Another 100 genes

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

[1]*Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA*
[3]*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 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 allow 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 allow 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).

3 Methods

4 Results

4.1 Improved Method and summary of cloning results

4.2 Transcription Factor identification

4.3 Growth Conditions

4.4 Gold Standard genes

4.5 Ethanol iModulon

4.6 DNA damage repair iModulon

4.7 Antitoxin/Antibiotic genes

4.8 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).

Supplemental Information for: Whatever the title will be

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

[1] *Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA*
[3] *Department of Physics, California Institute of Technology, Pasadena, CA 91125, USA*
[+] *Correspondence: phillips@pboc.caltech.edu*

S1 Reporter Sequence design