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 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 [2], 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 **??**). For other prokaryotic model organisms the numbers are similar (see **??**, **??**), while higher order model organisms such as *Saccharomyces cerevisiae* (see **??**) and *C. elegans* (see **??**) 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 [3] (TR: Any additional citations other than 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 [4], 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 [5, 6] and DNase-Seq [7] 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 [8–10]. (TR: Add sentences about DAPseq and the method we recently reviewed (as much as we can say))

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 [10–13]. Additionally, we present major improvements to the method called Reg-Seq [10], 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: paragraph about scaling to 1000)

References

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

[2]National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2022 Oct 23]. Available from: https://www.ncbi.nlm.nih.gov/.

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

[4]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).

[5]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).

[6]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).

[7]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).

[8]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).

[9]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).

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

[11]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).

[12]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).

[13]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).

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 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 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)

S1.1 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.

S1.2 Expression Shifts

Belliveau et al. (2018)[12] 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 $\langle c \rangle_i$ per promoter variant across all of its unique barcodes,

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

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

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

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 S1. Example dataset, arbitrarily generated. Different counts for the same sequence come from unique barcodes, are therefore separate measurements.

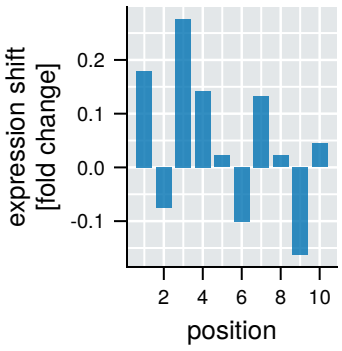

Figure S1. Expression shift for example dataset. (TR: Add expression shift for real dataset.)

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 S1 shows the expression shift footprint that is obtained for the test dataset. (TR: as well as the footprint for a real data set).

S1.3 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.)

S1.4 Mutual Information

Mutual information is a measure of 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 $k_B T$.

S1.4.1 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 S2. Then, mutual information is given by

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

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

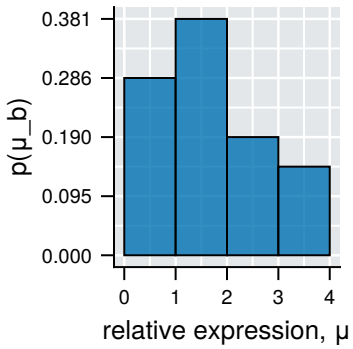

Figure S2. Possible binning of expression counts for example data set. (TR: Add footprint for real dataset.)

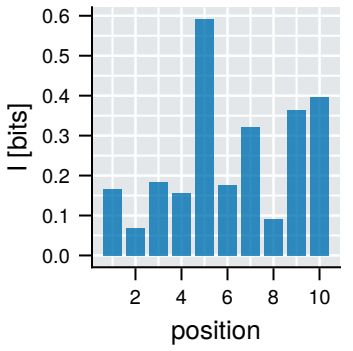

Figure S3. Mutual information based on base identity for example dataset. (TR: Add footprint for real dataset.)

S1.4.2 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. The phenotype can be computed from the sequence using e.g. an additive model, where each position independently contributes to the total value of the phenotype Φ,

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

where $\Theta_{l:c}$ is the phenotype matrix, Θ_0 an additive constant and $x_{l:c}$ is a one-hot representation of the sequence with

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

where the notation is adapted from [14]. 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.)

S1.4.3 Phenotype Matrices and Neural Networks, MaveNN

S1.4.4 Identifying Binding Energy Matrices

Supplemental References

[14] 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).