Another 100 genes

Tom Röschinger[1], Grace Solini[1], Anika Nawar Choudhury[2], Bob Jones[3], 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 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 about 10^{17} bases have been deposited in the SRA database (TR: cite SRA website (https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/)), 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 (see S1.1). For other prokaryotic model organisms the numbers are similar (see S1.2, S1.3), while higher order model organisms such as *Saccharomyces cerevisiae* (see S1.4) and *C. elegans* (see S1.6) have close to no regulatory annotations, given the arguably more complex nature of gene regulation in eukaryotes. 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 [2–4].

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 or for a broad spectrum of DNA binding factors. In methods like ChIP-Seq [5], proteins have to be cross linked to DNA, which often requires changing residues in the amino-acid sequence, such as for LacI in *E. coli* [6]. In addition, antibodies against the protein of interest have to be available, or the protein has to be modified to include a tag which can be targeted by antibodies. While the resolution of these methods is ever improving, it does not allow for a nucleotide resolution yet, making it difficult to identify changes in binding affinity caused by single mutations. Other methods such as ATAC-Seq [7, 8] and DNase-Seq [9] rely on open chromatin for binding site identification, and have almost exclusively been used for eukaryotic organisms. In general, identifying regulatory interactions from transcription factor occupancy alone can be misleading, since there can be high affinity binding sites in the genome, where there is no change in expression levels upon binding [10]. DAP-Seq [11] is a method similar to ChIP-Seq, however, instead of using immunoprecipitation to obtain DNA-TF pairs, purified and tagged TFs are incubated with fragmented genomic DNA. The method has been used to identify genome wide binding sites for TFs in *Clostridium thermocellum* [12] and *Riemerella anatipestifer* [13].

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 [14–16]. 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 [16–19]. Additionally, we present major improvements to the method called Reg-Seq [16], 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 on 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: To do:

- Transcription Factor identification using DAP-Seq, and possibly how we ended up scaling it
- final paragraph about the outlook to make it Reg-Seq genome wide
- Write about Rosalinds stuff, either in detail if part of the paper or mention and cite it if it is something different

) RP: TEST

2 Methods

2.1 Promoter sequence import

2.2 Reporter construct design

2.3 Barcode Mapping

2.4 Genome Integration

3 Results

3.1 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 [16, 18]. 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 [20]. 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 [20]. 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 [21] 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)

3.2 Barcode Mapping

For each gene studied here, we designed a library of 1500 mutated promoter variants with an average mutation rate of 0.1 for each promoter that was identified with the gene in EcoCyc. If a gene did not have a promoter identified, we first looked for a possible TSS identified by Urtecho et al., 2018 [14], which then was taken as initial sequence for generating mutated variants. In some cases, no TSS could be identified and we used the model of LaFleur et al., 2022 [22], which predicts trancsription start site given a genomic sequence of *E. coli*, to find the site in the intergenic region leading up to the first coding region in the operon the gene was part of that had the highes predicit affinity for σ70-factors. A total of 178619 sequences were ordered from Twist Biosciences. We added random barcodes to the sequences and cloned them into a plasmid vector for amplification and subsequent cloning steps. Details can be found in Supplementary information sections S3.1, S3.2, S4.1 and S5. 170192 (¿95%) of sequences were identified in the plasmid library (Figure S6), and an addional 17114 sequences were found, likely due to errors in the synthesis process of the oligonucleotide library.

3.3 Genome Integration of Reporters

3.4 Transcription Factor identification

3.5 Growth Conditions

3.6 Gold Standard genes

3.7 Ethanol iModulon

YgeV has been predicted to be a regulator involved in purine catabolism, leading to the production of allantoin, which can be used as a sole nitrogen source [23]. There are 16 putative regulatory targets [21] for YgeV, including the *xdhABC* operon, which degrades xanthine to uric acid [23] in the purine catabolism pathway. *E. coli* can survive exposure to low ethanol concentrations up to 5%, which can even lead to increased DNA synthesis [24], but mostly leads to varies stress responses such as an increased production of ROS. Growth in media supplemented with ethanol induced a change in gene expression for genes regulated by YgeV in a Δ*baeR* or Δ*cpxR* mutant strain. (TR: Is there a correlation between ethanol response and higher need for nitrogen?)

3.8 Oxidative stress response iModulon

The putative transcription factor YmfT regulates 14 out of 23 genes in the e14 prophage and is predicted to respond to oxidative stress [21]. Oxidative stress is caused by reactive oxygen species (ROS) such as H_2O_2, which are highly reactive and damage DNA, the cell wall, proteins [25] etc., however, oxidized amino acids can also lead to confirmational changes in transcription factors, such as OxyR and HypT, which induce DNA binding and subsequent regulation of genes involved in response to oxidative stress [25]. Hydrogen peroxide is produced endogenously in various pathways in *E. coli* and especially in high amounts when phenylethylamine is used as either carbon or nitrogen source [26], hence we used used minimal media supplemented with 10 mM 2-phenylethylamine hydrochloride (PEA) as sole carbon source to induce stress responses to H_2O_2 and therefore oxidative stress. (TR: discuss findings of ymfT modulon and how it relates to oxyR, look at oxyR iModulons and possibly do another run including this gene.)

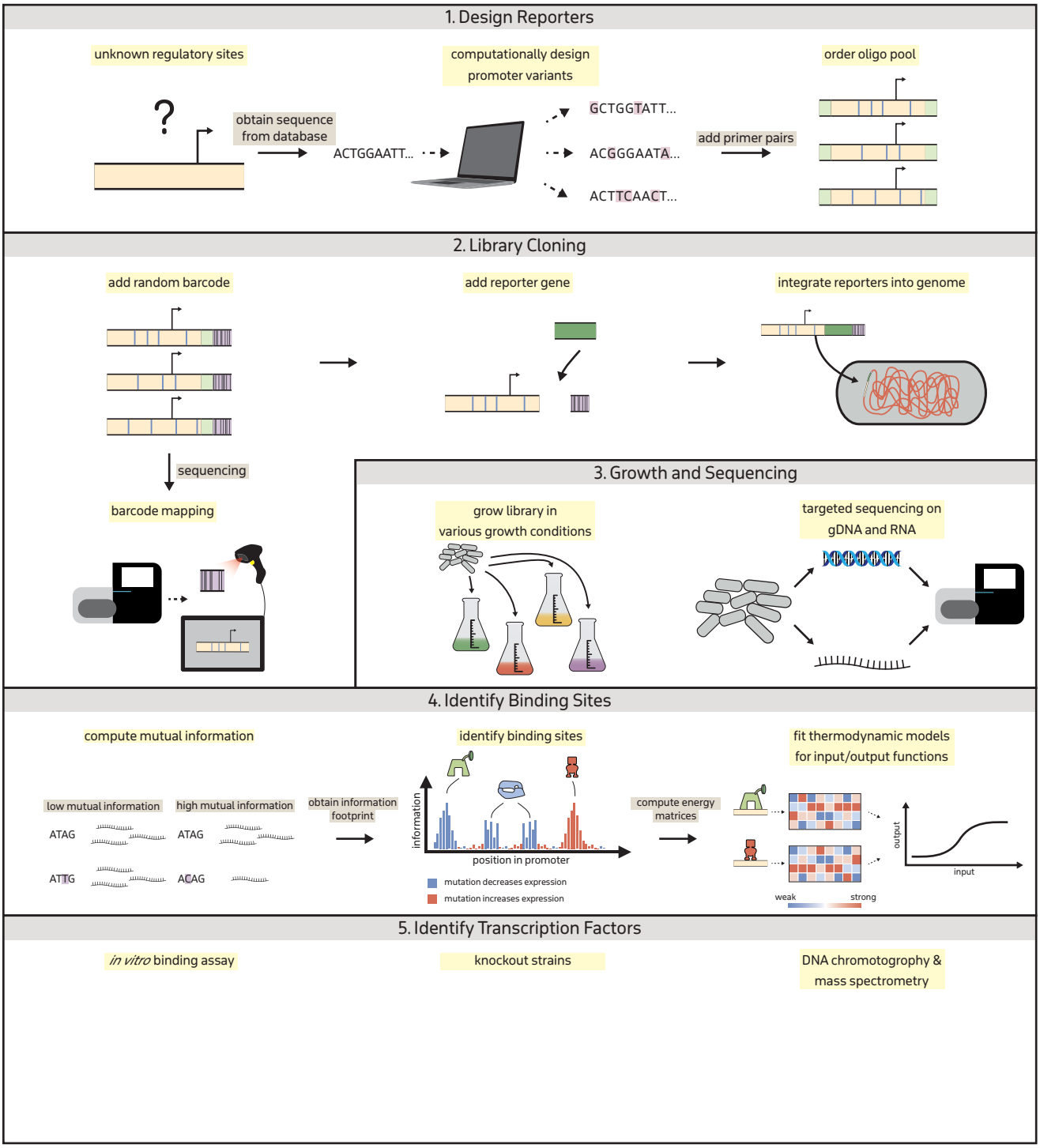

Figure 1. Method summary

3.9 Antitoxin/Antibiotic genes

3.10 other y-ome genes

4 Discussion

- discuss how to scale to 1000 genes

5 Acknowledgements

Bill Ireland, Justin Kinney, Stephan Grill, Frank Jülicher, Igor A. Antoshechkin This work was supported by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at California Institute of Technology

6 To do list

- Complete Introduction
- Write SI about mutual information and determining binding sites
- expand paragraphs for genes chosen
- check experimental details in methods

References

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

[2]M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators", Nature **403**, 335–338 (2000).

[3]S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif", Proceedings of the National Academy of Sciences **100**, 11980–11985 (2003).

[4]U. Alon, *An introduction to systems biology: design principles of biological circuits* (Chapman and Hall/CRC, 2006).

[5]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).

[6]D. Rutkauskas, H. Zhan, K. S. Matthews, F. S. Pavone, and F. Vanzi, "Tetramer opening in laci-mediated dna looping", Proceedings of the National Academy of Sciences **106**, 16627–16632 (2009).

[7]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).

[8]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).

[9]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).

[10]A. H. Yona, E. J. Alm, and J. Gore, "Random sequences rapidly evolve into de novo promoters", Nature communications **9**, 1530 (2018).

[11]A. Bartlett, R. C. O'Malley, S.-s. C. Huang, M. Galli, J. R. Nery, A. Gallavotti, and J. R. Ecker, "Mapping genome-wide transcription-factor binding sites using dap-seq", Nature protocols **12**, 1659–1672 (2017).

[12]S. D. Hebdon, A. T. Gerritsen, Y.-P. Chen, J. G. Marcano, and K. J. Chou, "Genome-wide transcription factor dna binding sites and gene regulatory networks in clostridium thermocellum", Frontiers in Microbiology **12**, 695517 (2021).

[13]Y. Zhang, Y. Wang, Y. Zhang, X. Jia, C. Li, Z. Zhou, S. Hu, and Z. Li, "Genome-wide analysis reveals that phop regulates pathogenicity in riemerella anatipestifer", Microbiology Spectrum **10**, e01883–22 (2022).

[14]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).

[15]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).

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

[17]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).

[18]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).

[19]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).

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

[21]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).

[22]T. L. LaFleur, A. Hossain, and H. M. Salis, "Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria", Nature communications **13**, 5159 (2022).

[23]Y. Iwadate and J.-i. Kato, "Identification of a formate-dependent uric acid degradation pathway in escherichia coli", Journal of bacteriology **201**, e00573–18 (2019).

[24]T Basu and R. Poddar, "Effect of ethanol onescherichia coli cells. enhancement of dna synthesis due to ethanol treatment", Folia microbiologica **39**, 3–6 (1994).

[25]B. Ezraty, A. Gennaris, F. Barras, and J.-F. Collet, "Oxidative stress, protein damage and repair in bacteria", Nature Reviews Microbiology **15**, 385–396 (2017).

[26]S. Ravindra Kumar and J. A. Imlay, "How escherichia coli tolerates profuse hydrogen peroxide formation by a catabolic pathway", Journal of bacteriology **195**, 4569–4579 (2013).

Supplemental Information for: Whatever the title will be

Tom Röschinger[1], Grace Solini[1], Anika Nawar Choudhury[2], Bob Jones[3], 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 [27] (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[15]. 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 [22] 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. S3. 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 [28]. Based on the approach by [17], 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 [16], which creates sufficient mutational coverage across the window. The promoter oligonucleotides are flanked

by restriction enzyme sites (*rs1* and *rs2* in Fig. S4) 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 [29]. 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. S4). 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. S4). 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-HF	ApaI
Barcoding Primer	SbfI-HF	SalI-HF

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. The code used to perform processing of sequencing data can be found in the associated Github repository https://github.com/RPGroup-PBoC/1000_genes_ecoli/tree/main/code/processing/20220514_mapping. Processing is done with the help of various software modules [30–32]. Custom Julia code used for analyis and visualization of results can be found in the associated Github repository https://github.com/RPGroup-PBoC/1000_genes_ecoli/tree/main/code/processing/20220514_mapping

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 at an OD600 of 1. The frozen stocks are then ready for growth experiments.

S6 Growth Conditions and Culture Growth

For each growth condition, cultures were inoculated from frozen stocks in 200ml of LB media (details) and grown overnight at 37C shaken at 250rpm. In the morning, cultures were diluted 1:100 into the growth media of choice, which, unless noted otherwise is at a volume of 200ml. Cultures are grown at 37C until reaching exponential phase (OD600 of 0.4). Once the culture reaches exponential phase, 1ml aliquots are spun down at 5000g for 10min for subsequent gDNA extraction. Supernatant was discarded and pellets were frozen at -20C overnight. For RNA extraction, culture was diluted in RNAprotect (Quiagen, add info) and 1ml aliquots are spun down at 5000g for 10min Supernatant was discarded and pellets were frozen at -20C overnight. (TR: Add details for special growth conditions)

S6.1 gDNA and RNA extractions

S7 Barcode Sequencing

S8 Promoter footprints

If a base of a binding site for a regulatory element in a promoter is mutated, the expression of the downstream gene is changed due to differences in binding affinity of the regulatory element (TR: cite Kinney, 2010 and Garcia, 2011; but maybe find some older/more original references). One can generate so called *footprints*, where the effect of a mutation in the promoter on expression levels

c_{dna}	c_{rna}	sequence
10	2	ACGTACGTAC
1	2	ACGTACGTTC
3	5	ACGTACGTTC
4	9	ACGTACGTTC
3	5	ACGTAAGAAC
3	6	ACGTAAGAAC
15	12	GCGTACGTAC
5	3	GCGTACGTAC
12	14	ACATACGTAC
2	3	ACATACGTAC
20	40	ACATACGTAC
5	3	ACGGATGTAC
5	1	ACGTACGTGA
10	1	ACGTACGTGA
2	10	ACGTCCATAC
2	10	ACGTCCATAC
4	13	ACGTCCGTAC
18	25	ACAAACGTAC
17	19	GCGTACGTAG
10	11	GCGTACGTAG
2	3	GGGTACGTAG

Table S2. Example dataset, arbitrarily generated. For each sequence, there are counts from RNA and DNA sequencing. Different counts for the same sequence come from unique barcodes, are therefore separate measurements. (TR: has to be updated to have the correct sequences for the figures below)

can be quantified by various metrics. Here, we explore various ways to compute footprints and explain each method in detail. (TR: add the footprints from one real dataset to compare)

S8.1 Dataset

For a given promoter, there are $i = 1, .., n$ promoter variants, where each variant has m_i unique barcodes. Per barcode, there are c_{dna} counts from genomic DNA sequencing, as well as c_{rna} counts from RNAseq. DNA sequencing is performed to normalize the RNA sequencing data by the abundance of cells in the culture expressing the reporter from a specific promoter variant.

S8.2 Frequency Matrices

(TR: Not sure if I will actually write about it, just a different way of computing footprints I came up with based on comments by Frank Jülicher and Stephan Grill. Have try it on old data set.)

S8.3 Expression Shifts

Belliveau et al. (2018)[18] used so called *expression shifts* to compute footprints for mutagenized promoters. In their experiments, cells were sorted based on fluorescence, where the fluorescent reporter gene was expressed under the control of a mutagenized promoter variant, and subsequently sequenced. Therefore, each sequence had a bin associated with it, which is a read out for how strong the reporter is expressed relatively to the other promoter variants in the library. This approach can be adapted to our data set, where we first compute the average relative expression $\langle c \rangle_i$ for the i-th promoter variant across all of its unique barcodes,

$$\langle c \rangle_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{c_{\mathrm{rna},j}}{c_{\mathrm{dna},j}}. \tag{S1}$$

Then, we determine how much relative expression is changed at each position if there is a mutation. If a base at position ℓ in promoter variant i is mutated, we denote that as $\sigma_{i,\ell} = 1$. Otherwise, if the base is wild type, we write $\sigma_{i,\ell} = 0$. Then, the change in relative expression due to mutation, the expression shift Δc_ℓ, at position ℓ is given by

$$\Delta c_\ell = \frac{1}{n} \sum_{i=1}^{n} \sigma_{i,\ell} \left(\langle c \rangle_i - \frac{1}{n} \sum_{k=1}^{n} \langle c \rangle_k \right). \tag{S2}$$

The absolute value of expression shift can be hard to interpret, so indeed one can present it in terms of relative change to the mean expression, i.e., fold-change,

$$\delta c_\ell = \frac{\Delta c_\ell}{\langle c \rangle} = \frac{1}{n} \sum_{i=1}^{n} \sigma_{i,\ell} \left(\frac{\langle c \rangle_i}{\langle c \rangle} - 1 \right). \tag{S3}$$

Figure **??** shows the expression shift footprint that is obtained for the test dataset. (TR: as well as the footprint for a real data set).

S8.4 Mutual Information

Mutual information is a measure of how much information is obtained about a random variable by measuring a different random variable. In the context of gene expression, this can be understand as the ability to predict changes in gene expression given a certain mutation on the promoter sequence. If there is no annotation, meaning it us unknown where RNAP or transcription factors bind, one can not make any predictions on the expression level of the downstream gene when observing a mutation in the promoter. In this case, there is low mutual information between sequence and expression level. On the other hand, if the promoter is annotated and on has binding energy matrices for all transcription factor binding sites and the RNAP binding site in hand, then one can precisely predict the change in gene expression given any point mutation based on thermodynamic models (TR: could cite a bunch of papers here), which is a case of high mutual information. Hence, by maximizing the mutual information between a model for the regulatory architecture and observed levels of gene expression, we can discover binding sites for transcription factors and subsequently, using equilibrium thermodynamic models and neural networks, compute binding energy matrices in real units of $k_B T$.

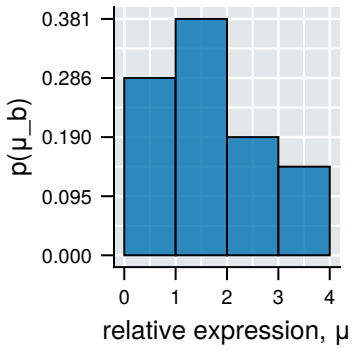

Figure S1. Possible binning of expression counts for example data set. (TR: Add footprint for real dataset.)

S8.4.1 Mutual Information based on Sequence Counts

The first way of computing mutual information at each position in the promoter is to take the base at each position as one random variable, and the expression of each sequence as other random variable. As measure for expression, we use RNA counts for each sequence normalized by DNA counts. In order to compute mutual information, we need to obtain a probability distribution $p_\ell(c, \mu)$, which gives the probability of finding a certain base c at position ℓ, and corresponding expression μ. One way of obtaining such a distribution is to find bins for the values of μ, denoted as μ_b, as shown in Figure S1. Then, mutual information is given by

$$I_\ell = \sum_{c=\mathrm{A,C,G,T}} \sum_{\mu_b} p_\ell(c, \mu_b) \log_2 \left(\frac{p_\ell(c, \mu_b)}{p_\ell(c) p(\mu_b)} \right), \tag{S4}$$

where $p_\ell(c)$ and $p(\mu_b)$ are the marginal distributions.

S8.4.2 Mutual Information based on Phenotype Matrices

A different way to utilize mutual information is to choose a phenotype as random variable instead of base identity. In this case, the phenotype Φ is a real number and is additive across the sequence, meaning that each position l with base c contributes $\Theta_{l:c}$ to the total phenotype. The contributions are independent, i.e., no epistasis effects are considered for this model. The phenotype Φ is then determined by the sum across all positions with a possible offset Θ_0,

$$\Phi = \Theta_0 + \sum_{l=1}^{L} \sum_{c} \Theta_{l:c} x_{l:c}, \tag{S5}$$

where $x_{l:c}$ is a one-hot representation of the sequence with

$$x_{l:c} = \begin{cases} 1 & \text{if character c occurs at position l,} \\ 0 & \text{otherwise} \end{cases} \tag{S6}$$

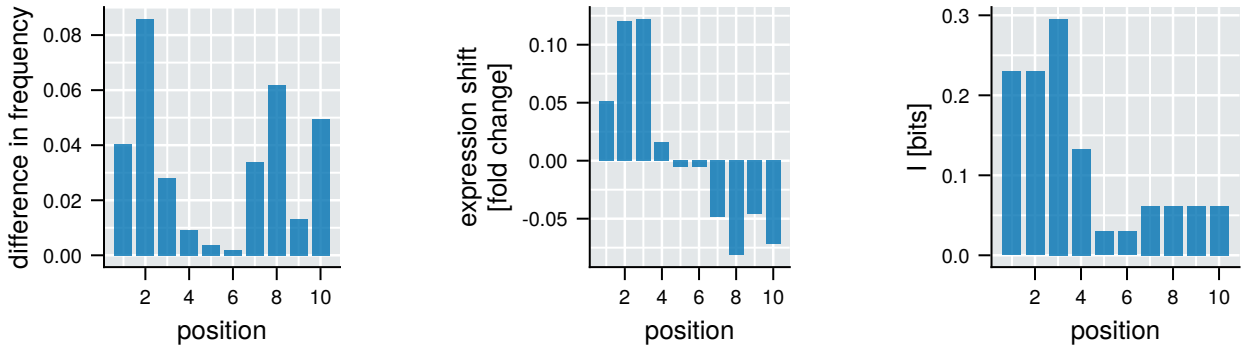

Figure S2. Different ways of computing footprints for test data set from table S2. Frequency Matrix left, Expression Shift middle, Mutual information right

where the notation is adapted from [33]. Without any knowledge of the regulatory architecture of the promoter, one can only make random guesses for the phenotype matrix. However, either using Metropolis-Hasting algorithms (TR: Reg-Seq, gotta decide how much to write about it) or Neural Networks (TR: MaveNN, will be included if we get good results with it), the phenotype matrix can be optimized in the sense its entries are more extreme where there are binding sites for regulatory elements in the sequence, since a mutation in that part of the sequence will have the strongest effect on gene expression. How extreme entries are can be quantified by using relative entropy, where the entries for each position on the sequence are first converted to a probability distribution using exponential weights, and then Kullback-Leiback-Divergence (KLD) between the resulting distribution and a uniform distribution is calculated. (TR: Expand by explaining how peaks are identified as binding sites.)

S8.4.3 Phenotype Matrices and Neural Networks, MaveNN

S8.4.4 Identifying Binding Energy Matrices

S9 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

Supplementary Figures

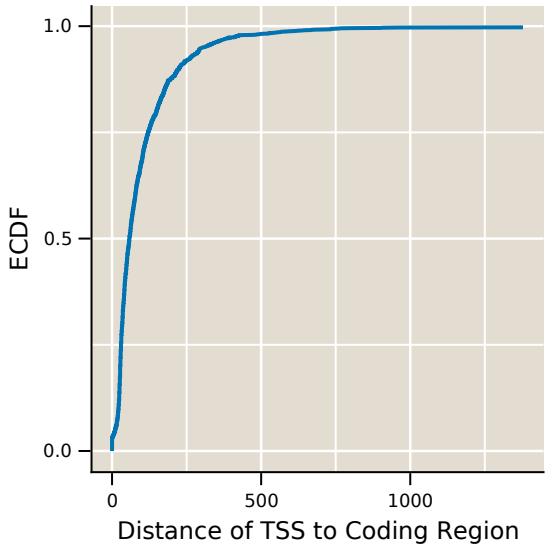

Figure S3. ECDF of distances of transcription start sites to the coding region for every operon in *E. coli* that has a transcription start site annotated in EcoCyc.

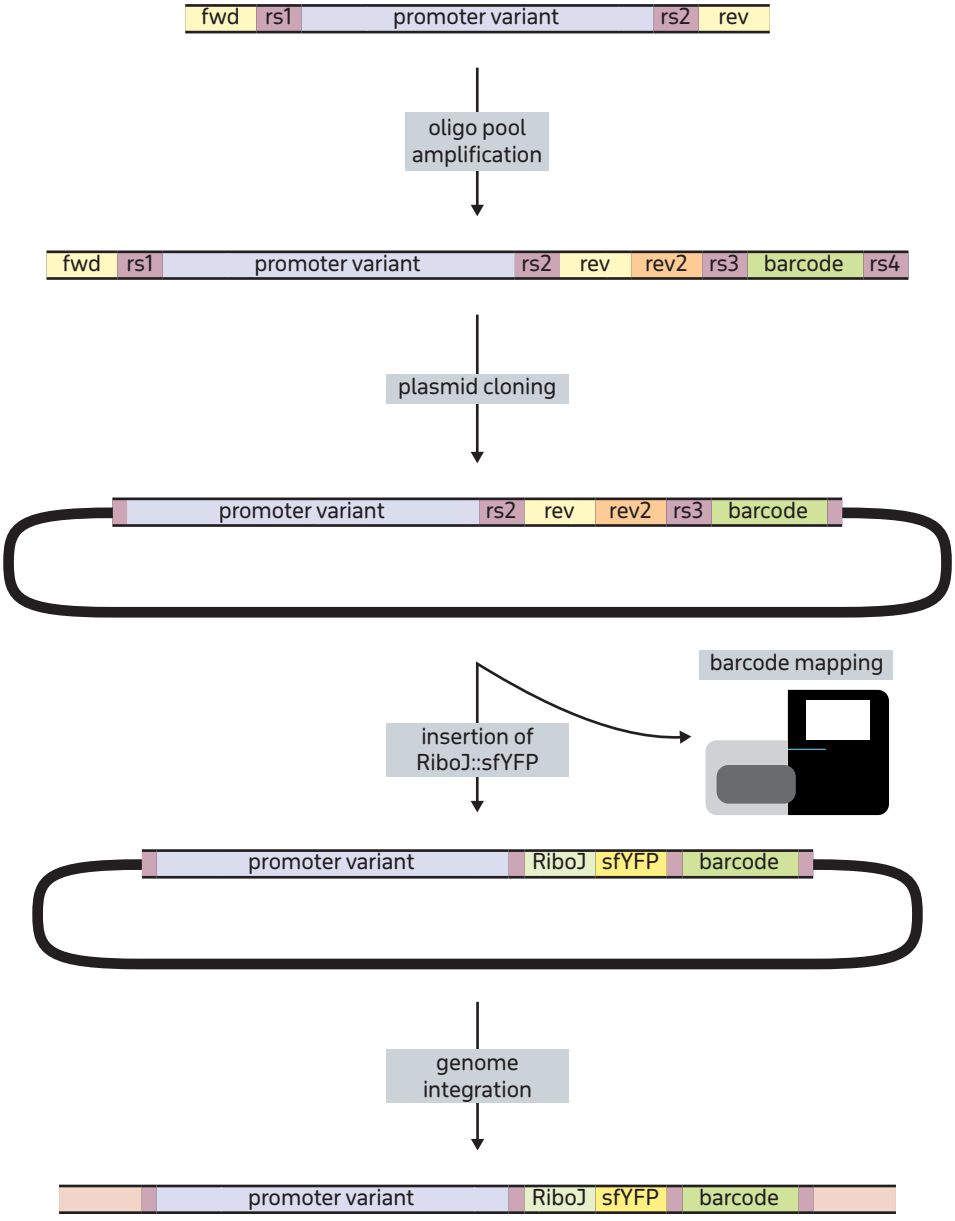

Figure S4. Placeholder figure for cloning scheme.

Figure S5. 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.

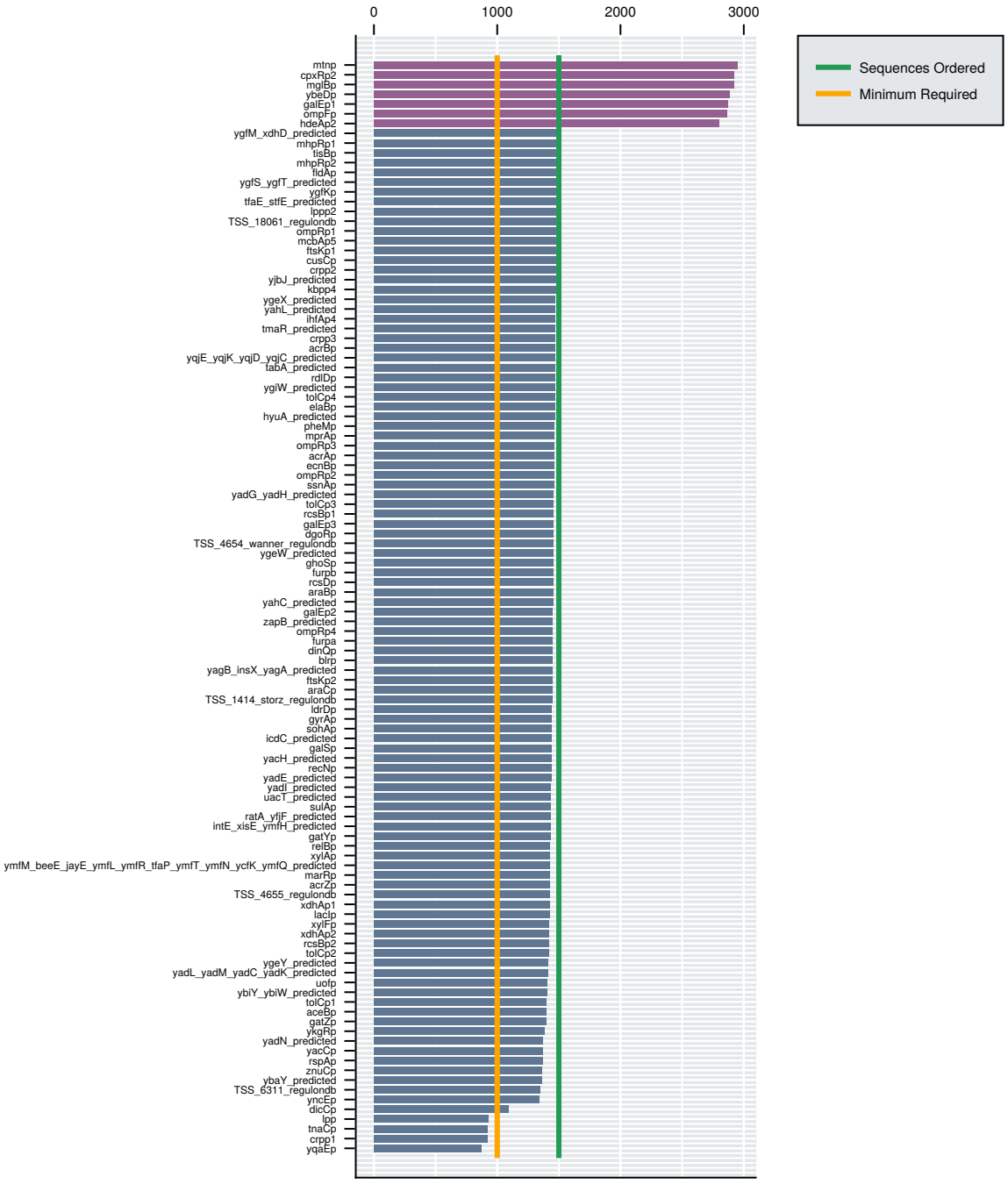

Figure S6. Number of unique mutated variants for each promoter after barcode mapping. Bars colored in purple indicate that a promoter was duplicated, i.e., there was a second promoter annotated to the same TSS in EcoCyc, for which 1500 variants were created as well, effectively doubling the number of mutated variants for the same promoter. Shown are promoter variants that could be mapped to at least 5 different barcodes.

Supplemental References

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

[28]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).

[29]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).

[30]B. Bushnell, *Bbmap: a fast, accurate, splice-aware aligner*, tech. rep. (Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014).

[31]S. Chen, Y. Zhou, Y. Chen, and J. Gu, "Fastp: an ultra-fast all-in-one fastq preprocessor", Bioinformatics **34**, i884–i890 (2018).

[32]O. Tange, *Gnu parallel 20230322 ('arrest warrant')*, GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them., Mar. 2023.

[33]A. Tareen, M. Kooshkbaghi, A. Posfai, W. T. Ireland, D. M. McCandlish, and J. B. Kinney, "Mavenn: learning genotype-phenotype maps from multiplex assays of variant effect", Genome biology **23**, 1–27 (2022).