Another 100 genes

Tom Röschinger[1], Grace Solini[1], Anika Nawar Choudhury[2], Stephen Quake[2, 3, 4], and Rob Phillips[1, 5, +]

[1]*Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA*
[2]*Chan Zuckerberg Biohub, San Francisco, CA 94158, USA*
[3]*Department of Bioengineering, Stanford University, Stanford, CA 94305, USA*
[4]*Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[5]*Department of Physics, California Institute of Technology, Pasadena, CA 91125, USA*
[+]*Correspondence: phillips@pboc.caltech.edu*

1 Abstract

2 Introduction

It has been more than sixty years since Jacob and Monod [1] shaped the way we think about transcriptional regulation in prokaryotes, yet, although more than one trillion bases have been stored in the NIH database (TR: find right citation format), we have yet to obtain a full understanding of how all the genes of a single organism are regulated. Even in the case of one of biology's best studied model organism, *Escherichia coli*, about two thirds of the genes lack any regulatory annotation (TR: add section to supp with details). For other prokaryotic model organisms the numbers are similar, while higher order model organisms such as *Saccharomyces cerevisiae* and *C. elegans* have close to no regulatory annotations, given the arguably more complex nature of gene regulation in eukaryotes (TR: also add section to supp for these organisms). Understanding how genes are regulated is required to understand how an organism adapts its physiology on short time scales to environmental stresses, as well as evolutionary adaption on long time scales. In addition, gene regulation networks and their building blocks, such as transcription factor binding sites and RNA polymerase (RNAP) promoters, are key elements in the design of synthetic gene circuits (TR: cite something here too, guess there is a ton. Repressilator?).

With its ever increasing availability, Next Gen Sequencing (NGS) is primed to be the method of choice to discover transcription factor and RNAP binding sites. A vast array of methods exists that make it possible to identify binding sites of either specific proteins (TR: cite) or for a broad spectrum of DNA binding factors (TR: cite). In methods like ChIP-Seq [2], proteins have to be cross linked to DNA, which does not work for all transcription factors, such as LacI in *E. coli* (TR: cite). While the resolution of these methods is ever improving, it does not allow for a nucleotide resolution yet (TR: cite), making it difficult to identify changes in binding affinity caused by single mutations. Other methods such as ATAC-seq [3, 4] and DNase-Seq [5] rely on open chromatin for binding site identification, and are therefore limited to mostly eukaryotic organisms (TR: look deeper for possible applications in bacteria, haven't found them yet). Another approach is to use RNA-seq as readout for mutagenised promoter regions, where binding sites are identified as regions that, when mutated, lead to significant increase or decrease in expression of a repressor gene [6–8].

Here we present the regulatory architecture of x (TR: depends on how many we end up showing) genes, including energy matrices with nucleotide resolution that make it possible to build thermodynamic models to predict gene expression [8–11]. Additionally, we present major improvements to the method called Reg-Seq [8], making further steps towards obtaining a method allowing to discover regulatory architectures genome wide. Reporter genes are chromosomally integrated into the *E. coli*

genome, and reduced diversity in mRNA stability lead to more precise identification of binding sites. A vast array of growth conditions is used to show how certain binding sites can only be identified in a certain growth condition, such as (TR: name example). The identification of transcription factors was moved away from laborious mass spectrometry experiments, using *in vitro* binding assays as well as a library of transcription factor knockout strains. Finally, improved computational analysis increases the speed of data analysis and the accuracy of parameters that are used for thermodynamic models (TR: here I am thinking Rosalinds stuff).

(TR: paragraph about scaling to 1000)

3 Methods

3.1 Promoter sequence import

3.2 Reporter construct design

3.3 Barcode Mapping

3.4 Genome Integration

4 Results

4.1 Genes studied

(TR: Does this belong into results or introduction?) In total we present the regulatory architecture of x promoters, which tells us how a total of y genes are regulated. 18 promoters were chosen as so called "gold standards". These genes have well annotated promoters and have been studied in detail in previous experiments [8, 10]. Including this set of genes allows us to compare the method presented in this work to previous iterations and verify the results, as well as find possible derivations or improvements. x promoters were chosen for genes that have been identified to have a high variation in protein copy number (TR: Probably should show that in the SI) across a set of 22 growth conditions by Schmidt et al., 2016 [12]. These genes were chosen since a high variation in copy number suggests that there are regulatory proteins controlling the expression of the gene. From the same dataset, a set of x promoters was chosen for genes with unidentified function, as annotated by the Schmidt et al., 2016 [12]. None of these promoters had any regulatory annotation prior to the experiments. Another set of x promoters were chosen for genes that were identified in EcoCyc as not having any functional annotation. Two groups of genes, so called iModulons [13] were chosen from the work of Lamoureux et al. 2021, where ca. 800 RNA-Seq datasets were evaluated to find genes that were regulated in a distinct network. (TR: Give details for iModulons and their function in their respective paragraphs?) The (TR: Give summary of all genes at the end of this paragragh.) (TR: Continue with genes of defined circuits and toxin/antitoxin genes)

4.2 Genome Integration of Reporters

A major improvement to the Reg-Seq method is the genome integration of the reporter using ORBIT (see supplementary information S5.1). By integrating the mutated promoter into the genome instead of keeping it on a plasmid, the variation in copy number per cell is reduced and the total number of reporters per cell is at an absolute minimum. Transcription factors that bind to the native version of the promoter in the chromosome also bind to the promoter copy, unless its binding site is heavily mutated. Hence, if many promoter copies are added to a cell, e.g. via plasmids, transcription

factors are more likely to bind the promoter of the reporter [14], leading to possible fitness defects of the cell. During the cloning process the reporters are carried on a plasmid with a copy number on the order of 10 (r6k replication origin). This effect can be seen in the results of the barcode mapping for the promoter of the gene *dicC*. In the promoter there is a binding site for DicA between the -9 and +11 positions (relative to the transcription start site), which was not identifiable in the previous iteration of Reg-Seq [8]. The sequences obtained for this promoter on our barcode mapping show an increased mutation rate in exactly the site predicted to be the transcription factor binding site, see Supplementary Figure S3. This means that sequences without mutations in this part of the promoter were not able to be cloned and lead to fitness defects, presumably due to the transcription factor being titrated away from the native promoter in the chromosome.

4.3 Barcode Mapping

4.4 Transcription Factor identification

4.5 Growth Conditions

4.6 Gold Standard genes

4.7 Ethanol iModulon

4.8 DNA damage repair iModulon

4.9 Antitoxin/Antibiotic genes

4.10 other y-ome genes

5 Discussion

- discuss how to scale to 1000 genes

6 To do list

- Write Introduction
- Collect references from reg-seq paper and new references
- write paragraphs about genes chosen
-

References

[1]F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins", Journal of molecular biology **3**, 318–356 (1961).

[2]H. S. Rhee and B. F. Pugh, "Chip-exo method for identifying genomic location of dna-binding proteins with near-single-nucleotide accuracy", Current protocols in molecular biology **100**, 21–24 (2012).

[3]J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf, "Atac-seq: a method for assaying chromatin accessibility genome-wide", Current protocols in molecular biology **109**, 21–29 (2015).

[4]Z. Li, M. H. Schulz, T. Look, M. Begemann, M. Zenke, and I. G. Costa, "Identification of transcription factor binding sites using atac-seq", Genome biology **20**, 1–21 (2019).

[5]A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-resolution mapping and characterization of open chromatin across the genome", Cell **132**, 311–322 (2008).

[6]G. Urtecho, A. D. Tripp, K. D. Insigne, H. Kim, and S. Kosuri, "Systematic dissection of sequence elements controlling σ70 promoters using a genomically encoded multiplexed reporter assay in escherichia coli", Biochemistry **58**, 1539–1551 (2018).

[7]G. Urtecho, K. D. Insigne, A. D. Tripp, M. Brinck, N. B. Lubock, H. Kim, T. Chan, and S. Kosuri, "Genome-wide functional characterization of escherichia coli promoters and regulatory elements responsible for their function", BioRxiv (2020).

[8]W. T. Ireland et al., "Deciphering the regulatory genome of escherichia coli, one hundred promoters at a time", Elife **9**, e55308 (2020).

[9]J. B. Kinney, A. Murugan, C. G. Callan Jr, and E. C. Cox, "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence", Proceedings of the National Academy of Sciences **107**, 9158–9163 (2010).

[10]N. M. Belliveau, S. L. Barnes, W. T. Ireland, D. L. Jones, M. J. Sweredoski, A. Moradian, S. Hess, J. B. Kinney, and R. Phillips, "Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria", Proceedings of the National Academy of Sciences **115**, E4796–E4805 (2018).

[11]S. L. Barnes, N. M. Belliveau, W. T. Ireland, J. B. Kinney, and R. Phillips, "Mapping dna sequence to transcription factor binding energy in vivo", PLoS computational biology **15**, e1006226 (2019).

[12]A. Schmidt et al., "The quantitative and condition-dependent escherichia coli proteome", Nature biotechnology **34**, 104–110 (2016).

[13]C. R. Lamoureux, K. T. Decker, A. V. Sastry, J. L. McConn, Y. Gao, and B. O. Palsson, "Precise 2.0-an expanded high-quality rna-seq compendium for escherichia coli k-12 reveals high-resolution transcriptional regulatory structure", BioRxiv (2021).

[14]R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips, "The transcription factor titration effect dictates level of gene expression", Cell **156**, 1312–1323 (2014).

Supplemental Information for: Whatever the title will be

Tom Röschinger[1], Grace Solini[1], Anika Nawar Choudhury[2], Stephen Quake[2, 3, 4], and Rob Phillips[1, 5, +]

[1]*Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA*
[2]*Chan Zuckerberg Biohub, San Francisco, CA 94158, USA*
[3]*Department of Bioengineering, Stanford University, Stanford, CA 94305, USA*
[4]*Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[5]*Department of Physics, California Institute of Technology, Pasadena, CA 91125, USA*
[+]*Correspondence: phillips@pboc.caltech.edu*

S1 Finding number of genes without regulatory annotation

S1.1 *E. coli* K12 MG1655

S1.2 *Bacillus Subtilis*

S1.3 *Pseudomonas Aeruginosa*

S1.4 *Saccharomyces cerevisiae*

S1.5 *Drosophila Melanogaster*

S1.6 *C. elegans*

S2 Reporter Sequence design

S3 Oligo Pool Design

S3.1 Identification of Transcription Start Sites

All oligo pools used in this work were manually designed. For each gene in our list we looked for promoters in Ecocyc [15] (accessed 12/08/2021) using the transcription start site if the promoter was found. If multiple promoters were identified, each promoter was included in the experiment. If no promoter was found, we looked for transcriptionally active sites in the data set from Urtecho et al, 2020[7]. In their work, every part of the genome was tested for transcription initiation in LB. If we could find a site that was identified as active close to the gene of interest, we chose this site as origin for computational promoter mutagenisis. If no transcription start site could be identified for a gene, the model from [16] was used to computationally predict a transcription start site in the intergenic region. The site predicted to be the most active within 500 bp upstream of the coding region was chosen as transcription start site since more than 99% of transcription start sites are within that region in *E. coli* K12 MG1655, see Fig. S1. Restriction enzymes leaving compatible sticky ends to the digested plasmid were used to cut the RiboJ::sfYFP element.

S3.2 Computational Promoter Mutagenesis

Once a TSS is identified, the 160 bp region from 115 bp upstream of the TSS to 45 bp downstream is taken from the genome. It has been shown that most cis-regulation is happening within this window [17]. Based on the approach by [9], each promoter sequence is mutated randomly at a rate of 0.1 per position. 1500 mutated sequences are created per promoter, following the approach from [8], which creates sufficient mutational coverage across the window. The promoter oligonucleotides are flanked

by restriction enzyme sites (*rs1* and *rs2* in Fig. S2) that are used in downstream cloning steps. The restriction sites are flanked by primer sites used to amplify the oligo pool. Primer sequences were chosen from a list of orthogonal primer pairs, designed to be optimal for cloning procedures [18]. oligo pools were synthesized (TwistBioscience, San Francisco, CA, USA) and used for subsequent cloning steps.

S4 Library Cloning

S4.1 Cloning oligo pool into plasmid vector

The oligo pool was amplified using a 20bp forward primer (SC142) and a 40 bp reverse primer (SC143), which consists of 20bp primer binding site and 20bp overhang. PCR amplifications were run to minimal amplification to minimize amplification bias. PCR products were cleaned and concentrated (DNA Clean & Concentrator-5, ZymoResearch) and used for a second amplification step. The 20 bp overhang from the first amplification was used as primer site for a reverse primer (SC172), which contains randomized 20 bp barcode, flanked by two restriction enzyme sites (*rs3* and *rs4* in Fig. S2). The forward primer is the same as in the first amplification step. PCR amplification is run again to minimal amplification to minimize amplification bias. PCR products are run on a 2% agarose TAE gel and subsequently extracted and purified (Zymoclean Gel DNA Recovery Kit, ZymoResearch). In the next step, restriction digest is performed on the outer restriction enzyme sites (*rs1* and *rs4* in Fig. S2). Unless noted otherwise, all restriction digests were run for 15 minutes at 37C. The plasmid vector was digested with different restriction enzymes which create compatible sticky ends. Most restriction enzyme sites are palindromes, so by choosing different enzymes with compatible ends, we avoid having palindromes flanking the plasmid inserts. This is important, since these sites are used for amplifications in the library preparation steps later in the protocol. (Maybe not needed to say). The oligo pool is combined with the plasmid vector using T7 DNA ligase (New England Biolabs, Ipswich, MA, USA) following the suppliers protocol. Ligation products were cleaned and concentrated (DNA Clean & Concentrator-5, ZymoResearch) and drop dialysis (MF-Millipore VSWP02500, MilliporeSigma, Burlington, MA, USA) was performed for 1h to improve sample purity. Electroporation using *E. coli* pir116 electrocompetent cells (Lucigen, Middleton, WI) was performed at 1.8kV in 1mm electroporation cuvettes, followed by 1h recovery at 37C and 250rpm in 1 ml LB-media (details here, the same for all following mentionings of LB). The entire cultures were plated on 150mm kanamycin (50μg/ml) + LB petri dishes and grown overnight. The following day, plates were scraped and the colonies resuspended. Freezer stocks were prepared using a 1:1 dilution of resuspended colonies and 50% glycerol. Cultures were inoculated with 5×10^8 cells in 200ml of LB + kanamycin (50μg/ml) and grown at 37C until saturation. Plasmid was extracted (ZymoPURE II Plasmid Maxiprep Kit, ZymoResearch) and used subsequent sequencing (see S5). The plasmid library is then used as template in a restriction digest using restriction enzymes $rs2$ and $rs3$. The resulting product was cleaned and concentrated (NEB Monarch) and concentration measured on a Nanodrop. Similarly, the riboJ::YFP element was PCR amplified (primers SC191 and SC192), adding restriction sites as overhangs (see table S1). The PCR product was cleaned and concentrated (NEB Monarch) and digested with the respective restriction enzymes. The plasmid library is combined with the RiboJ::sfYFP element using 7 DNA ligase (New England Biolabs, Ipswich, MA, USA) following the suppliers protocol. Ligation products were cleaned and concentrated (NEB Monarch) and drop dialysis (MF-Millipore VSWP02500, MilliporeSigma, Burlington, MA, USA) was performed for 1h to improve sample purity. Electroporation using *E. coli* pir116 electrocompetent cells (Lucigen, Middleton, WI) was performed at 1.8kV in 1mm

Part	5' restriction site	3' restriction site
Plasmid Vector	XbaI	XhoI
RiboJ::YFP	ApaI	PtsI
Oligo Pool	SpeI	ApaI
Barcoding Primer	SbfI	SalI

Table S1. Restriction sites used. All enzymes were ordered from NEB (check which ones are high fidelity versions)

electroporation cuvettes, followed by 1h recovery at 37C and 250rpm in 1 ml LB-media. The entire cultures were plated on 150mm kanamycin (50μg/ml) + LB petri dishes and grown overnight. The following day, plates were scraped and the colonies resuspended. Freezer stocks were prepared using a 1:1 dilution of resuspended colonies and 50% glycerol. Cultures were inoculated with 5×10^8 cells in 200ml of LB + kanamycin (50μg/ml) and grown at 37C until saturation. Plasmid was extracted (ZymoPURE II Plasmid Maxiprep Kit, ZymoResearch) and used for subsequent genome integration.

S5 Barcode Mapping

The plasmid library is used for barcode mapping. Purified plasmid is PCR amplified using forward primer (SC185) outside the promoter region and a reverse primer outside the 20bp barcode (SC184). The PCR is run to minimal amplification (until a band is visible on an ararose gel), and the product is gel purified (NEB Monarch). The purified DNA was used as template for a second PCR using a primer (SC196) adding an Illumina P5 adapter to the promoter side, and a primer (SC199) adding an Illumina P7 adapter. The PCR is again run to minimal amplification and gel purified (NEB Monarch). The product was used for sequencing on a Illumina NextSeq P2 flow cell with pair end reads using primers SC185 for read 1, SC184 for read 2 and SC201 for the index read. Reads were filtered and merged using custom bash scripts, which are available in the Github repository. After processing, each promoter/barcode pair was identified in each read, and pairs with less than 3 total reads were discarded. An alignment algorithm was used to identify the identity of each sequenced promoter variant. This allowed to include additional promoter variants that were in the initial oligo pool due to synthesis errors in the production of the oligos. The barcode mapping was used in analysis of libraries grown in various growth conditions. (TR: Give summary statistics and show plot maybe)

S5.1 Genome Integration

We used ORBIT to integrate the reporter libraries into the chromosome. A detailed description of the method and its efficiencies can be found in (Add scotts paper here). Wild type *E. coli* (K12 MG1655) are streaked on a LB plate and grown overnight at 37C. A single colony is picked and grown in 3ml of LB at 37C and shaken at 250rpm overnight. The overnight culture is diluted 1:1000 into fresh LB (e.g. 200ml) and grown at 37C and 250rpm until exponential phase ($\sim$ 0.4 OD 600nm). The cultures are then immediately put on ice and spun in a centrifuge at 5000g for 10min. Following the spin, the supernatant is discarded, and the cells are resuspended in deionized water at 4C at the same volume as the initial culture. The cells are spun again at 5000g for 10 min. This wash step is repeated 4 times with 10% glycerol. After the last wash, supernatant is discarded and cells

are resuspended in the remaining liquid and distributed into 50μl aliquots. Aliquots are frozen on dry ice and kept at -80C until used for electroporation. For electroporation, aliquots are thawed on ice and 1mm electroporation cuvettes are pre-chilled on ice. 100ng of helper plasmid (link to helper plasmid file) is added to a 50μl cell aliquot and mixed by slowly pipetting up and down. The aliquot is then added to the electroporation cuvette and electroporation is performed at 1.8kV. The aliquot is recovered with 1ml of LB media prewarmed to 37C for an 1h prior to electroporation. The culture is recovered for 1h at 37C and shaken at 250rpm. After recovery, aliquots at various dilutions are plated on LB + gentamycin (check gent concentration). Plates are grown overnight and a single colony is picked to prepare frozen stocks as described above. To perform genome integration, the host strain carrying the helper plasmid is made electrocompetent (follow growing and washing steps described above), and the plasmid library is electroporated into the host strain. The cells are recovered in 3ml of prewarmed LB + 1% arabinosea and shaken at 37C at 250rpm for 1h. The entire volume is plated on LB + kanamycin plates (TR: add concentration) and colonies are grown over night. The next day, colonies are scraped, resuspended in LB and diluted to optical density of 1 at 600nm. The helper plasmid used for genome integration causes growth deficits, hence, the library needs to get cured of the plasmid. Therefore, the library is inoculated with 0.5ml of culture at 1 OD in 200ml of LB, and grown until exponential phase at 37C shaken at 250 rpm. The helper plasmid carries the *sacB* gene, which is used for negative selection in the presence of sucrose. At exponential phase, the culture is plated on LB + 7.5% sucrose agarose plates. Plates are grown overnight, scraped and made into frozen stocks. The frozen stocks are then ready for growth experiments.

S6 Growth Conditions and Culture Growth

S6.1 gDNA and RNA extractions

S7 Barcode Sequencing

S8 Supplementary Files

- Plasmid Sequences with annotations + RiboJ::YFP
- pHelper sequence
- Primers
- list of restriction sites used in cloning
- Gene list
- Sequencing Data
- List of ordered sequences

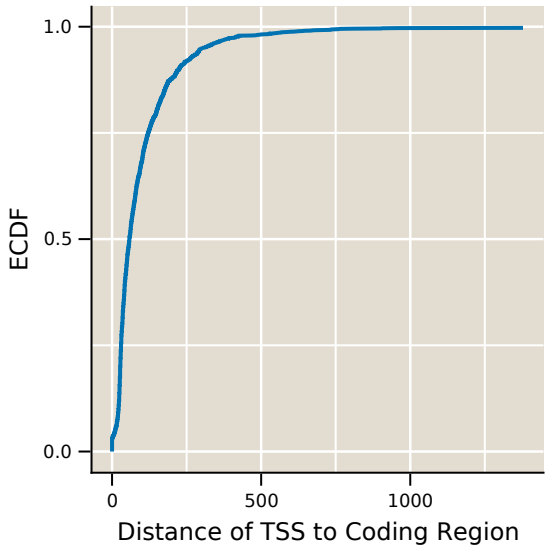

Figure S1. ECDF of distances of transcription start sites to the coding region for every operon in *E. coli* that has a transcription start site annotated in EcoCyc.

Supplementary Figures

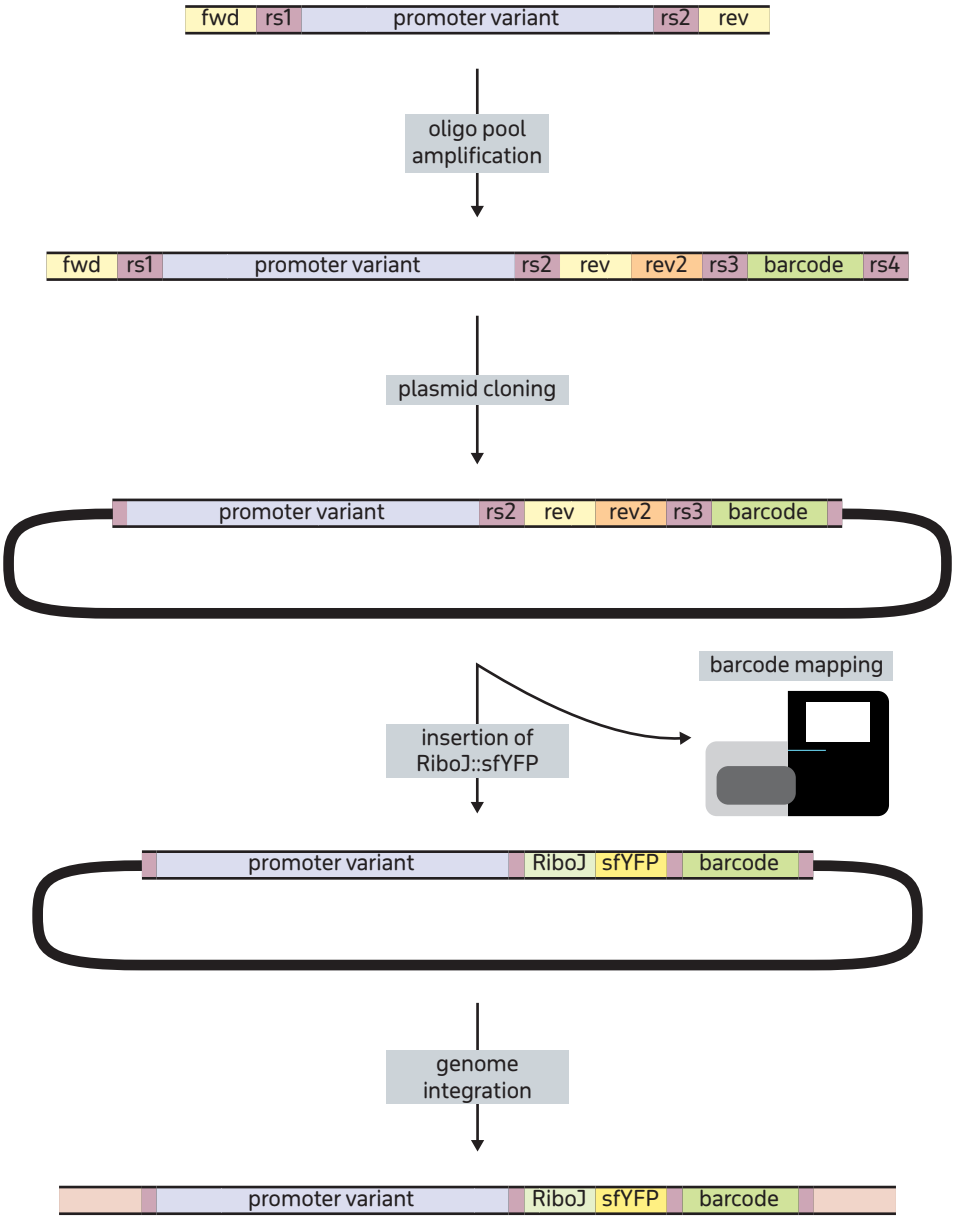

Figure S2. Placeholder figure for cloning scheme.

Figure S3. Mutation rate profile for the promoter of *dicC*. Mutation rate per position (blue) with rolling average over 11 positions (orange) compared to expected average mutation rate of 0.1 (purple). Predicted repressor binding site indicated by grey vertical lines.

Supplemental References

[15]I. M. Keseler et al., "Ecocyc: a comprehensive database of escherichia coli biology", Nucleic acids research **39**, D583–D590 (2010).

[16]T. L. La Fleur, A. Hossain, and H. M. Salis, "Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria", bioRxiv (2021).

[17]M. Rydenfelt, R. S. Cox III, H. Garcia, and R. Phillips, "Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration", Physical Review E **89**, 012702 (2014).

[18]S. K. Subramanian, W. P. Russ, and R. Ranganathan, "A set of experimentally validated, mutually orthogonal primers for combinatorially specifying genetic components", Synthetic Biology **3**, ysx008 (2018).