely promote general DNA affinity.
108 Specific interactions with the nucleotide bases are formed between R584:Gua-31* and E585:Ade-34* in
109 the major groove of the -35 element. Although oriented towards the major groove, Q589 does not directly
110 interact with the -35 element. Instead, Q589 is sterically packed between residues E585, R586 and R588
111 and, likely plays a role in properly orienting these critical residues within the recognition helix. Given their
112 structural roles, we hypothesized that redesigning these five key positions could engineer orthogonal
113 promoter specificity of σ70.
114 Our design workflow begins with identifying -35 element mutants that are not recognized by wildtype sigma-
115 70. These mutants became recognition targets for the computational redesign of the major groove helix
116 residues of wildtype sigma-70 (**Fig. 1b**). The top 1000 ranked designs were synthesized as chip-based
117 oligonucleotide library and evaluated in pooled fashion using a cell-based GFP screen. In this assay, high
118 GFP fluorescence signifies a functional redesign of σ70, as the variant must recognize the orthogonal -35
119 element to initiate transcription of GFP. To identify functional redesigns from the multiplex screen, we
120 compared the distribution of variants between sorted and presorted states by deep sequencing (**Fig. 1c**).
121 Clonal screens were then used to validate several successful redesigns for each targeted -35 element.
122

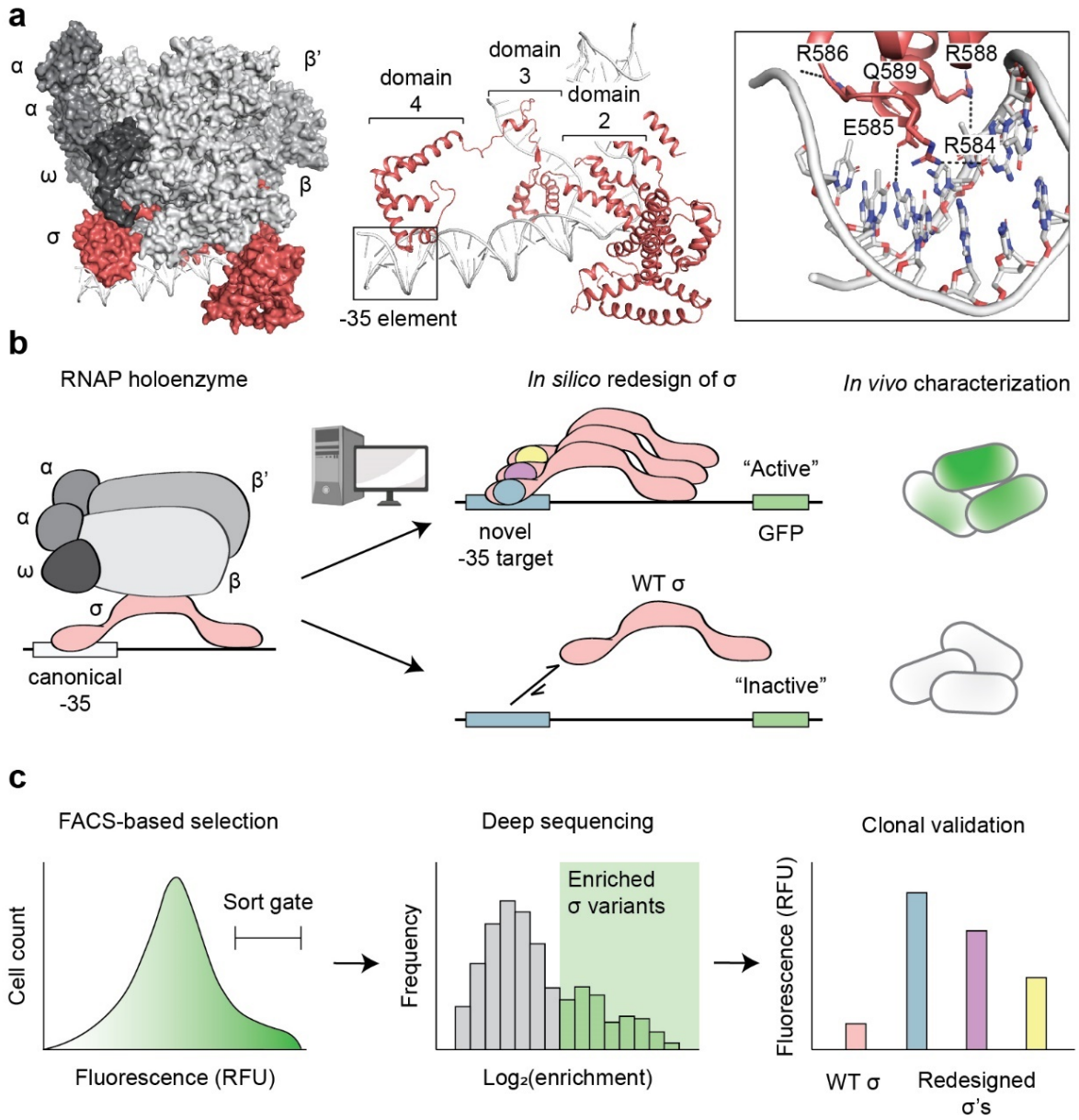

Fig. 1. Engineering σ70 variants with novel promoter specificity. (**a**), Structure of the *E. coli* transcription initiation complex (PDB ID: 4YLN). σ70 (red) bound to the -10 and -35 DNA elements (gray, cartoon representation) and RNAP core complex (gray, surface representation), composed of α (two copies), β, β', and ω subunits (left). Canonical -35 element recognized by the helix-turn-helix motif of σ70 domain 4 and -10 element recognized by domains 2 and 3 (middle). Orientation and h-bond contacts of residues R584, E585, R586, R588, and Q589 of the domain 4 recognition helix in the -35 major groove (right). (**b**), Workflow for altering σ70 promoter specificity *in silico* and characterizing function *in vivo*. Cartoon schematic of the RNAP holoenzyme with WT σ70 bound to the canonical -35 element (left). Replacement of the canonical -35 with a novel target and computational redesign of the domain 4 recognition helix to engineer σ70 variants with orthogonal promoter specificities (middle). *In vivo* functional

136 characterization using a GFP gene placed downstream of the promoter variants containing novel -35 targets
137 (right). (**c**), Platform for high-throughput selection and characterization of redesigned σ70 variants. FACS
138 to isolate GFP positive (i.e., functional variants) (left), deep sequencing to identify enriched variants
139 (middle), and clonal testing of redesigned σ70s to validate function (right).
140

141 **Synthetic promoter modifications enhance σ70 dependence on the -35 element**

142 As mentioned above, σ70 utilizes a two target DNA recognition system, with domain 4 recognizing the
143 canonical -35 element, and domains 2 and 3 recognizing the canonical -10 element. Due to this two-target
144 dependence, engineering an orthogonal sigma factor would require redesign of multiple σ70 DNA binding
145 domains as well as their respective consensus DNA sequences. However, by artificially increasing
146 dependence of DNA recognition on the -35 DNA element, redesign is simplified to a one domain-one target
147 problem. There are several advantages to engineering orthogonality through the -35 element rather than
148 the -10 element: (1) -35 recognition is mediated by a single domain of σ70, (2) interactions between the
149 helix-turn-helix motif of domain 4 with the -35 major groove are well characterized, and (3) interactions
150 between σ70 and the -10 element are much more complex as transcription initiation is mediated by DNA
151 melting near this region. Thus, we sought to increase σ70's dependence on the -35 element, then engineer
152 σ70 variants that recognize novel -35 targets.
153

154 We first placed a gene encoding GFP downstream of the strong constitutive *E. coli* promoter, apFab71, on
155 a "reporter plasmid" to utilize fluorescence as a measure of transcriptional activity. As expected, a negative
156 control lacking both canonical -10 and -35 promoter elements resulted in very low fluorescence relative to
157 the apFab71 promoter with the canonical -35 element (TTGACA) (**Fig. 1a**). However, moderate
158 fluorescence was still observed after the -35 was randomized (NNNNNN), suggesting that the canonical -
159 10 element alone is partially sufficient to recruit σ70. We aimed to identify a promoter variant which has
160 high transcriptional activity with the canonical -35 sequence, but significantly lower activity when paired with
161 randomized -35 sequences. To establish greater dependence on the -35 element, we characterized a small
162 subset of nucleotide deletion variants (Δ(-11), Δ(-16,-7), Δ(-8), Δ(-7)) near the -10 region of apFab71. As
163 desired, all four apFab71 promoter variants retained -35 dependence (i.e., lower fluorescence in the
164 randomized condition) and displayed approximately a 10-fold reduction in fluorescence relative to the native
165 apFab71 promoter with a randomized -35 sequence (**Fig. 2a**). The fold-change of median fluorescence

between canonical and randomized -35 states was used to evaluate -35 dependence. The Δ(-7) promoter exhibited the greatest -35 dependence with the fluorescence fold-change increasing from 9.3 in the native apFab71 promoter to 83 after the Δ(-7) nucleotide deletion. Thus, the Δ(-7) promoter was selected as the starting construct for redesigning promoter specificity of σ70.

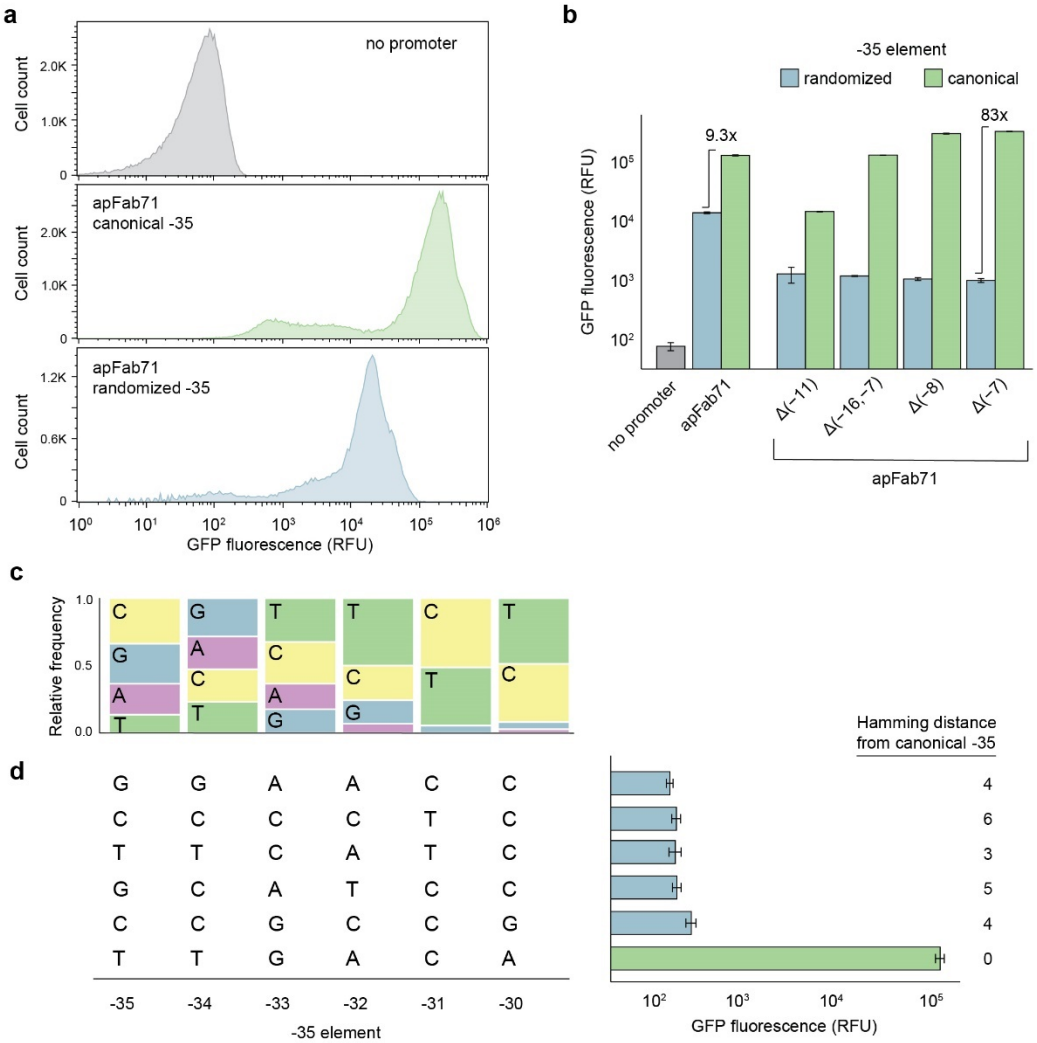

Fig. 2. σ70 dependence on the -35 element and selection of orthogonal promoter targets. (**a**), Fluorescence distributions of *E. coli* cells expressing native σ70 and containing plasmid constructs with a GFP gene placed downstream of no promoter (gray), the strong constitutive apFab71 promoter (green), or the apFab71 promoter with a randomized -35 element (blue). (**b**), Median fluorescence of promoter variants containing nucleotide deletions (Δ) near the -10 region, and either a canonical or randomized -35 element. Fold-change between canonical and randomized -35 states was used to evaluate -35 transcriptional dependence. (**c**), Sequence logos (-35 region) of 74 unique clonal isolates with low GFP fluorescence (<10^3 RFU). Clones were screened in 96-well format from the apFab71 Δ(-7) randomized -35 library. (**d**), Selected -35 DNA targets for engineering orthogonal σ70 variants. Nucleotide sequences (left) and median fluorescence (right) using native σ70. Error bars denote the standard deviation of replicate flow cytometry experiments (n≥2). RFU, relative fluorescence units.

183 **Endogenous *E. coli* σ70 displays very weak activity on selected -35 targets**

184 We next sought to identify a set of five promoter targets orthogonal to the canonical -35 element recognized
185 by endogenous *E. coli* σ70. From the library of randomized -35 sequences of the apFab71 Δ(-7) promoter,
186 we used a high-throughput 96-well screening approach to isolate and characterize promoter variants not
187 recognized by native σ70. We isolated 74 unique -35 variants with low σ70 transcriptional activity (GFP
188 fluorescence $<10^3$ RFU). To graphically evaluate nucleotide sequence conservation among low activity -35
189 variants, we generated a sequence logo. As expected, there are many unique ways to disrupt σ70 promoter
190 recognition, and therefore, little sequence conservation exists among low activity variants (**Fig. 2c**).
191 Interestingly, all nucleotides except Cyt-31 of the canonical -35 (TTGACA) were the least frequently
192 observed bases at each position among these low activity variants, which is also reflected by their median
193 hamming distance of 5. This suggests that as sequences diverge further from the canonical -35 sequence,
194 the probability of recognition by native σ70 decreases. Among these low activity -35 promoter variants, we
195 selected GGAACC, CCCCTC, TTCATC, GCATCC, and CCGCCG as target sequences for engineering
196 σ70 variants with novel specificities. These promoter targets all displayed low transcriptional activity,
197 comparable to the no promoter control, and vary in their hamming distances from the canonical -35
198 sequence (**Fig. 2d**). Due to their varying hamming distances from the canonical sequence, we hypothesized
199 that these 5 promoter sequences present would present unique design challenges and demonstrate the
200 breadth of our redesign capabilities.

201 **Computational tools and high-throughput screening facilitate the design of orthogonal σ70 variants**

202 The mutant combinatorial sequence space of σ70 positions 584, 585, 586, 588, and 589 is vast (20^5 =
203 3,200,000 variants). Synthesizing and functionally characterizing a library of this size would be challenging
204 and inefficient, as most combinatorial sequences are expected to have no detectable affinity for a targeted
205 orthogonal -35 element. Using the state-of-the-art Rosetta macromolecular modeling software suite, we
206 refined the sequence space to a small set of 1000-2000 variants with high probabilities of function (i.e.
207 favorable interface scores) on several novel -35 targets (CCGCCG, GCATCC, TTCATC, CCCCTC, and
208 GGAACC). For these DNA binding energy calculations, the canonical -35 element (TTGACA) was first
209 replaced with an orthogonal target that we empirically validated was unable to initiate transcription using
210 native σ70 (**Fig. 1b**). Then, we performed an *in silico* scan to evaluate the stability of the protein-DNA

211 complex for all single, double, triple, and quadruple mutants of the σ70 recognition helix. Calculations for
212 all 723,901 σ70 variants against the five target -35 elements were performed on a high-throughput
213 computing (HTC) system, requiring approximately 400,000 computing hours (2-3 weeks real-time). The
214 redesigned σ70 variants were ranked by binding energy, and the highest affinity 1000 variants for each
215 target -35 element were selected for experimental testing. In addition, we selected 1000 variants with
216 binding energies comparable to native σ70 and the canonical -35 element. We synthesized these variant
217 libraries on an IPTG-inducible plasmid expression system using oligonucleotide chip synthesis and one-pot
218 cloning.
219
220 **Rosetta calculations reveal the σ70 sequence preferences for each promoter target**
221 To engineer σ70 variants that recognize each of the five -35 targets, we first performed large-scale
222 combinatorial mutagenesis of residues R584, E585, R586, R588, and Q589 in the helix-turn-helix motif of
223 domain 4 using the Rosetta macromolecular modeling suite. Computed interface scores for the ~724,000
224 sigma variants in complex with the promoter targets were used to curate designs for experimental testing
225 (**Fig. 3a**). For these calculations, lower interface scores indicate a more stable complex. Interestingly,
226 differences between median interface scores among σ70 variants revealed a hierarchy of promoter
227 preference of the σ70 scaffold (TTGACA < TTCATC < GGAACC < CCGCCG < GCATCC < CCCCTC) with
228 interface scores decreasing with increasing hamming distance. This suggests an evolutionary and
229 structurally derived preference of σ70 to recognize the canonical -35 element. To assess σ70 sequence
230 preferences with each promoter, we computed position-specific amino acid enrichment scores (E_{score}) for
231 the top 1000 scoring interfaces (**Fig. 3b**). Due to the limited variety of natural protein-DNA interactions,
232 subtle differences exist between amino acid preferences across -35 targets.

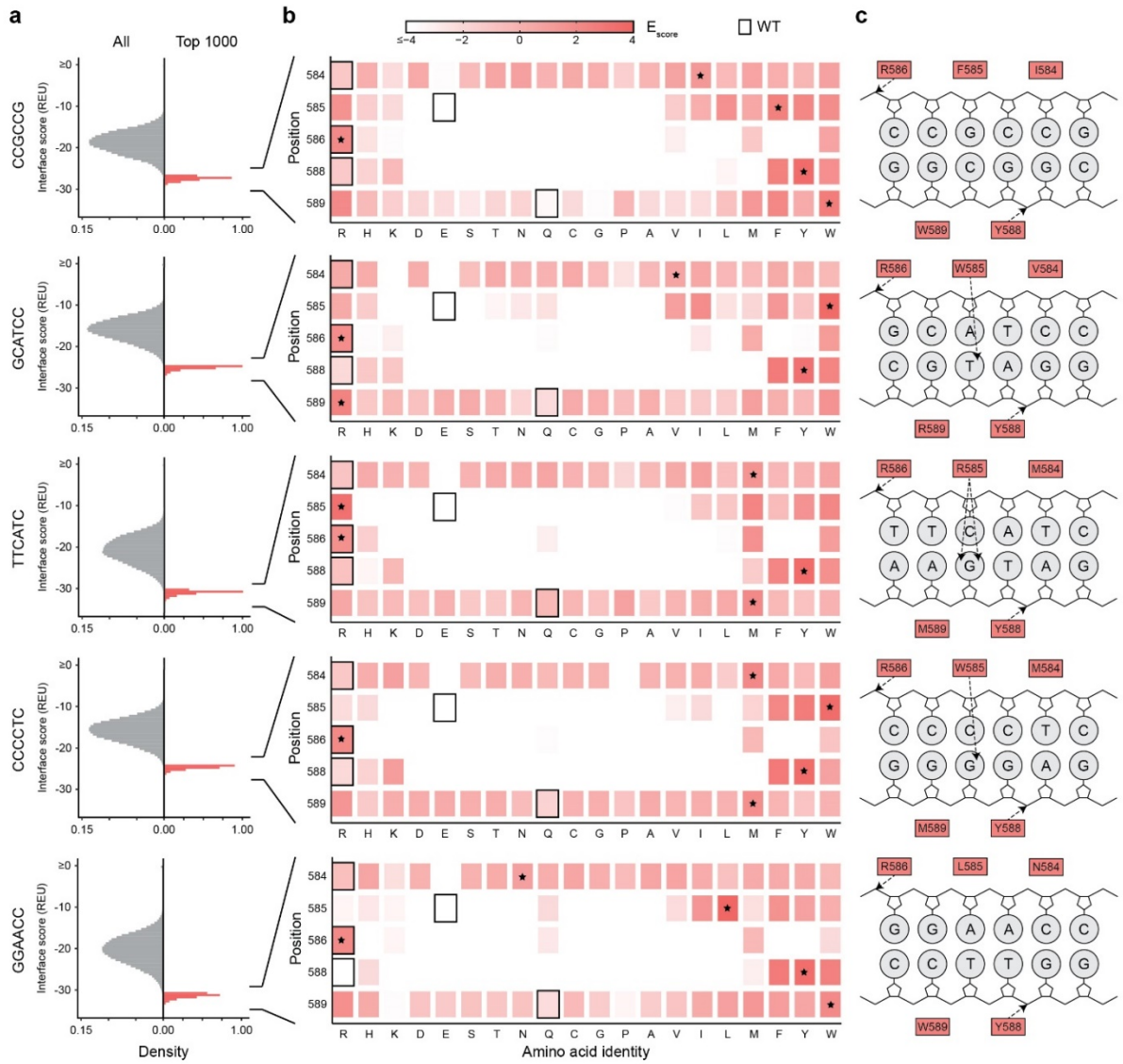

Fig. 3. Computation-based selection of redesigned σ70 variants for each promoter target. (**a**), Interface scores of all (gray) or 1000 top scoring (red) σ70 variants in complex with each -35 DNA target. All single, double, triple, and quadruple combinatorial variants of σ70 positions 584, 585, 586, 588, and 589 were modeled using Rosetta. (**b**), Position-specific amino acid enrichment scores (red gradient) among selected top scoring σ70 variants. WT identity (boxed outline) and most enriched amino acid (*) at each mutable position. (**c**), Cartoon schematic showing H-bonds formed between each enriched σ70 consensus sequence and -35 DNA target in the Rosetta structural models.

Enrichment heat maps indicate σ70 positions 585, 586, and 588 undergo stringent selection across all -35 targets. These positions directly contact DNA in the crystal structure of native σ70 in complex with the canonical (TTGACA) -35 element (**Fig.1a**). As expected, basic and/or large hydrophobic amino acids are generally preferred at these positions, which can facilitate highly favorable hydrogen bonding, electrostatic and/or van der Waals interactions with the DNA. For all promoters, arginine and tyrosine were the most

247 enriched amino acids at positions 586 and 588, respectively. Both residues make non-specific hydrogen
248 bonds with the sugar-phosphate backbone of DNA (**Fig. 3c**). Clear differences in amino acid preference
249 were evident for position 585. For the -35 target TTCATC, arginine was the most enriched amino acid,
250 which forms two hydrogen bonds with Gua-35* in the Rosetta structural models. Similarly, a single hydrogen
251 bond is formed between W585 with Gua-35* or Thy-35* for promoters CCCCTC and GCATCC,
252 respectively. In contrast, bulky hydrophobic residues F585 and L585 make van der Waals contacts with the
253 CCGCCG and GGAACC promoters. Because positions 584 and 589 do not directly interact DNA,
254 substitutions at these positions had less distinct effects on the Rosetta computed interface scores.
255
256 **Isolation and identification of σ70 variants with novel promoter specificities**
257 To synthesize each computationally curated library of σ70 designs, we used chip-based oligonucleotides
258 and one-pot cloning. The gene encoding each σ70 variant was placed under transcriptional control of the
259 lac operon on an expression plasmid, and these assembled libraries were used to transform *E. coli*
260 harboring their cognate reporter plasmid (i.e., GFP gene placed downstream of the target -35 element).
261 This IPTG inducible system provides several key advantages: (1) transcriptional activity on a target
262 promoter is linked directly to expression of a σ70 variant, and (2) the level of overexpression can be tuned
263 to minimize toxicity to the *E. coli* host. Due to the "housekeeping" role of σ70, overexpression of a σ70
264 variant can competitively sequester the RNAP core enzyme to inhibit cell viability. This toxicity challenge
265 has been reported in previous works using heterologous σ's and domain swaps. To identify an expression
266 level with minimal toxicity, we first varied the IPTG inducer concentration in a clonal growth assay. Minimal
267 growth deficiency was observed using 50μM IPTG, and thus, this concentration was used for all subsequent
268 experiments.

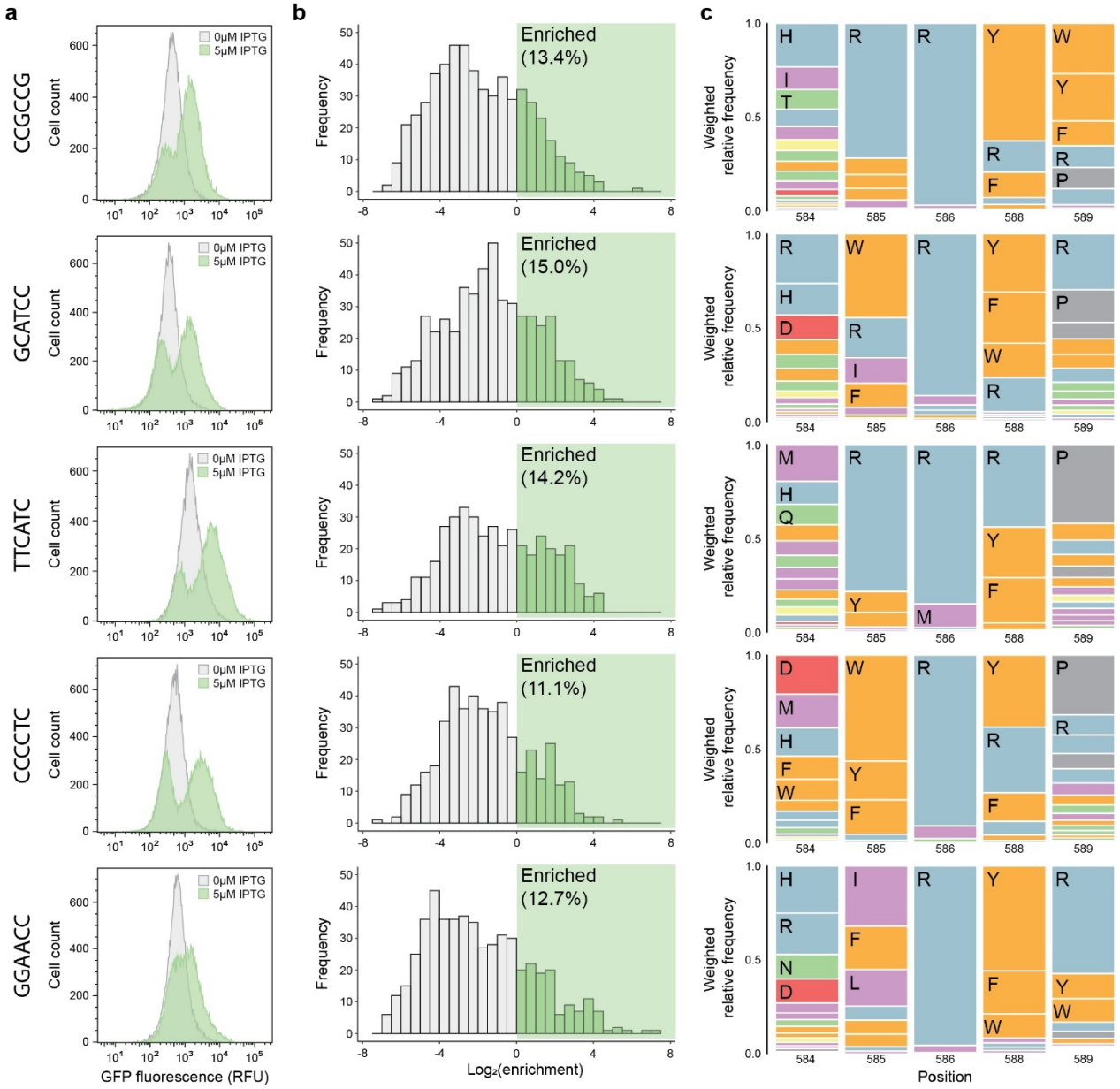

Fig. 4. Selection and identification of successfully redesigned σ70 variants. (**a**), Flow cytometry fluorescence distributions of uninduced (gray) and IPTG induced (green) σ70 variant populations for each -35 target after FACS-based selection of functional redesigns. Transcriptionally active variants were enriched using two rounds of sequential GFP positive cell sorting (**Supplementary Fig. X**). (**b**), Distributions of log-transformed enrichment scores of all characterized σ70 variants after selection. Deep sequencing was performed on the presorted and sorted libraries to compute enrichment ratios. (**c**), Sequence logos showing the weighted amino acid frequencies at each mutable position among functionally enriched σ70 variants. Amino acid identities are colored by chemical properties: polar amino acids (N, Q, S, T) shown in green, basic (H, K, R) blue, acidic (D, E) red, hydrophobic (A, I, L, M, V) purple, aromatic (F, W, Y) orange, and other (C, G, P) gray.

281 To isolate and enrich high activity σ70 redesigns, we sorted the high fluorescence subpopulation of each
282 library in the IPTG induced condition. After two rounds of sorting, flow cytometry was used to compare the
283 fluorescence distributions of uninduced and IPTG induced cell states and thereby, demonstrate successful
284 selection of transcriptionally active variants (**Fig. 4a**). The presorted and sorted libraries were deep
285 sequenced, and we computed enrichment ratios to identify functional σ70 redesigns from each library.
286 Approximately 10-15% of the σ70 variants within each 1000-member library were enriched after selection
287 (**Fig. 4b**). Taken together, these results validate our *in silico* redesign approach and high-throughput
288 screening platform.
289
290 **Sequence profiles of redesigned σ70 variants are unique to each -35 target**
291 Following deep sequencing, we evaluated the σ70 sequence profiles among enriched library variants of
292 each -35 target. Given the intragroup median hamming distance of 4 across the five promoter targets and
293 variations in distance relative to the canonical -35 promoter, we expected the mutable positions to undergo
294 unique sequence selection (i.e., position-specific amino acid enrichment) for each library. Across all
295 redesigned libraries, selection was less stringent at positions 584 and 589 of σ70, as a broad range of
296 amino acid identities were found among enriched variants (**Fig. 4c**). While H584 performed well across all
297 targets, basic residues appear more favorable at this position for targets GCATCC and GGAACC (hamming
298 distance separation of 2, D=2). In contrast, the other targets had less preference in terms of chemical
299 properties at position 584. Both P589 and R589 were commonly enriched across most promoter targets,
300 with P589 being heavily favored for TTCATC and CCCCTC (D=3) and R589 for GCATCC and GGAACC
301 (D=2). Interestingly, aromatic hydrophobic residues (W, Y, F) were uniquely favored at position 589 for the
302 CCGCCG target.
303
304 We observed more extreme sequence preferences at positions 585, 586, and 588 (**Fig. 4c**). R585 was
305 highly favored by targets CCGCCG and TTCATC (D=6), W585 by GCATCC and CCCCTC (D=4), and
306 I/L585 by (GGAACC). In contrast, R586 was universally preferred across all targets. At position 588, all
307 target libraries displayed a similar trend among enriched amino acids. Varying in their relative enrichment
308 and order of most enriched amino acids, the libraries all exhibited an obvious preference for aromatic or

basic amino acids (Y, F, W, R). While the target -35 elements are orthogonal, these results demonstrate that promoter specificity of σ70 is often dictated by subtle and compounding differences in primary sequence.

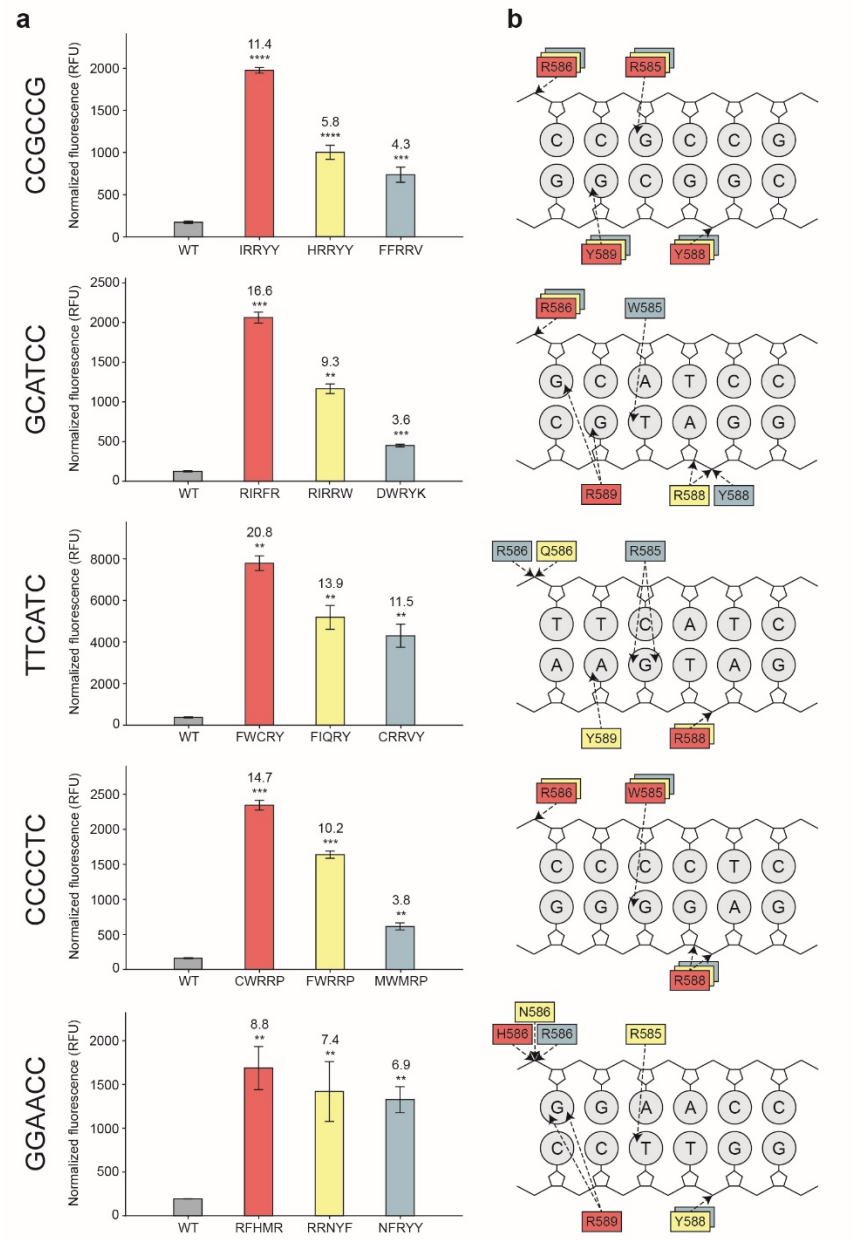

Fig. 5. Clonal validation of redesigned σ70 variants. (**a**), Normalized fluorescence of WT σ70 (gray) and the top three performing variants (red, yellow, or blue) for each promoter target. Clones were selected from the sorted libraries (**Supplementary Fig. X**) and assayed in a 96-well fluorescence plate reader. Error bars denote the standard deviation (**P ≤ 0.01, ***P ≤ 0.001) of replicates (n≥3) and computed fold-changes are relative to WT. RFU, relative fluorescence units. (**b**), Cartoon schematic showing H-bonds formed between each σ70 variant and the cognate -35 DNA target in the Rosetta structural models.

320 **Clonal hits show varying degrees of redesign success**

321 Due to the varied divergence from the -35 consensus sequence seen in our set of five orthogonal promoters,
322 it was expected that some target sequences would be more challenging to design compatible sigma factors
323 for than others, with the possibility of some proving too difficult for design of a compatible sigma. Through
324 clonal testing we were able to confirm that some target sequences yielded higher performing redesigned
325 sigma variants than other targets. However, we were able to find multiple sigma variants capable of
326 promoting modest gene expression for each orthogonal target sequence. To assess the performance of
327 each computationally designed sigma variant, we first measured the fluorescence of cells containing the
328 target sequences upstream of GFP in the presence of the cell's endogenous sigma factors. In the presence
329 of only the cells' endogenous sigma factors, the promoter variants yielded fluorescence values of 173, 125,
330 374, 160, and 193 RFU for the CCGCCG, GCATCC, TTCATC, CCCCTC, and GGAACC target sequences
331 respectively (**Fig. 5a**). After recording this baseline measurement, we measured the fluorescence of each
332 high-performing sigma variant identified during clonal testing and calculated each variants improvement in
333 activity over the endogenous sigma factors.

334

335 For the target sequence CCGCCG, which diverges from the canonical -35 sequence at 4/6 bases, we found
336 that the designed sigma variants IRRYY, HRRYY, and FFRRV lead to GFP expression levels on the target
337 promoter that are 11.4, 5.8 and 4.3 times higher than those observed from the endogenous sigma factors.
338 For the target sequence GCATCC, which diverges from the canonical -35 sequence at 5/6 base positions,
339 we found that the designed sigma variants DWRYK, RIRRW and RIRFR, display gene expression that is
340 3.6, 9.3, and 16.6-fold improvements over the WT sigma factors at this target sequence. For the target
341 sequence TTCATC, which diverges from the canonical -35 sequence at 3/6 base positions, we found the
342 three highest performing computationally designed sigma factors, CRRVY, FIQRY, and FWCRY, exhibited
343 11.5, 13.9, and 20.8-fold improvements respectively over the WT sigma variants for this target sequence.
344 For the target sequence CCCCTC, which diverges from the canonical -35 sequence at all of its base
345 positions, we found the sigma variants, CWRRP, FWRRP, and MWMRP displayed activities that were a
346 modest 14.7, 10.2, and 3.8-fold improvements over the WT sigma's activity at this target sequence. Finally,
347 for the target sequence GGAACC, which diverges from the canonical sequence at 4/6 bases, we found that

348 the sigma variants RFHMR, RRNYF, and NFRYY displayed activities that were 8.8, 7.4, and 3.8 times
349 higher than the WT sigma's activity on that target sequence.
350
351 **DISCUSSION**
352 These results demonstrate that the promoter specificity of an essential primary transcription regulator can
353 be rationally redesigned. Although the process of transcription initiation (i.e. promoter recognition and DNA
354 melting) by σ factors is complex, the partitioned functional domains of σ factors enabled key interactions
355 between the domain 4 recognition helix and the -35 DNA element to be redesigned without perturbing other
356 functions. Using our high throughput *in silico* modeling and *in vivo* screening workflow, we identified multiple
357 *E. coli* σ70 variants that activate transcription on each of the 5 promoter targets, demonstrating the extant
358 at which promoter specificity can be altered and the generalizability of our approach. Furthermore, given
359 the median hamming distance separation of 4 (out of 6 nucleotide positions) between each target and
360 relative to the canonical -35 element, our approach enables the design of σ factors with maximal
361 orthogonality from natural σ factors and thereby, greater potential to minimize crosstalk between synthetic
362 and host expression systems. In contrast, σ factors generated in previous works using domain swaps of
363 alternative σ factors or introduced from non-host family bacteria, have been limited in their orthogonality
364 and host compatibility.
365
366 One advantage to the engineered σ70 variants presented is that they integrate seamlessly into the *E. coli*
367 host, as they utilize the same endogenous transcriptional machinery. While the level of activation varied
368 from 3-fold to 20-fold across our top performing redesigns, we show that this activity was also engineered
369 de novo, as the native σ70 exhibited little to no activity on each of the orthogonal promoters. Using structural
370 models of these successful redesigns in complex with their cognate promoters, we observe diverse
371 combinations of interactions that facilitate promoter recognition. This diversity extends even among
372 successful redesigns for the same promoter target, demonstrating the utility of computational tools to
373 navigate the energetic landscape and simplify the sequence space to be experimentally tested.
374

The transcription-level regulatory complexity of a cell is mediated by the concerted efforts of global and local regulators. While many existing synthetic expression systems are reliant solely on local regulators (i.e. natural or synthetic transcription factors) to control a sparse set of genes, they lack the complexity and capabilities of native systems. Only by incorporating global regulatory components, such as engineered σ factors, can we expand the utility and complexity of synthetic genetic circuits. However, building or isolating an extensive repertoire of global regulatory components is challenging because they must be (1) orthogonal to host systems to minimize crosstalk between the synthetic and native expression networks, (2) display dynamic control by activating gene expression in response to simple input signals, and (3) be highly customizable and modular, such that they can be easily integrated into established expression systems. Here, we developed and validated a generalizable approach to engineer σ70 variants that satisfy all these requirements and can be similarly applied to any host σ factor to expand the set of available global regulatory components. Future work could pair these engineered σ70 factors with existing local regulators of transcription in a biosynthetic expression system.

ACKNOWLEDGEMENTS

This work was supported in part by the Great Lakes Bioenergy Research Center, U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018409. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, Department of Defense, Department of Energy, or other federal agencies.

AUTHOR CONTRIBUTIONS

X.L., A.T.M., T.G., and S.R. conceptualized the project. X.L. devised the cell-based screening system, cloned the promoter library, and performed sorting and NGS experiments. A.T.M. performed the *in silico* calculations and designed the sigma variant libraries. T.G. performed clonal screens and identified the top performing sigma variants for each promoter target. X.L., A.T.M., T.G, and S.R. drafted the manuscript. A.T.M., X.L., and T.G. created the figures. S.R. supervised the work. All authors edited and approved the final manuscript.

405 **REFERENCES**

406 1 Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and
407 translation initiation elements. *Nature Methods* **10**, 354-360 (2013).
408 https://doi.org:10.1038/nmeth.2404
409 2 Grimbs, A., Klosik, D. F., Bornholdt, S. & Hütt, M.-T. A system-wide network reconstruction
410 of gene regulation and metabolism in Escherichia coli. *PLOS Computational Biology* **15**,
411 e1006962 (2019). https://doi.org:10.1371/journal.pcbi.1006962
412 3 Bradley, R. W., Buck, M. & Wang, B. Tools and Principles for Microbial Gene Circuit
413 Engineering. *Journal of Molecular Biology* **428**, 862-888 (2016).
414 https://doi.org:https://doi.org/10.1016/j.jmb.2015.10.004
415 4 Slusarczyk, A. L., Lin, A. & Weiss, R. Foundations for the design and implementation of
416 synthetic genetic circuits. *Nature Reviews Genetics* **13**, 406-420 (2012).
417 https://doi.org:10.1038/nrg3227
418 5 Feklístov, A., Sharon, B. D., Darst, S. A. & Gross, C. A. Bacterial Sigma Factors: A
419 Historical, Structural, and Genomic Perspective. *Annual Review of Microbiology* **68**, 357-
420 376 (2014). https://doi.org:10.1146/annurev-micro-092412-155737
421 6 Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators.
422 *Nature* **403**, 335-338 (2000). https://doi.org:10.1038/35002125
423 7 Stanton, B. C. *et al.* Genomic mining of prokaryotic repressors for orthogonal logic gates.
424 *Nature Chemical Biology* **10**, 99-105 (2014). https://doi.org:10.1038/nchembio.1411
425 8 De Paepe, B., Peters, G., Coussement, P., Maertens, J. & De Mey, M. Tailor-made
426 transcriptional biosensors for optimizing microbial cell factories. *Journal of Industrial*
427 *Microbiology and Biotechnology* **44**, 623-645 (2017). https://doi.org:10.1007/s10295-016-
428 1862-3
429 9 Browning, D. F. & Busby, S. J. W. The regulation of bacterial transcription initiation. *Nature*
430 *Reviews Microbiology* **2**, 57-65 (2004). https://doi.org:10.1038/nrmicro787
431 10 deHaseth Pieter, L., Zupancic Margaret, L. & Record, M. T. RNA Polymerase-Promoter
432 Interactions: the Comings and Goings of RNA Polymerase. *Journal of Bacteriology* **180**,
433 3019-3025 (1998). https://doi.org:10.1128/JB.180.12.3019-3025.1998
434 11 Feklistov, A. RNA polymerase: in search of promoters. *Annals of the New York Academy*
435 *of Sciences* **1293**, 25-32 (2013). https://doi.org:https://doi.org/10.1111/nyas.12197
436 12 Burgess, R. R. & Travers, A. A. Escherichia coli RNA polymerase: purification, subunit
437 structure, and factor requirements. *Fed Proc* **29**, 1164-1169 (1970).
438 13 A Darst, S. Bacterial RNA polymerase. *Current Opinion in Structural Biology* **11**, 155-162
439 (2001). https://doi.org:https://doi.org/10.1016/S0959-440X(00)00185-8
440 14 Helmann, J. D. & Chamberlin, M. J. STRUCTURE AND FUNCTION OF BACTERIAL
441 SIGMA FACTORS. *Annual Review of Biochemistry* **57**, 839-872 (1988).
442 https://doi.org:10.1146/annurev.bi.57.070188.004203
443 15 Paget, M. S. Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and
444 Distribution. *Biomolecules* **5**, 1245-1265 (2015).
445 16 Kazmierczak Mark, J., Wiedmann, M. & Boor Kathryn, J. Alternative Sigma Factors and
446 Their Roles in Bacterial Virulence. *Microbiology and Molecular Biology Reviews* **69**, 527-
447 543 (2005). https://doi.org:10.1128/MMBR.69.4.527-543.2005
448 17 Gruber, T. M. & Gross, C. A. Multiple Sigma Subunits and the Partitioning of Bacterial
449 Transcription Space. *Annual Review of Microbiology* **57**, 441-466 (2003).
450 https://doi.org:10.1146/annurev.micro.57.030502.090913
451 18 Lonetto, M., Gribskov, M. & Gross, C. A. The sigma 70 family: sequence conservation and
452 evolutionary relationships. *J Bacteriol* **174**, 3843-3849 (1992).
453 https://doi.org:10.1128/jb.174.12.3843-3849.1992

19 Marles-Wright, J. & Lewis, R. J. Stress responses of bacteria. *Current Opinion in Structural Biology* **17**, 755-760 (2007). https://doi.org:https://doi.org/10.1016/j.sbi.2007.08.004

20 Helmann, J. D. The extracytoplasmic function (ECF) sigma factors. *Adv Microb Physiol* **46**, 47-110 (2002). https://doi.org:10.1016/s0065-2911(02)46002-x

21 Maeda, H., Fujita, N. & Ishihama, A. Competition among seven Escherichia coli σ subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Research* **28**, 3497-3503 (2000). https://doi.org:10.1093/nar/28.18.3497

22 Cowing, D. W. & Gross, C. A. Interaction of Escherichia coli RNA polymerase holoenzyme containing σ32 with heat shock promoters: DNase I footprinting and methylation protection. *Journal of Molecular Biology* **210**, 513-520 (1989). https://doi.org:https://doi.org/10.1016/0022-2836(89)90127-7

23 Koo, B. M., Rhodius, V. A., Nonaka, G., deHaseth, P. L. & Gross, C. A. Reduced capacity of alternative sigmas to melt promoters ensures stringent promoter recognition. *Genes Dev* **23**, 2426-2436 (2009). https://doi.org:10.1101/gad.1843709

24 Rhodius, V. A. *et al.* Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. *Mol Syst Biol* **9**, 702 (2013). https://doi.org:10.1038/msb.2013.58

25 Gaida, S. M. *et al.* Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. *Nature Communications* **6**, 7045 (2015). https://doi.org:10.1038/ncomms8045

26 Bervoets, I. *et al.* A sigma factor toolbox for orthogonal gene expression in Escherichia coli. *Nucleic Acids Research* **46**, 2133-2144 (2018). https://doi.org:10.1093/nar/gky010

478 **METHODS**

479 **Rosetta design**

480 Rosetta design focused on σ^{70} positions R584, E585, R586, R588 and Q589 (PDB: 4YLN) as previous
481 research has demonstrated that these 5 positions interact with -35 region nucleotides through hydrogen
482 bonds (Campbell, Muzzin et al. 2002). Rosetta v3.9 was used for all calculations following existing Rosetta
483 design protocol (Thyme and Baker 2014). 5 promoters bearing different -35 sequences (TTCATC,
484 GGAACC, CCGCCG, GCATCC and CCCCTC) were chosen after clonal testing, and were desirable due
485 to their low transcription activity in vivo. To generate starting structures for combinatorial protein
486 mutagenesis, each promoter sequence was first introduced in the WT ΣDNA complex structure. Nucleotide
487 substitutions were made at the -35 region (WT=TTGACA).

488

489 Input structures were subjected to energy minimization, and lowest energy structures from 50 minimizations
490 were used as the input structure for the combinatorial protein mutagenesis scan at 5 positions
491 (WT=RER_RQ). Standard res files were used to denote mutations at each position for all variants
492 containing single (95), double (3610), triple (68,590), and quadruple (651,605) amino acid mutations.
493 Variants containing 5 mutations (2,476,099) were not tested due to the excessive burden on computational
494 resources. The structure of each protein variant in complex with all promoters (including WT) were
495 generated and the Protein-DNA binding affinity was scored. The binding energy was reported as an average
496 of 10 Rosetta minimized structures. In this minimization procedure, the amino acid substitutions were first
497 introduced followed by side chain rotomeric state optimization and protein backbone design. Lastly, the
498 Protein-DNA interface was optimized and scored using the DNA Interface Packer move. Out of these
499 designs for each target promoter sequence, 2000 designs were selected based on two criteria. The first
500 criteria sought to contain variants with the highest binding affinity for target promoter sequence, i.e. the
501 lowest Rosetta Energy Units (REU). The second criteria sought to contain variants with a binding energy
502 for target sequence which was closest to the observed binding energy of WT σ^{70} on its native target
503 sequence.

504

505 The reason behind the first criterium is that low REU confers tighter binding to the target promoter
506 sequence, which could translate into higher transcription level. On the other hand, σ^{70} amino acid
507 composition and WT -35 sequence are likely the local minima through evolution. Although the pair's REU
508 is not the lowest among all variants, it might indicate that there is an optimal REU to allow tight enough
509 binding while permitting the release when RNAP moves downstream. The assumption is that tightest
510 binding hinders release. For some promoters, there might be overlap between two groups chosen
511 independently.
512
513 **Library cloning**
514 A promoter library was ordered via Integrated DNA Technology and amplified by hybridizing with a reverse
515 primer (hybridization protocol: 95oc for 3 minutes, 55°C for 1 minute, 72°C for 1 minute). The product was
516 subsequently cleaned up using PCR cleanup kit (Omega Biotek) and stored in dH2O. Backbone preparation
517 started with PCR amplification of plasmid pXL-9. For Rosetta designed σ variants, 110-base pair (bp) single
518 strand DNA oligo pool containing Rosetta designed σ fragments were ordered from Agilent. Oligo design
519 features unique priming regions at 5' and 3' for individual libraries complemented with BsaI recognition sites
520 and cutting sites to enable Goldengate cloning. 10ng of the oligo pool was used in the PCR reaction to
521 amplify individual libraries (PCR protocol: 95°C for 3 minutes, 98°C for 20 seconds, 55°C for 15 seconds,
522 72°C for 8 seconds, cycling from step 2 for 20 cycles, 72°C for 30 seconds). PCR product was subsequently
523 cleaned up using PCR cleanup kit (Omega Biotek) and stored in dH2O. Backbone preparation started with
524 PCR amplification of plasmid SC101_LacI_ΣWT containing a WT σ 70 driven by pLacO promoter. Amplified
525 backbone bared the corresponding BsaI binding and cutting sites, enabling scarless cloning. Backbone
526 PCR was cleaned up and subject to sequential digestion of DPN1 and BsaI_HF_V2 (New England Biolab)
527 to remove the template DNA and expose sticky overhangs. The digested backbone was further incubated
528 with antarctic phosphatase (New England Biolab) to remove 5' and 3' phosphates. The removal of those
529 phosphates prevents backbones from circularizing and creating false positive transformants. The
530 Goldengate reaction was done as described on New England Biolab website. Briefly, 300ng of the
531 backbone was combined with 70ng of the library in a 20μL reaction. The mix was incubated at 37°C for 1

532 hour and 65°C for 5 minutes. Cooled reaction mix was dialyzed on 0.02μm filter in dH2O for 1 hour at room
533 temperature. The dialyzed sample was collected and stored at -20°C.
534
535 **Transformation of Promoter and Redesigned σ Libraries**
536 Library transformation was conducted using electro-component DH10B *E. coli* (New England Biolab). 2μL
537 of the assembled library reaction mix was transformed with 25 μL cells and recovered with 1mL Soc medium
538 at 37°C for 1 hour. Multiple logs of dilutions were plated on appropriate antibiotic plate to measure
539 transformation efficiency. 4mL of LB and antibiotics were added to ensure proper growth selection. Cultures
540 were grown overnight before being stored in 25% glycerol at -80°C. For non-library transformations, 10ng
541 DNA was added directly to 25 μLof cells and plated on agar plates containing antibiotics to confer selection.
542 Co-transformations were conducted by first growing cells containing one plasmid in LB medium with
543 appropriate antibiotics overnight. 1:50 back dilution was then done into 3mL LB without antibiotics. Cells
544 were allowed to grow until reaching an OD600 of 0.6 before being chilled on ice. Chilled cells were spun
545 down at 5000g and thoroughly washed with ice cold dH2O twice. The final cell pellet was suspended in 25
546 μL of ice cold dH2O. Electroporation was then repeated similarly to regular plasmid transformation.
547 Transformed cells were plated onto agar plates with appropriate antibiotics overnight at 37°C before storage
548 at 4°C.
549
550 **Induction and fluorescence measurement**
551 Multiple colonies were picked from agar plates and inoculated into 150μL LB medium in a 96-well plate
552 containing appropriate antibiotics. Cells were allowed to grow shaking (900rpm, multi-well plate shaker,
553 Southwest scientific) at 37°C for around 3 hours or until OD600 reached 0.6. Cultures were back inoculated
554 into two technical replicates of fresh LB medium with appropriate antibiotics at 1:20 dilution factor. After
555 around 3 hours or upon reaching an OD600 of 0.3, IPTG (at final concentration of 10μM) was added to one
556 of the two technical replicates. Cells were incubated for 5 hours before fluorescence measurement.
557 Fluorescence measurements were conducted using a multi-well platereader (HTX Biotek). Excitation and
558 emissions wavelengths of 485nm and 528nm (± 20) were used for all measurements. Measurements were
559 normalized by dividing fluorescence intensities by the OD^{600} before subtracting blank well fluorescence

560 (cells carrying no GFP). Fold-improvement was calculated by dividing normalized the IPTG induced
561 fluorescence of σ^{70} variants by the IPTG induced fluorescence of WT σ^{70} containing cells using the same
562 target promoter sequence. Average fold improvement was calculated from at least 3 biological replicates.
563
564 **Fluorescence activated cell sorting**
565 Around 25µl of reporter transformed libraries carrying a target promoter upstream of GFP were inoculated
566 into 3mL LB kan/spec 50 (kanamycin and spectinomycin at 50µg/mL) from glycerol stock and grown
567 overnight at 37°C in a shaking incubator. Cultures were inoculated into two separate 150µL fresh LB
568 kan/spec50 in a 96-well plate at 1:50 dilution. Fresh cultures grew shaking at 37°C for 3 hours or until
569 reaching an OD600 of 0.3. IPTG was added to a final concentration of 10 uM to one technical replicate for
570 induction. After the addition of IPTG, cultures were grown for 5 hours before being chilled on ice. Chilled
571 cultures were added to ice cold phosphate buffer saline (PBS) at 1:50 ratio and mixed well before being put
572 back onto the ice. Cell sorting was conducted using a Sony SH800 cell sorter (Sony Biotech) (condition:
573 FCS threshold=2500, 50% PMT on GFP, ultra-purity, round 8k events/second). Three independent
574 replicates were sorted. Briefly, cells in PBS were flowed through first to capture the distribution. Under 10µM
575 IPTG induced condition, top 10% cells were sorted out (at least 50k events). Sorted populations were flowed
576 through the sorter to capture sorting efficiency (percentage of cells fall back into the sorted gate). Sorted
577 cells with less than 80% efficiencies were sorted using the same gate again to improve the purity. Sorted
578 cells were recovered in 1mL LB medium shaking at 37C. 100µL (1:10 dilution) of cells were plated onto LB
579 agar plates with appropriate antibiotics. The rest were growing overnight in 3ml LB with appropriate
580 antibiotics before being miniprepped into plasmid. Promoter library sorting was done by simply isolating 5%
581 of cells with lowest fluorescence.
582
583 **Variant identification via next generation sequencing (NGS)**
584 A pair of primers upstream and downstream of the mutated σ region were used to add partial illumina
585 adaptors. Priming sites were at least 20 base pair away from the variable regions to ensure good quality
586 base calling during NGS. Amplicon size was 120 base pair. Amplicons were sequenced on Amplicon-EZ
587 Miseq service (Genewiz) with at least 50k reads for each σ library. Sequenced libraries were processed

588 with PEAR pair-ended merger (quality threshold set to 35). Merged reads were processed with custom
589 script to identify variants. Briefly, σ variable regions were extracted and converted into amino acids with
590 corresponding frequency. Reads with frequencies of less than 5 were removed. Variant frequency was
591 normalized against total reads to get percentages. To compute individual variant's enrichment, a variant's
592 percentage under 10μM IPTG induction was divided by that of the condition without IPTG. Weblogos for
593 variants with highest enrichment were generated with Weblogo (weblogo.berkeley.edu).
594
595 **Clonal identification**
596 To identify highly functional σ variants, 100 colonies were sampled from the sorted populations. Highly-
597 functional (defined by its fold improvement level) colonies were measured using the above protocol for plate
598 reader measurements, and high performing colonies were subsequently identified using Sanger
599 sequencing. Unique high-performing variants were then inoculated in 3 mL kan/spec LB media and grown
600 overnight. Following overnight growth, plasmids from these variants were extracted using the protocol for
601 plasmid mini-prep. Following plasmid extraction, plasmids from these variants were re-transformed into
602 DH10B *E. coli* in order to normalize for variations in the *E. coli* genome of sorted cells that may have led to
603 false positives in GFP transcription during clonal testing. Using three biological replicates, the re-
604 transformed σ^{70} variants were measured once again using the induction protocol from initial clonal screens.
605 High-performing σ^{70} variants were measured alongside cells containing WT σ^{70} and the orthogonal
606 promoter sequences to ensure that fold-improvement scores are consistent across measurements.
607