A Codon Constrained Method for Both Eliminating and Creating Intragenic Bacterial Promoters

Dominic Y. Logel, Ellina Trofimova, Paul R. Jaschke*.

Department of Molecular Sciences, Macquarie University, Sydney 2109, New South Wales, Australia

*Correspondence to Paul R Jaschke: paul.jaschke@mq.edu.au

Abstract:

Future applications of synthetic biology will require refactored genetic sequences devoid of internal regulatory elements within coding sequences. These regulatory elements include cryptic and intragenic promoters which may constitute up to a third of predicted *Escherichia coli* promoters. Promoter activity is dependent on the structural interaction of core bases with a σ factor. Rational engineering can be used to alter key promoter element nucleotides interacting with σ factors and eliminate downstream transcriptional activity. In this paper, we present COdon Restrained Promoter SilEncing (CORPSE), a system for removing intragenic promoters. CORPSE exploits the DNA-σ factor structural relationship to disrupt σ^{70} promoters embedded within gene coding sequences, with a minimum of synonymous codon changes. Additionally, we present an inverted CORPSE system, iCORPSE, which can create highly active promoters within a gene sequence while not perturbing the function of the modified gene.

Keywords:

Synthetic biology; refactoring; cryptic promoters; internal regulation; structural biology; σ factor; sigma factor

Introduction

Building predictable and orthogonal transcriptional networks is a central goal in synthetic biology. The mass re-engineering of complex native genetic sequences into predictable and rational systems, through a process known as refactoring, is dependent on maintaining biological function while removing cryptic regulation and genetic overlaps [1, 2]. Refactoring is a broad term in synthetic biology applying to multiple approaches which simplify genetic systems on a multi-gene or genomic level. Many examples of refactoring have been demonstrated in both multi-gene pathways [3-5], and bacteriophage [6-8].

The ongoing design of synthetic and modified genetic systems is complicated by internal genetic regulation, such as promoter sequences internal to a coding sequence, which can create deviations from predicted outcomes [3, 9, 10]. While the majority of bacterial promoters are in intergenic regions, many promoters are *intragenic*, overlapping an upstream coding sequence. Well-known examples of intragenic promoters include the *micLp* promoter within the *Escherichia coli cutC* gene driving *micL* small RNA transcription [11] and the *trpCp* promoter within *trpD* driving *trpC* transcription [12].

Recent advances in methods to detect transcription start sites (TSS) across the genome has suggested that *E. coli* transcription under multiple conditions contains nearly 15,000 TSS [13] with 32 % located internal to coding sequences. It is currently unclear how biologically relevant these newly discovered promoters are, and recent work with other experimental methods has shown far fewer promoters in *E. coli* [14].

Bacterial promoter sequences are typically made of two core sequence elements, the –35 and –10 elements. The core elements bind to the primary specification component within the RNA polymerase (RNAP) holoenzyme, the σ factor. Transcription is initiated when a σ factor binds to dsDNA at the –35 element to promote DNA isomerisation to ssDNA at the –10 element which drives transcription events. Promoter DNA-protein interactions are not limited to the two core elements as holoenzyme interactions can be enhanced or replaced through

accessory sequences such as the extended –10 TGn motif and UP elements, which interact with the bacterial σ factor and α subunits respectively [15]. Promoter elements interacting with the housekeeping σ^{70} are the most well-characterised [16], and have been demonstrated to bind specifically with their target promoters through DNA-protein interactions such as hydrogen bonding, Van der Waals forces, and stacked cation-π bonds [17, 18]. The –35 element nucleotide sequence identity is a relatively promiscuous interaction because the σ factor mainly recognises the –35 element's nucleotide backbone [18], while in comparison the –10 element DNA-protein interaction is more discriminatory. The –10 element DNA-protein interaction relies on two bases in the sequence, $^{-11}$A and $^{-7}$T, which base-flip and change orientation from the DNA sequence stack to interact within two highly specific binding pockets in the σ factor [17]. The specific nature of DNA-σ factor interactions leaves opens a rational approach to silently eliminate intragenic promoter sequences within a coding sequence by eliminating DNA-σ interactions through synonymous mutations.

Typically, intragenic promoters have been eliminated during refactoring by extensive randomization of synonymous codons of overlapping coding sequences [5]. However, the mass randomisation of codons in a coding sequence can have unpredictable deleterious effects when specific codons play a play a critical role in translation rate [19], or protein folding and function [20-23]. Furthermore, synonymous codon changes have been seen to create mRNA toxicity independent of translation [24]. Therefore, a rational approach that minimizes codon changes while still effectively disrupting internal promoters would be highly desirable.

In this work, we present a codon-aware promoter removal method called COdon Restrained Promoter SilEncing, or CORPSE. CORPSE aims to disrupt internal promoters through minimal synonymous codon changes. We show the potential of the CORPSE method through erasing the *trpCp* intragenic promoter without perturbing the overlapping *trpD* gene function in *E. coli*. We also present inverted CORPSE (iCORPSE), which is capable of silently inserting a promoter within a coding sequence with only a few synonymous codon

modifications. We demonstrate iCORPSE by creating a new promoter within the fluorescent reporter gene mCherry without disrupting its function. Together, these two methods present a platform for precisely eliminating intragenic promoter elements from coding regions during refactoring, and the construction of new intragenic promoters for next-generation compressed genetic circuit design.

Results and Discussion

The CORPSE method can eliminate promoter activity using only synonymous mutations

To create the CORPSE method, we first gathered σ^{70} promoter sequence data from the *E. coli* K-12 database RegulonDB [25]. From this list of promoters we removed all promoter elements varying from the standard hexameric nucleotide length. The remaining promoters were used to create a position specific scoring matrix (PSSM). We used the PSSM to score promoters based on how close they conformed to the σ^{70} consensus.

Next, we created an algorithm to consider the hexameric sequences in a promoter as codons, as would be the case for intragenic promoters, and then to generate alternative sequences that are limited to only synonymous codon variants at each position.

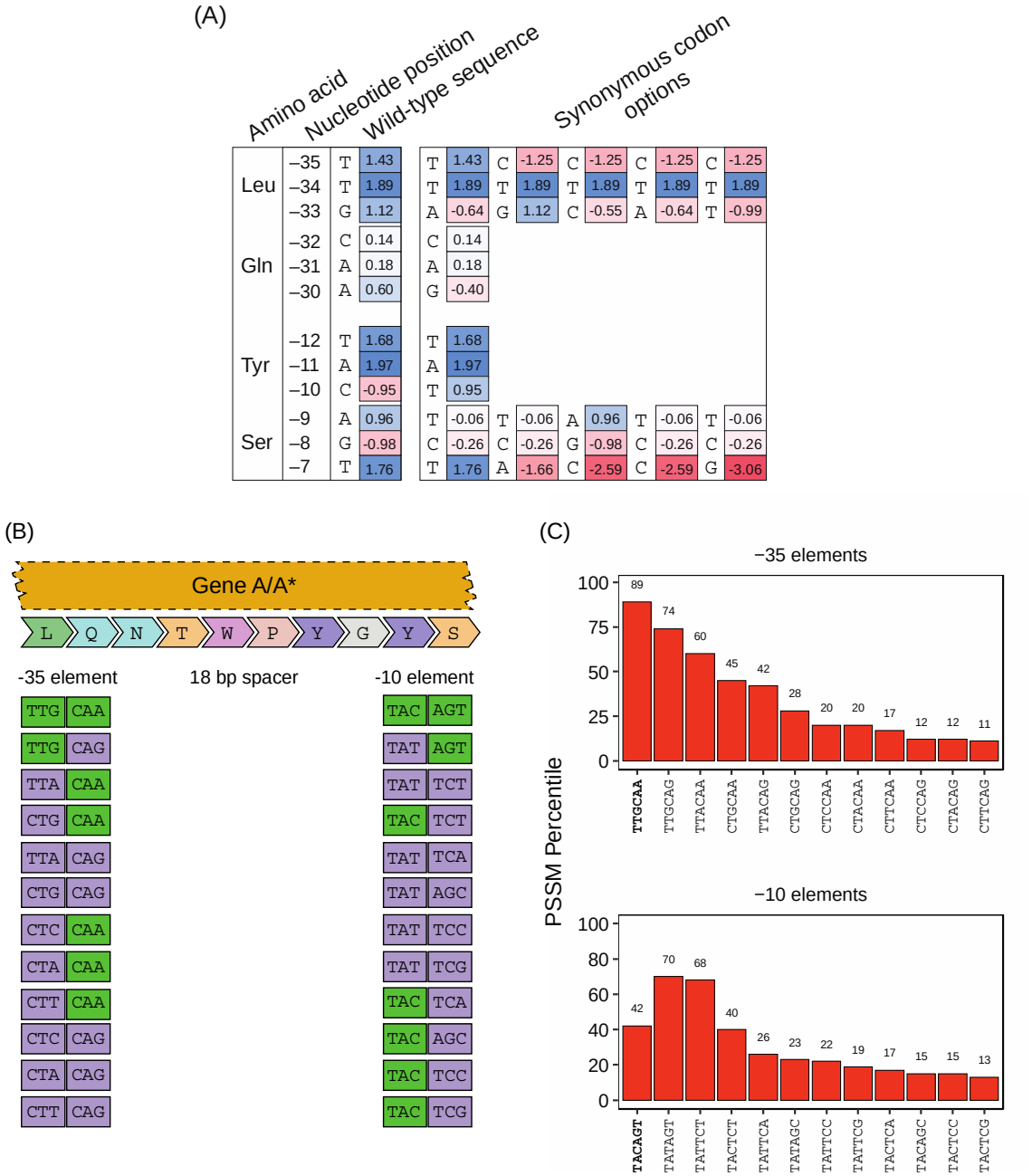

Figure 1: CORPSE algorithm. (A) List of all possible CORPSE modifications for *pB* with their PSSM score noted. Colors denote calculated score within each position in PSSM, high to low scores are shown as blue to red gradient. (B) List of all selected CORPSE variants used to eliminate *pB* activity with wild-type codons shown in green and variants in purple. Codons shown corresponded to translated *pB* sequence within ϕX174 genes A/A* (C) PSSM percentile for each possible variant with percentile shown above each sequence. The wild-type sequences are shown in bold.

To test the CORPSE algorithm's ability to reduce or eliminate a σ^{70} promoter, we chose the intragenic promoter *pB* within the bacteriophage ϕX174 genes A/A* as a model [26]. We used

the CORPSE algorithm to generate a list of 11 possible alternative sequences each for the –35 and –10 promoter elements that would be expected to reduce promoter activity while maintaining synonymous codons of overlapping coding sequences A/A* (Figure 1A and 1B). The –35 element synonymous codon PSSM scores ranged from -0.43 to 4.36 which are in the 25th - 70th percentiles of all *E. coli* σ^{70} promoters (Figure S1A), while the unmodified sequence score (5.36) is in the 89th percentile. The –10 element PSSM scores ranged from -0.68 to 6.34, placing them in the 10th - 70th percentiles of all *E. coli* promoters, while the unmodified sequence score (4.44) was in the 40th percentile (Figure S1B). We generated promoter elements using CORPSE that reduced the score to less than 2 (33rd percentile) for the –35 and less than 1 (18th percentile) for the –10 element, respectively (Figure 1C). We combined these selected elements combinatorically to generate 16 variants to examine further (Table S1).

To determine if CORPSE mutations reduced the activity of the *pB* promoter, the 16 variants were assembled into the pJ804 plasmid upstream of superfolder GFP (sfGFP). The plasmid also contained mCherry in the reverse direction to sfGFP under control of a constitutive promoter (Figure 2A) for expression capacity normalization [27].

Plasmids were built, transformed into *E. coli*, and their fluorescence measured using flow cytometry. The median sfGFP values for each variant were normalized to mCherry and compared to the wild-type construct. In all variants, sfGFP expression was significantly reduced to 0.9 % -5.3 % of wild-type (p values < 0.005) (Figure 2B). The CORPSE alterations to *pB* were highly effective as the minimal mutations within CORPSE-01, a $^{-9}$AGT^{-7} to $^{-9}$TCC^{-7} modification, were disruptive enough to remove 95 % of the promoter's function. These results were not unexpected as the two most conserved sequences in the –10 element, $^{-11}$A and $^{-7}$T, are found 88.5 % and 80.2 % of the time respectively in σ^{70} promoters [25]. These two nucleotides base flip out of the DNA stack during isomerisation and bind with unique pockets within the σ factor.

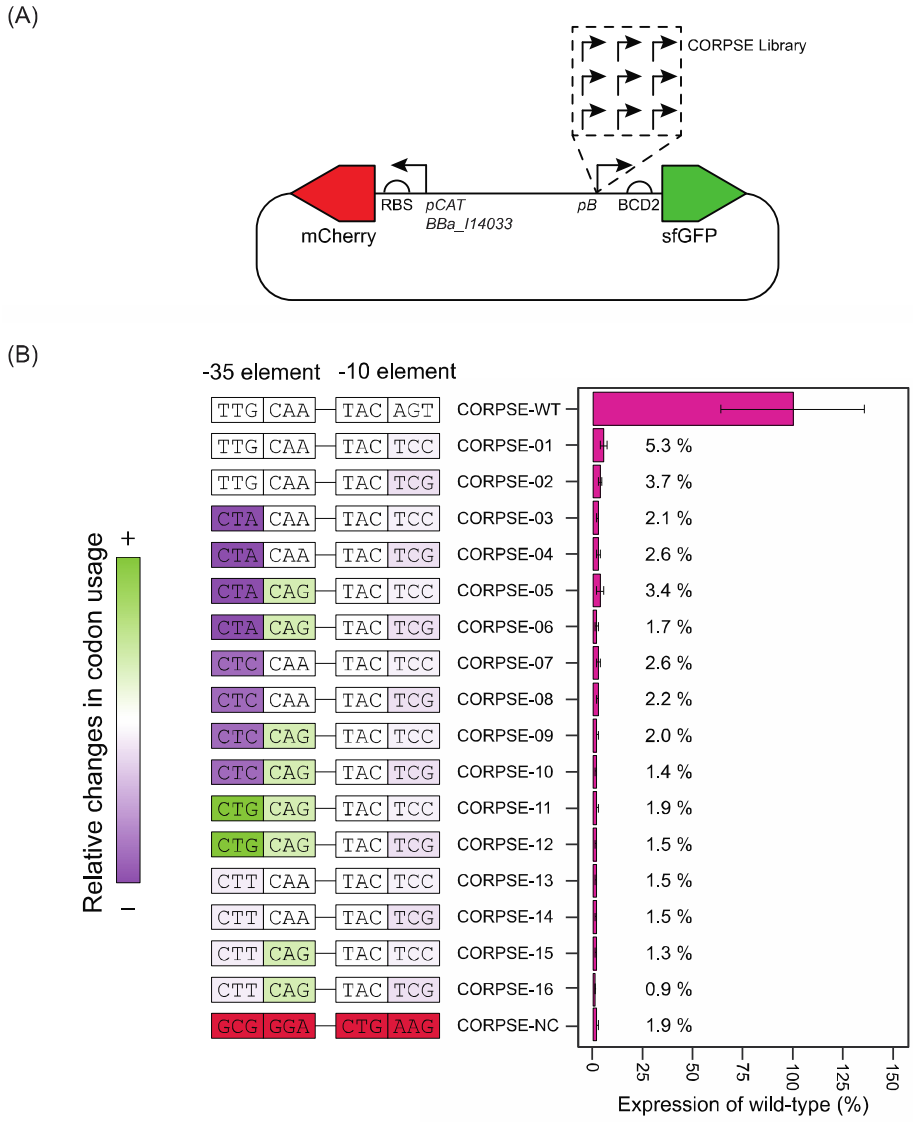

Figure 2: Low scoring promoter variants dramatically reduce *pB* promoter activity, while only making synonymous changes to overlapping codons. (A) CORPSE reporter plasmid contains constitutively expression mCherry as an internal control while wild-type and CORPSE mutants of *pB* driven variable sfGFP expression. (B) All mutations to *pB* reduced activity of the promoter within the reporter plasmid with the minimal changes to the –10 element in CORPSE-02 displaying only 5.3 % of wild-type activity. Broader modifications to the sequence changing both promoter elements further decreased expression with normalized expression being reduced to 0.9 % of wild-type. CORPSE-NC represents a negative control with selected nucleotides being worse scoring choices at each position, ignoring codon restraints. Normalised sfGFP compared to wild-type is shown as %.

The likely driver behind the reduced expression was the altered –7 base which canonically binds within an hydrophilic pocket in the σ factor which disallows pyrimidine

nucleotides binding and structurally excludes $^{-7}$C bases, a mutation which was present in half of the CORPSE variants (Figure 2B) [17]. While the –11 base is a useful target for rational engineering due to its binding constraints [28], the synonymous codons for tyrosine (TAT or TAC) within the gene A/A* sequence overlapping the promoter only allowed the third nucleotide to vary, and changing the C to a T would have increased, not decreased promoter strength.

While the –7 modifications likely drove expression changes, the non-base flipping nucleotides, –10 to –8, still bind to the σ factor through sugar-phosphate backbone interactions, as well as potential $^{-8}$A nucleotide van der Waal force contact with the σ factor [17]. These structural interactions bring potential for further reductions in ideal DNA-σ factor interactions, however, are likely limited as they interact via their backbone, not nucleotide structure.

CORPSE can silently eliminate *trpCp* intragenic promoter without affecting overlapping *trpD* gene function

To determine if this method could be more broadly applied to *E. coli* refactoring, and to test if the codon changes affect the overlapping gene function, we applied the CORPSE method to the *trpCp* promoter within the tryptophan biosynthesis operon (Figure 3A). The *trpCp* promoter is located within the *trpD* gene sequence [12] and provides additional transcriptional current to drive basal expression of the *trpCBA* genes [12, 29]. The wild-type sequence of *trpCp* has a –35 and –10 PSSM score of 3.2 and 1.6 respectively which placed them within the 32nd and 21st percentiles (Figure S1A and S1B), corresponded with the known low strength of *trpCp* [12, 29]. We applied the CORPSE method to generate a set of all possible variants that would significantly reduce the promoter strength while retaining the same amino acids of the overlapping TrpD protein.

We selected one variant which reduced the –35 and –10 scores to the 1st and 4th percentiles respectively for further testing. We assembled the wild-type (pGERC::*trpCp-WT*) and variant *trpD* (pGERC::*trpCp-V1*) gene sequences into the pGERC plasmid upstream of a sfGFP gene to probe the expression effects of our modifications (Figure 3B). Again, an mCherry gene under constitutive promoter was located in the opposite orientation on the plasmid as expression normalization control. The two strains were analysed using flow cytometry and the results showed pGERC::*trpCp-V1* displayed an 81 % reduction in sfGFP fluorescence compared to pGERC::*trpCp-WT* (Figure 3C). As with the *pB* variants, the pGERC::*trpCp-V1* variant contained minimal changes to the wild-type sequence with only two nucleotide modifications in the –35 element and three modifications in the –10 element.

To test whether the synonymous mutations made to disrupt the *trpCp* promoter had any deleterious effects on the overlapping *trpD* gene function, we assembled the wild-type and modified *trpD* genes into a pQE60 plasmid (Figure 3D) and attempted to complement a strain from the KEIO collection with a *trpD* disruption (JW1255-1). The two complemented strains, along with an empty pQE60 plasmid control, were grown overnight in LB media followed by plating on M9 minimal media lacking tryptophan. After overnight growth we observed identical growth patterns for strains pQE60::*trpCp-WT*) and (pQE60::*trpCp-V1*) on both LB and tryptophan deficient M9 minimal media (Figure 3E) demonstrating that modified *trpD* was functionally equivalent to wild-type *trpD* under these conditions. While both complemented strains grew on M9 minimal medium lacking tryptophan, the strain containing the empty plasmid control did not grow.

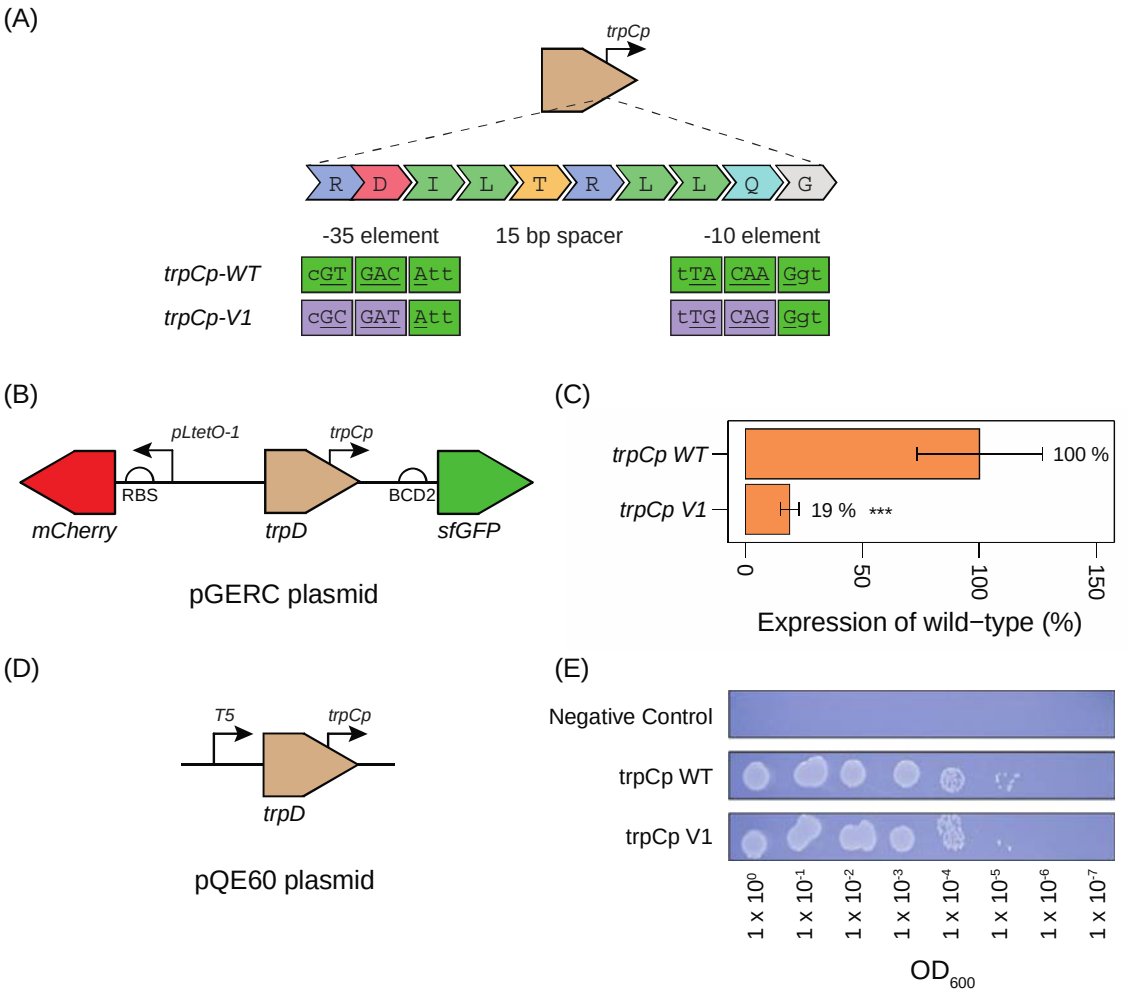

Figure 3. Wild-type and CORPSE variant supplementation of *trpCp* containing *trpD* constructs reduced reporter expression and maintained gene functionality. (A) The *trpCp* promoter within *trpD* was modified through CORPSE to eliminate the promoter. Wild-type codons shown in green and synonymous mutations in purple. Promoter hexamers underlined within codons. (B) Wild-type and CORPSE V1 *trpD* genes were inserted upstream of *sfGFP* in a pGERC plasmid to drive reporter expression. (C) CORPSE modification of *trpCp* successfully reduced reporter expression by 81 % when normalized to internal mCherry expression with p values < 0.005 shown as ***. (D) Wild-type and CORPSE V1 *trpD* gene were inserted into qQE-60 plasmids and expressed in the Δ*trpD* KEIO collection strain JW1255-1 to assess gene function effects of the modification. (E) Overnight cultures of empty pQE60 and the two *trpD* constructs were grown on tryptophan deficient M9 media after normalization and serial dilution. No phenotypic effects were detected from the *trpCp* modifications as no growth defects were apparent.

While wild-type *trpCp* already contained a non-optimal nucleotide in the –7 position, $^{-7}$G, the promoter contained a consensus $^{-11}$T. We were able to break this consensus sequence by altering $^{-13}$TTA^{-11} to $^{-13}$TTG^{-11} in the –10 element. The nucleotide with the largest predicted effect was the alteration of $^{-37}$AGT^{-35} codon to $^{-37}$AGC^{-35}. This in theory affected the interactions of the σ subunit to the nucleotide and backbone atoms at the –36 and –35 positions [18]. In addition to significantly reducing sfGFP expression, when the wild-type and variant *trpD* genes were expressed in a Δ*trpD* cell strain, no changes in cell growth were detected when grown on minimal media. This suggests that the specific synonymous mutations used had no effect on TrpD functionality.

Inverted CORPSE (iCORPSE) can silently create new promoters within gene sequences

We next tested the feasibility of inverting CORPSE to create a functional promoter within a coding sequence without disturbing the function of the overlapping gene. We first used the promoter prediction software BPROM [30] to analyse the complementary strand of the mCherry gene to identify any sequences already displaying promoter-like characteristics that could be modified to generate a *de novo* promoter directing transcription in the opposite direction to mCherry and its promoter. Two promoters were predicted through that software, designated PROM-1 and PROM-2 (Figure 4A). In our experience the BPROM software often predicts sites that are not actually promoters *in vivo* in addition to identifying sites that are true promoters[26]. We also manually identified a site (PROM-3) with two hexameric sequences separated by 17 bp that could be synonymously modified towards the σ^{70} consensus sequence (Figure 4A and B, and Table 1). Analysis of the proto-promoters showed they had a range of PSSM scores that put the –35 elements in the 10th – 35th percentiles and the –10 elements in the 4th – 95th percentiles of *E. coli* σ^{70} promoters (Table 1).

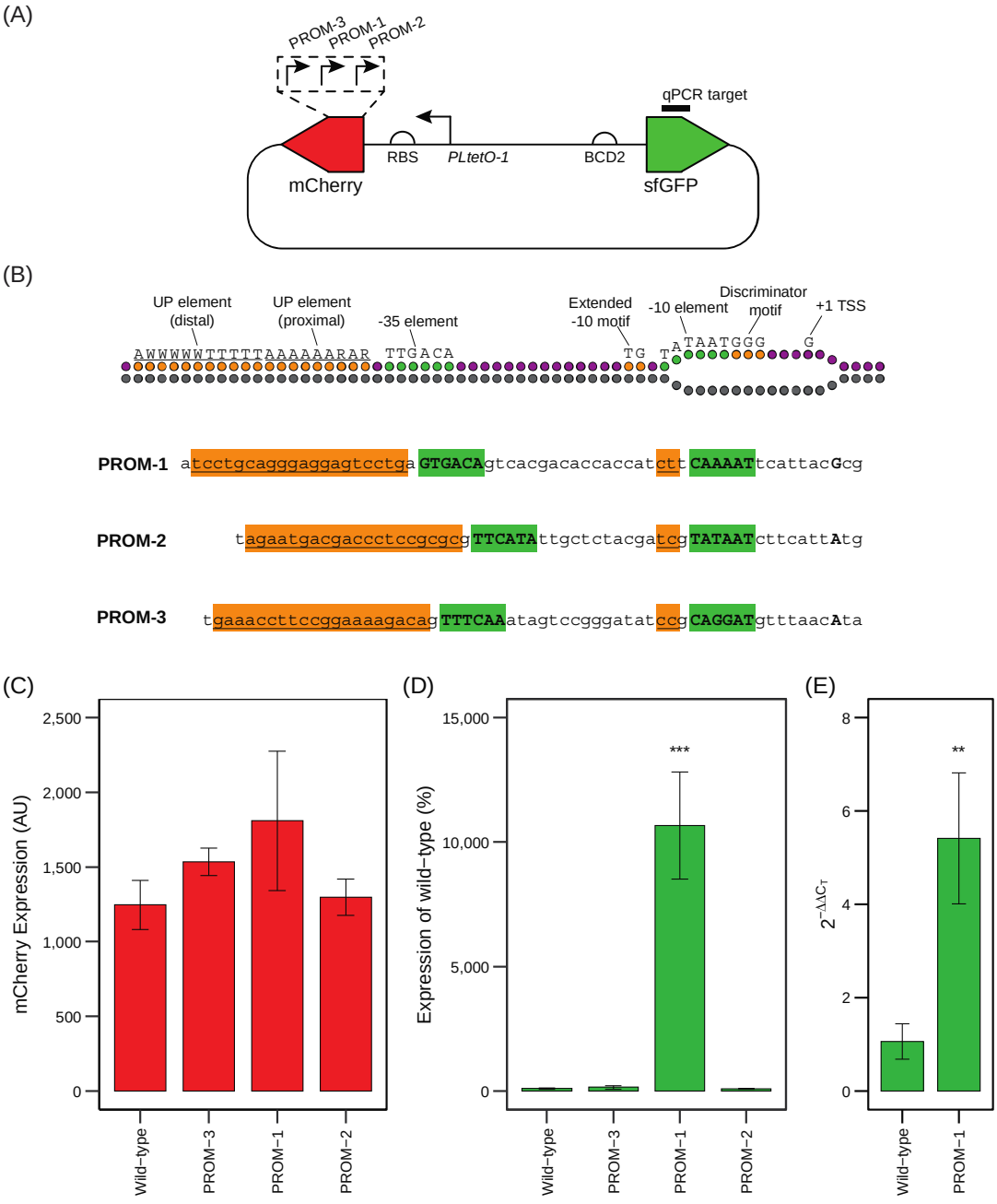

Fig. 4: iCORPSE driven sfGFP expression. (A) Reporter plasmid for iCORPSE has constitutively driven mCherry expression and variable sfGFP expression driven by synonymous mutations within the mCherry antisense reading frames. Three potential promoters were engineered into mCherry to drive sfGFP expression. The qPCR target for sfGFP is shown. (B) Engineered iCORPSE sequence aligned against structural map of σ^{70} promoter sequences. Core binding elements shown in green and accessory binding elements shown in yellow. (C) Fluorescence from sfGFP expression across mCherry variants compared to wild-type mCherry sequence. (D) Mean mCherry expression and standard deviation. (E) Log2 fold-change results for sfGFP expression normalized against housekeeping cysG. P values < 0.01 and <0.001 are shown as ** and ***, respectively.

Table 1: iCORPSE sequences

Promoter variant	–35 element	–10 element	PSSM score –35	Percentile	PSSM score –10	Percentile	PSSM score combined
PROM-1	GTGACG (wt)	CAAAAT (wt)	2.24	37th	4.07	39th	6.31
	GTGACA	CAAAAT	3.24	54th	4.07	39th	7.31
PROM-2	CTCATA (wt)	TATAAT (wt)	1.27	24th	8.35	95th	9.62
	TTCATA	TATAAT	3.95	65th	8.35	95th	12.3
PROM-3	CTTCAG (wt)	CTGGGT (wt)	-0.43	10th	-7.37	4th	-7.80
	TTTCAA	CAGGAT	3.25	54th	1.30	19th	4.55

Next, we used the CORPSE algorithm to generate a list of synonymous codon changes increasing the cumulative PSSM score across both elements, followed by modifying the mCherry sequence within the pGERC plasmid [31] (Figure 4A) to reflect the recoded PROM-1, PROM-2, and PROM-3 sequences. Measurements of mCherry and sfGFP fluorescence from cells carrying the plasmids showed PROM-2 and PROM-3 modifications resulted in no detectable change to sfGFP expression compared to the control plasmid (Figure 4D). In contrast, PROM-1 dramatically increased sfGFP expression to 10,664 % over the control plasmid (Figure 4D). The synonymous codon changes used to create the promoter had no significant (p value ≤ 0.01) effect on mCherry fluorescence (Figure 4C). To confirm the increased sfGFP expression was due to increased transcription of the sfGFP gene we extracted RNA and performed qPCR with a primer set targeted to the sfGFP coding sequence (Figure 4E). RNA abundance of the sfGFP transcript was found to be 5.4-fold increased when the modified PROM-1 sequence was present in mCherry versus the wild-type mCherry sequence (Figure 4E).

We cannot fully explain the differences in sfGFP expression from the different iCORPSE variants using the CORPSE scoring scheme. All iCORPSE sequences contain the crucial –10 element nucleotides $^{-11}$A and $^{-7}$T bases (Figure 4B). In addition, no iCORPSE sequence displays the canonical A/T tract repeats of an UP region [32] or has a canonical extended –10 TGn motif [33] and therefore has not clear enhancer motifs. The difference in spacer length may explain the differential effects in promoter activity because BPROM-1

contains 19 bp, BPROM-2 contains 13 bp, and PROM-3 contains a 17 bp spacer. Through a spacer sequence explanation, the extended BPROM-1 spacer could permit more dynamism in RNAP binding which was more restricted in the other two iCORPSE variants [34].

Most attempts to predict strength from sequence identities have utilised statistical methods with mixed success [35-37]. The overall biological complexity of the system complicates prediction which in the future may be eliminated by large measurement sets and machine learning approaches that are only now becoming available [38-41].

A future direction for synthetic genomics research will be to progress from ordered genome architectures where all regulatory elements and coding domains are separate from each other, and towards compressed architectures with overlapping coding and regulatory sequences. The advantages of a compressed architecture would be reducing overall genome size [42], and translationally couple gene expression [43, 44] permitting greater genetic content within a reduced length. Intragenic overlapping promoters could potentially shield them from mutation but any changes created with synonymous changes could be easily undone by synonymous mutation.

In conclusion, in this study we used a rational system to eliminate function from a promoter internal to a known coding domain through ranked synonymous codon modifications. With the CORPSE method we demonstrated the 99 % reduction of *pB* activity within a reporter system without altering amino acid sequence of the overlapping gene. The CORPSE method was also applied to the intragenic promoter *trpCp* within *trpD* and eliminated 81 % of *trpCp* activity. We also inverted the CORPSE ranking system to create novel reverse promoters internal to the mCherry coding sequence. In one of our constructs, iCORPSE increased sfGFP reporter expression by > 10,000 %.

The CORPSE system shows promise in the future refactoring of complex genetic circuits by eliminating intragenic promoter sequences without potentially deleterious large codon usage changes which facilitates increased standardisation of parts. The CORPSE system

has also demonstrated a future application in creating compressed and theoretically evolutionarily robust circuits through the embedding of a reverse promoter sequence internal to an existing coding sequence.

Materials and Methods

Calculating promoter frequencies

All σ^{70} promoter sequences present in the database were extracted from RegulonDB [25] (http://regulondb.ccg.unam.mx) via the σ70 sigmulon dataset. The sequence datasets were manually curated to remove non-hexameric promoter sequences. Non-hexameric sequences accounted for 8.5 % and 4.9 % of the –35 and –10 datasets, respectively. The –35 and –10 elements were extracted and used to calculate individual position specific scoring matrices (PSSM) [45] to generate log odds scores for each nucleotide in a position within the element, these log odds score were then reported as PSSM scores. The calculations were performed by curating –35 and –10 elements to only retain hexameric sequences. The nucleotides in each position in the sequence were counted and divided by the total number of sequences. These ratios were then divided by 0.25 to account for random frequency of an individual nucleotide occurring in the position. The normalised ratios had their binary logarithm values calculated to determine the log odds score of each nucleotide in the matrix. The frequency and PSSM (log odds) scores of each nucleotide at each position is presented in Supplementary File S1.

Codon Restrained Promoter Silencing (CORPSE) algorithm

The CORPSE algorithm consisted of the following steps: (1) calculate log odd scores for hexameric –10 and –35 sequences, (2) separate sequences into codon triplets, (3) determine synonymous codon changes possible at each location, (4) combinatorically assemble all possible synonymous codon options, (5) calculate all log odds scores for each sequence, (6) output ranked sequences by log odd score.

Log odd scores used as a proxy for predicted promoter strength by assuming that a higher score contained more nucleotides required to facilitate the σ^{70}-DNA interaction driving transcription. The CORPSE algorithm for *E. coli* is available as a webtool (https://bio-tools.com.au/promoter.html).

Bacterial strains and plasmids

Three plasmid constructs were used throughout this study: pJ804 [8] used as the basis for CORPSE, pGERC [31] for the *trpCp* and iCORPSE studies, and pQE60 for the phenotypic analysis of *trpCp* modifications. All strains were grown in lysogeny broth (LB) and supplemented with either 100 µg/mL carbenicillin (pJ804), or 50 µg/mL kanamycin (pGERC) depending on which plasmid was required All plasmids were originally transformed and sequence verified in NEB Turbo (New England Biosciences, #C2984H). Plasmids were purified and transformed into the one of three bacterial strains if required. The NCTC122 *E. coli* strain, was used for all plasmids used in the CORPSE experimentation using TSS heat transformation [46]. The pQE60::*trpCp* variants were transformed into the Δ*trpD* JW1255-1 strain (BW25113) from the KEIO collection [47] using the *Mix & Go!* protocol from Zymo Research (Zymo Research, #T3001).

Designing CORPSE promoter constructs

Wild-type and CORPSE-designed promoter sequences were analysed in a reporter plasmid (pJ804) which drove sfGFP expression under BCD2 control [44] whilst mCherry under pCAT (BBa_I14033) constitutive expression with a consensus RBS sequence acted as an internal control. All promoter sequences from CORPSE analyse were embedded with 5` primer overhangs containing the variant promoter elements and native spacer sequence for assembly through round the horn (RTH) PCR [44, 48, 49] with a universal reverse phosphorylated primer. RTH variant primers were phosphorylated with NEB T4 Polynucleotide kits (New England Biosciences, #M0201) and 10 µM of RTH primers were used to amplify and mutate pJ804 backbone with Q5 Hot Start HF 2x Master Mix (New England Biosciences, #M0494). Template DNA was digested via DpnI (New England Biosciences, #R0176). Linear mutated DNA was

ligated with T4 ligase (New England Biosciences, #M0202). Ligated plasmids were heat transformed into TSS competent cells [46] and recovered at 37 °C for 1 h. All recovered cells were centrifuged at 6,000 RCF for 5 min (Eppendorf 5424), resuspended in 100 µL of media, and plated onto selective media.

Designing *trpCp* promoter constructs

The *trpCp* sequence within *trpD* was analysed through CORPSE which generated two alternative promoter sequences. Only one of these sequences reduced the CORPSE score and was selected as the variant sequence for experimentation. Two gene constructs for *trpD* were synthesized (IDT), one containing the wild-type sequence (*trpCp-WT*) and one containing the variant sequence (*trpCp-V1*). The *trpD* sequences were inserted upstream of *sfGFP* in the pGERC plasmid immediately adjacent to the BCD2 RBS sequence, and immediately downstream of the T5 promoter in the pQE60 plasmid.

Designing iCORPSE promoter constructs

The reverse complement sequence of mCherry was analysed through BPROM bacterial promoter prediction software [30] and two promoter sites were predicted (BPROM-1, and BPROM-2). These, along with two hexameric sequences separated by 17 nt and chosen for their alternative codons being similar to the σ^{70} consensus sequence (PROM-3), were analysed using the CORPSE method and the strongest promoter sequences were selected.

Using the pGERC plasmid [31, 44], separate DNA fragments were designed to remove the sfGFP EM7 promoter to silence sfGFP expression, and the four iCORPSE mutations were added (wild-type, V1, V2, and V3). These were assembled via NEBuilder HiFi DNA Assembly Master Mix (New England Biosciences, #E2621) and transformed into NEB Turbo competent cells (New England Biosciences, #C2984H), plasmid purified (QIAgen, #27104), and transformed into *E. coli* C strain NCTC122 TSS competent cells via the method described above.

Flow Cytometry

Sequence confirmed strains were isolated as single colonies on LB agar plates (2 mM carbenicillin) and 1 mL cultures were grown overnight in a 96 well, square-welled, round bottomed deep well plate at 40 % liquid volume and sealed with a Breathe-easy sealing membrane (Merck, #2380059-1PAK). Overnight growth occurred in an Infors MT multitron pro shaker at 250 RPM rotating orbitally at 25 mm diametres at 37 °C for 17 h.

The density of 200 µL overnight cells were measured on a BMG Labtech SPECTROstar Nano plate reader with a CellStar clear 96 well plate. Cells were passaged to an OD600 of 1 in 1 mL and then diluted 1:100 into another deep well plate.

Cells were grown for 5 h to reach mid-log growth and the OD600 of 200 µL was measured on a BMG Labtech SPECTROstar Nano plate reader to confirm growth stage.

A further 200 µL was aliquoted into a CellStar clear 96 well plate and centrifuged in Eppendorf 5430R centrifuge with the plate swing bucket attachment at 2,240 RCF for 10 min to pellet cells. LB supernatant was removed and the cells resuspended in 1x phosphate buffered saline (PBS) media, and diluted 1:200 into 1x PBS media in a CellStar clear 96 well plate. Bacterial growth parameters have been reported to the best of our knowledge conforming with the MIEO v0.1.0 standard [50].

Fluorescence Measurements

Fluorescence readings were performed as previously [51] on a Beckman Coulter CytoFLEX S (FITC, 488 nm excitation laser, 525/40 nm emission band-pass filter). 10,000 events were collected with acquisition settings of: FSC – 264, SSC – 2000, FITC – 299, and ECD – 150. Data was processed using CytExpert and data analysis conducted with median FITC and ECD scores and their generated standard deviations. FITC scores were normalized against ECD scores to create sfGFP/mCherry expression ratios, which were then divided against wild-type ratios to determine percentages of expression.

RNA purification

RNA was purified from 20 mL *E. coli* cultures at early growth states (30 min post inoculation) using the RNeasy Mini kit (Qiagen: #74106) according to the manufacturer's instructions, with the optional DNase on-column digestion step (Qiagen: #79254). Cultures were grown in an Infors MT multitron pro shaker at 250 RPM rotating orbitally at 25 mm diametres at 37 °C for both 17 h overnight step and for 30 min growth step.

Quantitative PCR analysis

Purified RNA samples were reverse transcribed (RT) into cDNA according to the manufacturer's instructions (ThermoFisher Scientific, #4368814). qPCR of the RT cDNA was performed according to the manufacturer's instructions in a LightCycler 480 II (Roche Life Science using SYBR GREEN (Roche Life Science, #04707516001) with a 10 µL total reaction volume. *sfGFP* RNA expression was quantified using primers (F: GGTGAAGGTGACGCAACTAATGGTA, R: TTGGCCGACTCTGGTAACGACGCTG) normalized to the housekeeping gene *cysG* (F: GAAAGCCTTCTCGACACCTG, R: CGTTACAGAAGATGCGACGA) using the $2^{-\Delta\Delta C}{}_{T}$ method [52].

TrpD phenotype assay

Cultures of pQE60 containing strains (pQE60::*trpCp-WT*, pQE60::*trpCp-V1*, pQE60::*empty*) were grown overnight in an Infors MT multitron pro shaker at 250 RPM rotating orbitally at 25 mm diametres at 37 °C for 17 h in LB selective media. All cultures were normalized to an OD_{600} of 1 and serial diluted ten-fold until a final concentration of 1×10^{-8} was achieved. Dilutions were pipetted (2 µL aliquots) onto pre-warmed (37 °C) M9 agar plates. Plates were incubated overnight at 37 °C and assessed for strain growth.

Acknowledgements

We recognize that the intellectual and physical labour of this research was conducted on the traditional lands of the Wattamattagal clan of the Darug nation and of the Gadigal and

Wangal peoples of the Eora nation. We thank Thomas Williams for his suggestion to try inverting the CORPSE system to create new promoters via iCORPSE and Hannah Zhu for helpful suggestions to improve the manuscript.

Funding

DYL is a recipient of the Macquarie University Research Excellence PhD scholarship (MQRES) and CSIRO PhD Scholarship Program in Synthetic Biology (Synthetic Biology Future Science Platform). PRJ was supported by the Molecular Sciences Department, Faculty of Science & Engineering, the Deputy Vice-Chancellor (Research) of Macquarie University, and NHMRC Ideas Grant APP1185399.

Conflicts of interest

The authors declare no competing conflicts of interest.

References

1. Müller, K. M.; Arndt, K. M., Standardization in Synthetic Biology. In *Synthetic Gene Networks: Methods and Protocols*, Weber, W.; Fussenegger, M., Eds. Humana Press: Totowa, NJ, 2012; pp 23-43.
2. Vecchio, D. D.; Dy, A. J.; Qian, Y., Control theory meets synthetic biology. *Journal of The Royal Society Interface* **2016,** *13* (120), 20160380.
3. Gorochowski, T. E.; Espah Borujeni, A.; Park, Y.; Nielsen, A. A.; Zhang, J.; Der, B. S.; Gordon, D. B.; Voigt, C. A., Genetic circuit characterization and debugging using RNA-seq. *Mol Syst Biol* **2017,** *13* (11), 952.
4. Smanski, M. J.; Bhatia, S.; Zhao, D.; Park, Y.; B A Woodruff, L.; Giannoukos, G.; Ciulla, D.; Busby, M.; Calderon, J.; Nicol, R.; Gordon, D. B.; Densmore, D.; Voigt, C. A., Functional optimization of gene clusters by combinatorial design and assembly. *Nature Biotechnology* **2014,** *32* (12), 1241-1249.
5. Temme, K.; Zhao, D.; Voigt, C. A., Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** *109* (18), 7085-90.
6. Chan, L. Y.; Kosuri, S.; Endy, D., Refactoring bacteriophage T7. *Mol Syst Biol* **2005,** *1*, 2005.0018.
7. Ghosh, D.; Kohli, A. G.; Moser, F.; Endy, D.; Belcher, A. M., Refactored M13 bacteriophage as a platform for tumor cell imaging and drug delivery. *ACS synthetic biology* **2012,** *1* (12), 576-582.
8. Jaschke, P. R.; Lieberman, E. K.; Rodriguez, J.; Sierra, A.; Endy, D., A fully decompressed synthetic bacteriophage PhiX174 genome assembled and archived in yeast. *Virology* **2012,** *434* (2), 278-284.

9. Brophy, J. A.; Voigt, C. A., Principles of genetic circuit design. *Nature methods* **2014,** *11* (5), 508.
10. Song, M.; Sukovich, D. J.; Ciccarelli, L.; Mayr, J.; Fernandez-Rodriguez, J.; Mirsky, E. A.; Tucker, A. C.; Gordon, D. B.; Marlovits, T. C.; Voigt, C. A., Control of type III protein secretion using a minimal genetic system. *Nature Communications* **2017,** *8* (1), 14737.
11. Guo, M. S.; Updegrove, T. B.; Gogol, E. B.; Shabalina, S. A.; Gross, C. A.; Storz, G., MicL, a new σE-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes & development* **2014,** *28* (14), 1620-1634.
12. Horowitz, H.; Platt, T., Identification of trp-p2, an internal promoter in the tryptophan operon of *Escherichia coli*. *J Mol Biol* **1982,** *156* (2), 257-67.
13. Thomason, M. K.; Bischler, T.; Eisenbart, S. K.; Förstner, K. U.; Zhang, A.; Herbig, A.; Nieselt, K.; Sharma, C. M.; Storz, G., Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in *Escherichia coli*. *Journal of Bacteriology* **2015,** *197* (1), 18-28.
14. Urtecho, G.; Insigne, K. D.; Tripp, A. D.; Brinck, M.; Lubock, N. B.; Kim, H.; Chan, T.; Kosuri, S., Genome-wide Functional Characterization of *Escherichia coli* Promoters and Regulatory Elements Responsible for their Function. *bioRxiv* **2020**, 2020.01.04.894907.
15. Hook-Barnard, I. G.; Hinton, D. M., Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. *Gene regulation and systems biology* **2007,** *1*, 117762500700100020.
16. Murakami, K. S.; Darst, S. A., Bacterial RNA polymerases: the wholo story. *Current opinion in structural biology* **2003,** *13* (1), 31-39.
17. Feklistov, A.; Darst, S. A., Structural basis for promoter– 10 element recognition by the bacterial RNA polymerase σ subunit. *Cell* **2011,** *147* (6), 1257-1269.
18. Lane, W. J.; Darst, S. A., The structural basis for promoter– 35 element recognition by the group IV σ factors. *PLoS biology* **2006,** *4* (9).
19. Chevance, F. F. V.; Hughes, K. T., Case for the genetic code as a triplet of triplets. *Proceedings of the National Academy of Sciences* **2017,** *114* (18), 4745-4750.
20. Frumkin, I.; Lajoie, M. J.; Gregg, C. J.; Hornung, G.; Church, G. M.; Pilpel, Y., Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences* **2018,** *115* (21), E4940-E4949.
21. Yu, C.-H.; Dang, Y.; Zhou, Z.; Wu, C.; Zhao, F.; Sachs, Matthew S.; Liu, Y., Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell* **2015,** *59* (5), 744-754.
22. Hanson, G.; Coller, J., Codon optimality, bias and usage in translation and mRNA decay. *Nature reviews Molecular cell biology* **2018,** *19* (1), 20-30.
23. Goodman, D. B.; Church, G. M.; Kosuri, S., Causes and effects of N-terminal codon bias in bacterial genes. *Science* **2013,** *342* (6157), 475-9.
24. Mittal, P.; Brindle, J.; Stephen, J.; Plotkin, J. B.; Kudla, G., Codon usage influences fitness through RNA toxicity. *Proceedings of the National Academy of Sciences* **2018,** *115* (34), 8639-8644.
25. Santos-Zavaleta, A.; Salgado, H.; Gama-Castro, S.; Sánchez-Pérez, M.; Gómez-Romero, L.; Ledezma-Tejeida, D.; García-Sotelo, J. S.; Alquicira-Hernández, K.; Muñiz-Rascado, L. J.; Peña-Loredo, P., RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic acids research* **2019,** *47* (D1), D212-D220.
26. Logel, D. Y.; Jaschke, P. R., A high-resolution map of bacteriophage ϕX174 transcription. *Virology* **2020,** *547*, 47-56.
27. Ceroni, F.; Algar, R.; Stan, G.-B.; Ellis, T., Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nature Methods* **2015,** *12* (5), 415-418.

28. Tsujikawa, L.; Strainic, M. G.; Watrob, H.; Barkley, M. D.; deHaseth, P. L., RNA polymerase alters the mobility of an A-residue crucial to polymerase-induced melting of promoter DNA. *Biochemistry* **2002,** *41* (51), 15334-15341.
29. Horowitz, H.; Platt, T., Initiation in vivo at the internal trp p2 promoter of *Escherichia coli*. *Journal of Biological Chemistry* **1983,** *258* (13), 7890-7893.
30. Solovyev, V., Automatic annotation of microbial genomes and metagenomic sequences. Salamov, A., Ed. Nova Science Publishers: Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Li, R. W., Ed.),, 2011; pp 61 - 78.
31. Kosuri, S.; Goodman, D. B.; Cambray, G.; Mutalik, V. K.; Gao, Y.; Arkin, A. P.; Endy, D.; Church, G. M., Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* **2013,** *110* (34), 14024-9.
32. Gourse, R. L.; Ross, W.; Gaal, T., UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. *Molecular microbiology* **2000,** *37* (4), 687-695.
33. Mitchell, J. E.; Zheng, D.; Busby, S. J. W.; Minchin, S. D., Identification and analysis of 'extended –10' promoters in *Escherichia coli*. *Nucleic Acids Research* **2003,** *31* (16), 4689-4695.
34. Typas, A.; Hengge, R., Role of the spacer between the– 35 and– 10 regions in σs promoter selectivity in *Escherichia coli*. *Molecular microbiology* **2006,** *59* (3), 1037-1051.
35. Alper, H.; Fischer, C.; Nevoigt, E.; Stephanopoulos, G., Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (36), 12678-12683.
36. De Mey, M.; Maertens, J.; Lequeux, G. J.; Soetaert, W. K.; Vandamme, E. J., Construction and model-based analysis of a promoter library for *E. coli*: an indispensable tool for metabolic engineering. *BMC Biotechnology* **2007,** *7* (1), 34.
37. Jensen, K.; Alper, H.; Fischer, C.; Stephanopoulos, G., Identifying functionally important mutations from phenotypically diverse sequence data. *Appl Environ Microbiol* **2006,** *72* (5), 3696-3701.
38. Travis La Fleur, A. H., Howard Salis Massively Parallel Development of a Predictive Model of Transcription Rate for Sigma70 Promoter Sequences. https://salislab.net/software/predict_promoter_calculator.
39. Wang, Y.; Wang, H.; Wei, L.; Li, S.; Liu, L.; Wang, X., Synthetic promoter design in *Escherichia coli* based on a deep generative network. *Nucleic Acids Research* **2020**.
40. Zhao, M.; Zhou, S.; Wu, L.; Deng, Y., Machine learning-based promoter strength prediction derived from a fine-tuned synthetic promoter library in *Escherichia coli*. *bioRxiv [Preprint]* **2020**, 2020.06.25.170365.
41. Van Brempt, M.; Clauwaert, J.; Mey, F.; Stock, M.; Maertens, J.; Waegeman, W.; De Mey, M., Predictive design of sigma factor-specific promoters. *Nat Commun* **2020,** *11* (1), 5822.
42. Sakharkar, K. R.; Sakharkar, M. K.; Verma, C.; Chow, V. T., Comparative study of overlapping genes in bacteria, with special reference to Rickettsia prowazekii and Rickettsia conorii. *International Journal of Systematic and Evolutionary Microbiology* **2005,** *55* (3), 1205-1209.
43. Johnson, Z. I.; Chisholm, S. W., Properties of overlapping genes are conserved across microbial genomes. *Genome Research* **2004,** *14* (11), 2268-2272.
44. Mutalik, V. K.; Guimaraes, J. C.; Cambray, G.; Lam, C.; Christoffersen, M. J.; Mai, Q. A.; Tran, A. B.; Paull, M.; Keasling, J. D.; Arkin, A. P.; Endy, D., Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat Methods* **2013,** *10* (4), 354-60.
45. Schneider, T. D.; Stephens, R. M., Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **1990,** *18* (20), 6097-6100.

46. Chung, C. T.; Miller, R. H., Preparation and storage of competent *Escherichia coli* cells. *Methods Enzymol* **1993,** *218*, 621-7.
47. Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K. A.; Tomita, M.; Wanner, B. L.; Mori, H., Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2006,** *2*, 2006.0008.
48. Ochman, H.; Gerber, A. S.; Hartl, D. L., Genetic applications of an inverse polymerase chain reaction. *Genetics* **1988,** *120* (3), 621-3.
49. Reisch, C. R.; Prather, K. L. J., Scarless Cas9 Assisted Recombineering (no-SCAR) in *Escherichia coli*, an Easy-to-Use System for Genome Editing. *Curr Protoc Mol Biol* **2017,** *117*, 31 8 1-31 8 20.
50. Hecht, A.; Filliben, J.; Munro, S. A.; Salit, M., A minimum information standard for reproducing bench-scale bacterial cell growth and productivity. *Communications Biology* **2018,** *1* (1), 219.
51. Vincent, R. M.; Wright, B. W.; Jaschke, P. R., Measuring Amber Initiator tRNA Orthogonality in a Genomically Recoded Organism. *ACS synthetic biology* **2019,** *8* (4), 675-685.
52. Livak, K. J.; Schmittgen, T. D., Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **2001,** *25* (4), 402-8.