A Multiplexed Approach to Defining Sequence-Function Relationships in Gene Expression using a Model Human Transcription Factor Binding Site

Davis, Jessica Elizabeth

2019

UNIVERSITY OF CALIFORNIA

Los Angeles

A Multiplexed Approach to Defining Sequence-Function Relationships in Gene Expression using a Model Human Transcription Factor Binding Site

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Biochemistry, Molecular and Structural Biology

by

Jessica Elizabeth Davis

2019

ABSTRACT OF THE DISSERTATION

A Multiplexed Approach to Defining Sequence-Function Relationships in Gene Expression using a Model Human Transcription Factor Binding Site

by

Jessica Elizabeth Davis

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2019

Professor Sriram Kosuri, Chair

In this dissertation I present a complete characterization of the transcriptional activity of a model human transcription factor binding site (TFBS), the c-AMP Response Element (CRE), across varied cis-regulatory element architectures. The arrangement and assortment of TFBSs within cis-regulatory elements drive specific gene regulatory responses, yet it remains difficult to predict gene expression based on sequence alone. Part of this issue lies in our incomplete picture of how a single TFBS drives expression across various regulatory architectures differing in TFBS composition, TFBS affinity, TFBS number, distance between TFBSs, distance of TFBSs to transcription start sites, and sequence content surrounding TFBSs. To better our understanding on sequence-function relationships in eukaryotic gene expression, we designed and assayed 9,126 synthetic regulatory elements isolating such TFBS variables. We developed and employed massively-parallel reporter assays (MPRAs) to enable episomal and genomic interrogation of

synthetic regulatory element activities in a human cell line. Overall, we find CRE number and affinity within regulatory elements largely determines expression, and this relationship is shaped by CRE proximities to promoter elements. Expression is not only dependent upon CRE’s overall distance to a downstream promoter, but also on its precise positioning and follows a ~10 bp periodicity along regulatory elements. Additionally, in the episomal MPRA, we find the spacing between multiple CREs dictates the phasing of expression periodicity in addition to overall expression. Lastly, we indicate differences between a single-copy genomic and episomal assay, highlighting the varied role certain TFBS variables have across regulatory contexts.

The dissertation of Jessica Elizabeth Davis is approved.

Albert J. Courey

Stephen T. Smale

Jason Ernst

Sriram Kosuri, Committee Chair

University of California, Los Angeles

2019

DEDICATION

I dedicate this work first to my family.

To my mother, Lara, father, Jeffrey, and grandparents, Duane, Kathy, Robert, and Marie, for their support in letting me pursue my ambitions.

To my mother for her hard work and time over the years ensuring my happiness, education, and ability to achieve my goals.

To my grandparents, Robert and Marie, for showing me the world and encouraging my questioning of it.

To my father and grandparents, Duane and Kathy, for my fascination with biology and nature.

To my boyfriend, John, for his never-ending support, love, and sympathy during my struggles with research, writing, and life.

To my cat, Olive, and all others before her, for their emotional support.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

LIST OF TABLES

Chapter 2

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Sriram Kosuri, for his guidance and support throughout my PhD. With his help, I grew into an independent and confident scientist. This thesis and my journey would not have been accomplished without him. Thank you.

Furthermore, I would like to thank my committee members Professors Albert Courey, Guillaume Chanfreau, Stephen Smale, and Jason Ernst. Their guidance, comments, and suggestions were helpful in framing my analyses and writings.

I would like to thank members of the Kosuri Lab for my development as a scientist. First I would like to thank Kimberly Insigne for her contributions to library construction and expression analysis on my paper mentioned in Chapter 2. Both Kimberly and Dr. Nathan Lubock were instrumental to my growth as an independent computational scientist and were most patient in addressing all my questions. I would also like to thank Clifford Boldridge for his insight and sarcasm over the years, aiding my research and mentality. I would like to thank Angus Sidore for his work on a related project to my research and for his guidance while finishing my PhD and post-PhD career planning. I would also like to thank Guillaume Urtecho for inspiring me to be a better and more optimistic scientist and mentor. I would like to thank the postdoctoral researchers in the Kosuri Lab, Dr. Rocky Cheung, Dr. Calin Plesa, and Dr. Hwangbeom Kim, for their generous advice and time over the years. Also, Dr. Eric M. Jones was helpful in framing my research and analysis over the years in addition to helping with cell line work mentioned in Chapter 2. I would also like to thank Dr. Plesa, Guillaume, Cliff, and Dr. Cheung for their much-needed and helpful writing advice. I would also like to thank Quinn Hastings, Tim Yu, Winnie Liu, Grace Bowers, and Jeremy Shek for teaching me how to be a better mentor. You all excelled with minimal input and I am glad to have been part of your growth as scientists. Specifically, I would like to thank Quinn Hastings for her help with library cloning and assay conditions mentioned in Chapter 2. Lastly, I would like to thank all other and previous members of the Kosuri Lab for their support and discussions: Christina Burghard, Daniel Cancilla, David Yao, Jeff Wang, Arielle Tripp, Marcia Brink, Joyce Samson, Johnny Lee, Taylor Ward, Suraj Alva, and Michael Corrado.

I would like to thank Dr. Kathrin Plath and all members of her lab who helped with my initial work in their lab. Dr. Kathrin Plath, Robin McKee, Dr. Amy Pandya-Jones, and Jarrett Miller were fundamental to my training in genetic manipulations of mouse embryonic stem cells and encouraging my skepticism of scientific methods and results during this process. I would also like to thank all other members of the Plath Lab for their support: Dr. Konstantinos Chronis, Dr. Justin Langerman, Dr. Martina Roos, Dr. Bernadett Papp, Dr. Vincent Pasque, Dr. Jacques Serizay, and Dr. Anna Sahakyan.

This work and my publication in chapter 2 would not have been possible without funding from the Ruth L. Kirchstein National Research Award (GM007185).

I would like to thank the UCLA Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research Sequencing Cores, particularly Suhua Feng, and the Technology Center for Genomics and Bioinformatics for providing next-generation sequencing services and assistance. I would also like to thank UCLA Eli & Edythe Broad Center of Regenerative Medicine Flow Cytometry Core for assistance in flow cytometry. I would like to thank Laura Day from the Kruglyak lab at UCLA for use of and assistance on their MiSeq. Lastly, I would like to thank Dr. Michael R. Sawaya for his guidance in CREB-CRE structural modeling mentioned in Chapter 2.

I would like to also thank those from my studies prior to graduate school. For Dr. Javier Read de Alaniz for first introducing me to research and welcoming me into his group. Also for Dr. Norbert Reich for teaching me the responsibilities of an independent scientist. Lastly, for Dr. Kalju Kahn for pushing and training me throughout my time at UCSB.

Lastly, I would like to thank those outside of lab for their support. For my family for supporting me and ensuring I could focus on research in a stress-free environment. For my mother Lara in particular for her advice and emotional support. For my boyfriend John, for all of his support and sympathy, without which would have made my last two years at UCLA unbearable. For my board game friends, Kimberly, Guillaume, Cliff, and Danny, as well as my Dungeons & Dragons group, Guillaume, Aaron, David, Daniel, Ryan, Parker, and Arturo, you guys all helped manage the stress from research and life. Lastly for all that I have learned from friends made while living in LA, I thank Cinta, Dr. Jain, Gabriel, Jing, and Jason.

Chapter 2 is a version of the preprint: **Davis JE**, Insigne KD, Jones EM, Hastings Q, Kosuri S. Multiplexed dissection of a model human transcription factor binding site architecture. bioRxiv preprint, 2019 May 02. doi: 10.1101/625434.

VITA

EDUCATION

University of California, Santa Barbara, Santa Barbara, CA 2010-2014

College of Letters and Sciences

BS, Biochemistry

WORK EXPERIENCE

Octant, Inc. 02/2018 - present

Research consultant

RESEARCH EXPERIENCE

University of California, Los Angeles, 07/2014 - 08/2019

Department of Chemistry and Biochemistry

Advisor: Dr. Sriram Kosuri

University of California, Santa Barbara 07/2013 - 06/2014

Department of Chemistry and Biochemistry

Advisor: Norbert Reich

University of California, Santa Barbara 03/2012 - 12/2012

Department of Chemistry and Biochemistry

Advisor: Javier Read de Alaniz

AWARDS AND HONORS

Ruth L Kirschstein National Research Service Award 07/2015 - 07/2018

University of California, Los Angeles, Cellular and Molecular Biology Training Program

PUBLICATIONS

Davis JE, Insigne KD, Jones EM, Hastings Q, Kosuri S. Multiplexed dissection of a model human transcription factor binding site architecture. bioRxiv preprint, 2019 May 02. doi: 10.1101/625434

Jones EM, Lubock NB, Venkatakrishnan A, Wang J, Tseng AM, Paggi JM, Latorraca NR, Cancilla D, Satyadi M, **Davis JE**, Babu MM, Dror RO, Kosuri S. Structural and functional characterization of G protein-coupled receptors with deep-mutational scanning. bioRxiv preprint, 2019 April 30. doi: 10.1101/623108

Fisher D, Palmer LI, Cook JE, **Davis JE**, Read de Alaniz J. Efficient synthesis of 4-hydroxycyclopentenones: dysprosium(III) triflate catalyzed Piancatelli rearrangement. Tetrahedron 2014 July 08;70:4105–4110.

CHAPTER ONE

Introduction

Gene regulation

Genes encode for molecules that define a cell and, on a larger scale, an organism. Yet it is the precise control over this process that enables biological complexity and survival. This regulation allows a cell to grow, reproduce, respond to environmental changes, and even differentiate into performing specific functions within an organism. The central dogma defines the backbone of this process, with DNA encoding for RNA (transcription) and RNA in turn encoding for proteins (translation), with many control and modification checkpoints scattered throughout these steps. RNA polymerase transcribes genes into RNA at one of the earliest points of control in this process. In eukaryotes, RNA polymerase II and its associated general transcription factors (TFs) bind elements in the genome, termed promoters, where this regulatory complex initiates transcription (Kornberg, 2007). Transcription factors, including the general factors of the pre-initiation complex, control the frequency of RNA polymerase II recruitment through interactions with other TFs, coregulators, the mediator complex, as well as to RNA polymerase itself (Bryant and Ptashne, 2003; Kornberg, 2007; Krumm et al., 1995; Lambert et al., 2018) in addition to modulating local nucleosome occupancy by recruiting chromatin-modifying cofactors (Lambert et al., 2018; Spitz and Furlong, 2012). Cis-regulatory elements of the genome, such as promoters and enhancers, contain binding motifs for specific TFs, orchestrating the number, type, and arrangement of these interactions (ENCODE Project Consortium, 2012; Lambert et al., 2018). Thus, it is this assortment and placement of TF-binding sites (TFBS) that ultimately establishes interaction networks driving the first step in RNA polymerase II-directed transcription.

Transcription factors drive transcription from cis-regulatory elements

There have been many efforts to characterize how combinations of TFBSs recruit transcription factors, and in turn, RNA Polymerase II. At a basic level, TF affinity for a sequence determines the duration of TF localization to a region of the genome. *In vitro* binding assays across thousands of synthetic sequences can determine TF position weight matrices (PWM), or the preference of a TF for each base in a binding site (Slattery et al., 2011; Tuerk and Gold, 1990). Additionally, *in vitro* binding experiments have shown sequences flanking binding sites also influence TF-binding affinity (Levo et al., 2015). However, the orientation and spacing between these experimentally-derived motifs can further influence TF affinity for sequences with various motif combinations (Jolma et al., 2013, 2015). While this complicates predicting TF binding even in controlled environments, this may explain the large differences observed between the number of predicted TF-binding sites in the genome and how many regions are actually bound via ChIP-seq experiments (Wasserman and Sandelin, 2004). In addition, TFs indicating preferential binding based off of the GC content surrounding their site (Dror et al., 2015) may also explain this phenomenon. Indeed, the likelihood of transcription factors to co-bind correlates with their GC content preference, with regulators that prefer similar content co-binding more frequently (Dror et al., 2015). Yet, TF-binding preferences determined from ChIP-seq alone can be confounded by indirect binding events (Worsley Hunt et al., 2014), biases to open chromatin (Spitz and Furlong, 2012), and suffers from the lack of the sequence diversity that can be explored in binding assays that allow precise control over sequence composition. Furthermore, while these methods identify preferential motifs and TF-bound regions, these experiments do not determine the transcriptional impact of these sequences when bound.

TF recruitment to a cis-regulatory element has variable effects on transcription due to a number of factors. The exact placement of TFBSs can be highly conserved in close proximity to core transcriptional machinery, such as surrounding transcription start sites (TSSs) of genes (Tabach et al., 2007), and such placement can be critical for transcriptional activity (Kim and Maniatis, 1997; Kim et al., 1998). TFBS not only organize TFs, but they dictate the phasing of these sites along the DNA helix and determine where short-range interactions with RNA Polymerase can occur, following ~10 bp intervals of activity (Kim et al., 2013). In addition to small changes in TFBS placement, the activities of certain TF indicate a dependence on overall TFBS proximity to the promoter/TSS (Tabach et al., 2007; Tinti et al., 1997). Lastly, some TFs require specific binding partners, such as other TFs and coactivators, to drive or inhibit transcription from a cis-regulatory element (Stampfel et al., 2015). Coupled with the lack of sequence diversity in cis-regulatory elements, this complicates the challenge of understanding how the precise composition of TFBSs within these regions results in gene expression. In order to piece apart the influence of these variables on transcription, many approaches have focused on a small set of changes performed to natural regions of the genome. Yet these studies are limited and confounded by the tested diversity of TFBS compositions, TFBS affinities, TFBS arrangements, distance of TFBSs to the TSS, and the sequence content surrounding TFBSs. Thus, we still lack a complete understanding of how the exact sequence composition of cis-regulatory elements, and combination of TFBSs wherein, results in specific quantitative gene regulatory responses.

Multiplexed approaches to characterizing TFBS logic

Thousands of synthetic regulatory elements designed and tested *in vivo* can isolate the effects of various cis-regulatory architectures on TFBS activity. Oligonucleotide microarray synthesis enables the multiplexed construction of large libraries (Fodor et al., 1991; Kosuri and

Church, 2014) to target such regulatory variables. By ensuring some way to track the activity of each sequence in a cell, one-pot experiments can then be performed, characterizing the transcriptional activity of library members in a single assay. Massively-parallel reporter assays (MPRAs) have emerged as one such method that utilize the scale of synthetic DNA libraries and next-generation sequencing to determine the expression of thousands of individual regulatory elements in pooled expression measurements (White, 2015). In many of these assays, library members are placed upstream of a cellular reporter and a unique transcribed region, called a barcode. These barcodes are either co-synthesized with library members or introduced following synthesis and associated to regulatory elements using next-generation sequencing (Urtecho et al., 2019). This reporter plasmid pool is then transfected or integrated into a cell line or organism and DNA and mRNA is collected. Next-generation sequencing is then used to identify barcode abundance in each sample, which in turn determines relative expression levels of barcodes in the pooled assay, and thus the level of activity of the library members associated with those barcodes. Thus MPRAs enable a high-throughput and multiplexed analysis of cis-regulatory architecture, expanding existing biological sequence space and isolating the effects of potentially confounding variables within cis-regulatory elements.

MPRAs have started to isolate the regulatory effects of features shaping eukaryotic cis-regulatory architecture. Many have used synthetic regulatory elements to determine the transcriptional effects of TFBS location (Sharon et al., 2012), sequences flanking TFBSs (Kwasnieski et al., 2012; Melnikov et al., 2012), active TF levels (van Dijk et al., 2017), homotypic site number (van Dijk et al., 2017; Gertz et al., 2009; Levo et al., 2017; Sharon et al., 2012; Smith et al., 2013; Weingarten-Gabbay et al., 2019; White et al., 2016), site strength (van Dijk et al., 2017; Gertz et al., 2009; Kheradpour et al., 2013; Sharon et al., 2012), and heterotypic

TFBS combinations (Fiore and Cohen, 2016; Grossman et al., 2017; Levo et al., 2017; Smith et al., 2013; White et al., 2016). Typically such approaches use combinations of different TFBSs derived from natural regions of the genome or sometimes arranged between synthetic sequences. Many of these approaches have confirmed the findings of older, small-scale manipulations but at a higher-resolution enabled by the increase in sequences tested. From these a general picture emerges that combinations of different TFBSs generally illicit a greater response than combinations of the same TFBS, although no formal combinatorial logic has been presented across TFBSs (Fiore and Cohen, 2016; Grossman et al., 2017; Levo et al., 2017; Smith et al., 2013; White et al., 2016). Site number in general follows a non-linear increase and saturation of expression, presumably due to saturated recruitment of RNA Polymerase II (Gertz et al., 2009; Sharon et al., 2012; Weingarten-Gabbay et al., 2019) although this response is shaped by active TF abundance in the cells tested (van Dijk et al., 2017). Additionally, one TFBS tested indicated a relationship between overall promoter proximity and transcriptional activity and in another case a ~10 bp periodicity in expression from changes in local TFBS placement (Sharon et al., 2012). Many of these approaches have balanced the number of TFBSs assayed with the number of various architectures studied. As such, we still lack a complete characterization of the many variables acting upon a single TFBS in driving transcription and how these variables may interact in combination in cis-regulatory elements.

MPRAs are employed to approximate regulatory features and trends expected to act upon natural regions in the genome. Yet, in order to easily test the cellular activity of thousands of sequences, many MPRAs are performed episomally in which 10^1-10^4 copies of library members are present per nucleus (Cohen et al., 2009). This eases the burden of the cell numbers required to cover library complexity and collected RNA and/or DNA amounts needed for next-generation sequencing. Yet, library members are assayed in an episomal rather than chromosomal context and are also in the presence of many other library members, potentially introducing binding competition for limited pools of available TF (Brewster et al., 2014; Lee and Maheshri, 2012). As an alternative to this format, some MPRAs have employed lentiviral library integration to reduce cellular copy numbers in addition to testing sequences in a genomic context (Inoue et al., 2017; Klein et al., 2019). Yet, lentiviral vectors integrate randomly into the genome such that each library member is assayed in a different genomic context. Additionally, template switching prior to lentiviral-based integration results in inaccurate barcode-library member associations (Hill et al., 2018; Sack et al., 2016), further confounding results and placing constraints on library designs to avoid this issue (Klein et al., 2019). MPRAs integrating library members into a single, specific locus of the genome per cell avoid the pitfalls of lentiviral assays in addition to assaying sequences in a constant genomic context. A recent MPRA used CRISPR/Cas9 and homology-directed repair to integrate library members into the genome of a human cell line and observed robust results (Weingarten-Gabbay et al., 2019). Although such experiments are presumed to approximate chromosomal activity of library members, such assays drastically extend the amount of time to perform MPRAs, require greater experimental planning, and require handling of more cells in addition to processing greater amounts of biological material for next-generation sequencing. As of yet, no work has compared the impact of MPRA assay format on sequence-function relationships surrounding TFBS architecture within cis-regulatory elements. Therefore, it is still not apparent if these more intensive approaches in a genomic context are necessary for approximating more natural genomic trends.

A high-resolution analysis of c-AMP response element logic using a massively-parallel reporter assay

Here we simplify approaches to understand TFBS logic within cis-regulatory elements by focusing on a single model human TFBS, the c-AMP Response Element (CRE). The CRE Binding (CREB) protein binds CRE and drives expression downstream of adenylyl cyclase activation (Gonzalez and Montminy, 1989; Montminy et al., 1986) across most cell types (Mayr and Montminy, 2001). CRE is ideally suited for exploring associations between TFBS architecture and regulation due to its ability to drive expression without other TFBSs in regulatory elements (Melnikov et al., 2012) and its ease of inducibility in a cell (Gonzalez and Montminy, 1989; Montminy et al., 1986), allowing finer control over active concentrations of the CREB protein. The most conserved, and likely to be functional, CREs generally localize within 200 basepairs (bp) of a TSS in the human genome (Mayr and Montminy, 2001; Zhang et al., 2005). Additionally, a previous MPRA that performed scanning mutagenesis on a commercial CRE reporter found mutations to CREs in closer proximity to the promoter had a greater effect on expression in addition to mutations to sequences flanking CREs (Melnikov et al., 2012). Here we explore the relationship between CRE's distance to promoter elements and its activity in greater detail, placing CRE in 1 bp intervals away from a promoter. We further explore the role other regulatory features play in modulating CREB protein activity including: CRE affinity, number, the spacing between multiple CREs, and the surrounding sequence content. Finally, we test our library both transiently and in a newly developed, singly-integrated genomic MPRA to better understand the quantitative and sometimes subtle differences between genomic and episomal assay context.

REFERENCES

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

Bryant, G.O., and Ptashne, M. (2003). Independent Recruitment In Vivo by Gal4 of Two Complexes Required for Transcription. Mol. Cell *11*, 1301–1309.

Cohen, R.N., van der Aa, M.A.E.M., Macaraeg, N., Lee, A.P., Szoka, F.C., and Jr. (2009). Quantification of plasmid DNA copies in the nucleus after lipoplex and polyplex transfection. J. Control. Release *135*, 166–174.

van Dijk, D., Sharon, E., Lotan-Pompan, M., Weinberger, A., Segal, E., and Carey, L.B. (2017). Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators. Genome Res. *27*, 87–94.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment on transcription factor binding across diverse protein families. Genome Res. *25*, 1268–1280.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fiore, C., and Cohen, B.A. (2016). Interactions between pluripotency factors specify *cis* - regulation in embryonic stem cells. Genome Res. *26*, 778–786.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. Science *251*, 767–773.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.

Gonzalez, G.A., and Montminy, M.R. (1989). Cyclic AMP stimulates somatostatin gene transcription by phosphorylation of CREB at serine 133. Cell *59*, 675–680.

Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E., et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. U. S. A. 201621150.

Hill, A.J., McFaline-Figueroa, J.L., Starita, L.M., Gasperini, M.J., Matreyek, K.A., Packer, J., Jackson, D., Shendure, J., and Trapnell, C. (2018). On the design of CRISPR-based single-cell molecular screens. Nat. Methods *15*, 271–274.

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., and Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. *27*, 38–52.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *advance on*.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811.

Kim, T.K., and Maniatis, T. (1997). The mechanism of transcriptional synergy of an in vitro assembled interferon-beta enhanceosome. Mol. Cell *1*, 119–129.

Kim, S., Broströmer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y.Q., et al. (2013). Probing allostery through DNA. Science *339*, 816–819.

Kim, T.K., Kim, T.H., and Maniatis, T. (1998). Efficient recruitment of TFIIB and CBP-RNA polymerase II holoenzyme by an interferon-beta enhanceosome in vitro. Proc. Natl. Acad. Sci. U. S. A. *95*, 12191–12196.

Klein, J., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2019). A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays.

Kornberg, R.D. (2007). The molecular basis of eukaryotic transcription. Proc. Natl. Acad. Sci. U. S. A. *104*, 12955–12961.

Kosuri, S., and Church, G.M. (2014). Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods *11*, 499–507.

Krumm, A., Hickey, L.B., and Groudine, M. (1995). Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. Genes Dev. *9*, 559–572.

Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc. Natl. Acad. Sci. U. S. A. *109*, 19498–19503.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665.

Lee, T.-H., and Maheshri, N. (2012). A regulatory role for repeated decoy transcription factor binding sites in target gene expression. Mol. Syst. Biol. *8*, 576.

Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A.C., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. Genome Res. *25*, 1018–1029.

Levo, M., Avnit-Sagi, T., Lotan-Pompan, M., Kalma, Y., Weinberger, A., Yakhini, Z., and Segal, E. (2017). Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. Mol. Cell *65*, 604–617.e6.

Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor CREB. Nat. Rev. Mol. Cell Biol. *2*, 599–609.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–

277.

Montminy, M.R., Sevarino, K.A., Wagner, J.A., Mandel, G., and Goodman, R.H. (1986). Identification of a cyclic-AMP-responsive element within the rat somatostatin gene. Proc. Natl. Acad. Sci. U. S. A. *83*, 6682–6686.

Sack, L.M., Davoli, T., Xu, Q., Li, M.Z., and Elledge, S.J. (2016). Sources of Error in Mammalian Genetic Screens. G3 *6*, 2781–2790.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *advance on*.

Tabach, Y., Brosh, R., Buganim, Y., Reiner, A., Zuk, O., Yitzhaky, A., Koudritsky, M., Rotter, V., and Domany, E. (2007). Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. PLoS One *2*, e807.

Tinti, C., Yang, C., Seo, H., Conti, B., Kim, C., Joh, T.H., and Kim, K.S. (1997). Structure/function relationship of the cAMP response element in tyrosine hydroxylase gene transcription. J. Biol. Chem. *272*, 19158–19164.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505–510.

Urtecho, G., Tripp, A.D., Insigne, K.D., Kim, H., and Kosuri, S. (2019). Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. Biochemistry *58*, 1539–1551.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. *5*, 276–287.

Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A., and Segal, E. (2019). Systematic interrogation of human promoters. Genome Res. *29*, 171–183.

White, M.A. (2015). Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. Genomics *106*, 165–170.

White, M.A., Kwasnieski, J.C., Myers, C.A., Shen, S.Q., Corbo, J.C., and Cohen, B.A. (2016). A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors.

Worsley Hunt, R., Mathelier, A., Del Peso, L., and Wasserman, W.W. (2014). Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. BMC Genomics *15*, 472.

Zhang, X., Odom, D.T., Koo, S.-H., Conkright, M.D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., et al. (2005). Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. Proc. Natl. Acad. Sci. U. S. A. *102*, 4459–4464.

CHAPTER TWO

Multiplexed dissection of a model human transcription factor binding site architecture

Title: Multiplexed dissection of a model human transcription factor binding site architecture

Authors: Jessica E. Davis[1], Kimberly D. Insigne[1,2], Eric M. Jones[1,‡], Quinn B Hastings[1], Sriram Kosuri[1]*

Author affiliations: [1]Department of Chemistry and Biochemistry, UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, and Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA 90095, USA [2]Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, Los Angeles, CA 90095, USA

Current Address:

‡Octant, Inc. 570 Westwood Plaza, Los Angeles, CA 90095

*Correspondence should be addressed to S.K. (sri@ucla.edu)

Abstract

In eukaryotes, transcription factors orchestrate gene expression by binding to TF-Binding Sites (TFBSs) and localizing transcriptional co-regulators and RNA Polymerase II to cis-regulatory elements. The strength and regulation of transcription can be modulated by a variety of factors including TFBS composition, TFBS affinity and number, distance between TFBSs, distance of TFBSs to transcription start sites, and epigenetic modifications. We still lack a basic comprehension of how such variables shaping cis-regulatory architecture culminate in quantitative transcriptional responses. Here we explored how such factors determine the transcriptional activity of a model transcription factor, the c-AMP Response Element (CRE) binding protein. We measured expression driven by 9,126 synthetic regulatory elements in a massively parallel reporter assay (MPRA) exploring the impact of CRE number, affinity, distance to the promoter, and spacing between multiple CREs. We found the number and affinity of CREs within regulatory elements largely determines overall expression, and this relationship is shaped by the proximity of each CRE to the downstream promoter. In addition, while we observed expression periodicity as the CRE distance to the promoter varied, the spacing between multiple CREs altered this periodicity. Finally, we compare library expression between an episomal MPRA and a new, genomically-integrated MPRA in which a single synthetic regulatory element is present per cell at a defined locus. We observe that these largely recapitulate each other although weaker, non-canonical CREs exhibited greater activity in the genomic context.

Introduction

The ability for organisms to precisely control gene expression levels and responses is crucial for almost all biological processes. Expression levels are controlled by *cis*-regulatory elements such as promoters and enhancers, *trans*-acting factors such as transcription factors (TFs), and cell, epigenetic and environmental states. Cis-regulatory elements help direct transcription responses by localizing and orchestrating interactions between active transcription factors, co-regulators, and RNA Polymerase II (ENCODE Project Consortium, 2012; Lambert et al., 2018). For control of human gene expression, a variety of large-scale projects seek to determine gene expression levels across various cell lines and cell types (Lizio et al., 2015, 2017), identifying functional elements that might control expression (ENCODE Project

Consortium, 2012), the genome-wide characterization of epigenetic states of DNA (Roadmap Epigenomics Consortium et al., 2015), and the binding specificities of transcription factors (Jolma et al., 2013, 2015; Yin et al., 2017; Zhu et al., 2018). Collectively, while these efforts generally give us a parts list of putatively functional elements, understanding how these parts define quantitative levels of expression is still not well understood.

The combination of sequence motifs that recruit TFs, or TF-binding sites (TFBS), functionalize cis-regulatory elements via unique arrangements that help determine quantitative regulatory responses (Lambert et al., 2018; Spitz and Furlong, 2012). The consequences of subtle changes to TFBS compositions can be drastic. For example, clusters of weak-affinity Gal4 sites in yeast promoters increases expression synergistically, while stronger-affinity sites contributing to expression additively (Giniger and Ptashne, 1988). There can also be differences in TF occupancy of similar sequences in the genome that follow differences in the GC content of the surrounding sequence (Dror et al., 2015). Additionally, the placement of TFBSs can be highly conserved in close proximity to core transcriptional machinery, such as surrounding transcription start sites (TSSs) of genes (Tabach et al., 2007), and such placement can be critical for transcriptional activity (Kim and Maniatis, 1997; Kim et al., 1998). Lastly, the positional arrangement of TFBS combinations within cis-regulatory elements can modulate TF binding strength (Jolma et al., 2013, 2015) and TF activity can vary across the composition of TFBS combinations (Stampfel et al., 2015). Deciphering the logic imbued in cis-regulatory elements is difficult, as the limited set of natural variants and cell types are typically insufficient to control for variables such as sequence composition, TFBS composition and arrangements, and activity of trans-acting factors. Proving that particular sequences have causative effects on gene expression requires carefully controlled and high-throughput reverse-genetic studies.

The emergence of the massively parallel reporter assay (MPRA) allows for the testing of such reverse genetic transcriptional assays, and has become a powerful tool for the large-scale functional validation of regulatory elements across genomic and organismal contexts (White, 2015). These assays utilize the scale of synthetic DNA libraries and next-gen sequencing to determine the expression of thousands of individual regulatory elements in pooled expression measurements, enabling high-throughput functional characterizations of cis-regulatory logic. MPRAs have been used to quantify the transcriptional strengths of cis-regulatory elements and identify the motifs integral to element activity (Ernst et al., 2016; Kheradpour et al., 2013). Furthermore, several groups are using these systems to dissect how individual TFBSs drive quantitative regulatory responses in bacteria (Belliveau et al., 2018), yeast (van Dijk et al., 2017; Gertz et al., 2009; Levo et al., 2017; Sharon et al., 2012), human cell lines (Fiore and Cohen, 2016; Grossman et al., 2017; Weingarten-Gabbay et al., 2019), and animals (Kwasnieski et al., 2012; Smith et al., 2013; White et al., 2016). Collectively, these studies have begun to dissect TFBS logic by exploring how the regulatory grammar of different site combinations, numbers, and placements affect transcriptional activity.

Here we focus on how a range of factors guiding cis-regulatory architecture shape the activity of a single TFBS, the c-AMP Response Element (CRE). The CRE Binding (CREB) protein binds CRE and drives expression downstream of adenylyl cyclase activation (Gonzalez and Montminy, 1989; Montminy et al., 1986) across most cell types (Mayr and Montminy, 2001). CRE is ideally suited for exploring associations between TFBS architecture and regulation due to its ability to drive expression without other TFBSs in regulatory elements (Melnikov et al., 2012) and its ease of inducibility in a cell (Gonzalez and Montminy, 1989; Montminy et al., 1986), allowing finer control over active concentrations of the CREB protein.

The most conserved, and likely to be functional, CREs generally localize within 200 basepairs (bp) of a TSS in the human genome (Mayr and Montminy, 2001; Zhang et al., 2005). Additionally, a previous MPRA that performed scanning mutagenesis on a commercial CRE reporter found mutations to CREs in closer proximity to the promoter had a greater effect on expression in addition to mutations to sequences flanking CREs (Melnikov et al., 2012). Here we explore the relationship between CRE's distance to promoter elements and its activity in greater detail. We further explore the role other regulatory features play in modulating CREB protein activity including: CRE affinity, number, the spacing between multiple CREs, and the surrounding sequence content. Finally, although many MPRAs are performed episomally due to their ease and quickness (Fiore and Cohen, 2016; Grossman et al., 2017; Kheradpour et al., 2013; Melnikov et al., 2012), it's been observed that episomal cis-regulatory element expression does not always correlate with their genomic counterparts (Inoue et al., 2017; Klein et al., 2019). Thus, we test our library both transiently and in a newly developed, singly-integrated genomic MPRA to better understand the quantitative and sometimes subtle differences between genomic and episomal assay context.

Results

CRE MPRA design and assay

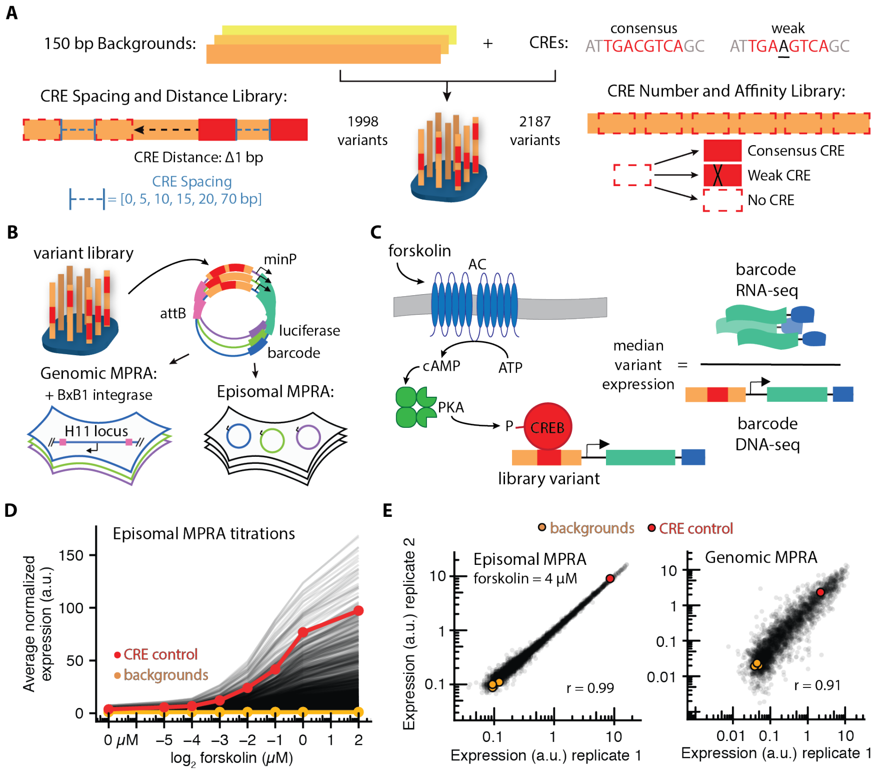


Figure 2.1 CRE regulatory library design, synthesis, and assays. (A) We replaced sequence within three putatively inactive backgrounds with consensus and/or mutant CREs to generate two libraries with varying spacing and distance to the minimal promoter (CRE Spacing and Distance Library) or the number and strength of the binding (CRE Number and Affinity Library). (B) We assembled the library reporter vectors with variants upstream of a minimal promoter, luciferase ORF, and a unique barcode sequence. The pool of library reporter vectors was assayed in a HEK293T cell line either by transient transfection or integrated at one allele of the Human H11 locus. (C) We stimulated CREB protein activity with forskolin, which activates cAMP signaling. Expression levels were determined as the ratio of barcodes reads in the RNA to that of the DNA sample. (D) Library expression following increasing forskolin concentrations in the episomal MPRA. Each variant is normalized to the expression of their corresponding background (no CREs), then averaged across replicates. (E) Variant expression exhibits strong correlation between biological replicates in both MPRAs, here shown at maximally-inducing concentrations of forskolin in both assays (episomal r = 0.99 and genomic r = 0.91).

We designed libraries with one or more CRE(s) by replacing sequence within three putatively inactive 150 bp background sequences to assay a range of features contributing to TFBS architecture (Figure 2.1 A). These backgrounds were adapted from sequences with little reported activity from the Vista Enhancer Database (Visel et al., 2007) or a commercial reporter modified by removing previously identified CREs (Fan and Wood, 2007). We generated regulatory variants by replacing 12 bp regions of the backgrounds with either the consensus CRE (AT *TGACGTCA* GC), in which the central 8 bp region binds a CREB dimer (two monomer binding sites), or a weaker CRE (AT *TGAAGTCA* GC), where one of the central dinucleotides bound by both monomers was mutated and has been previously shown to reduce activity (Mayr and Montminy, 2001; Melnikov et al., 2012). For the majority of analysis, we used two CRE libraries. The first library, the CRE Spacing and Distance Library, assays CREB activity as a function of both the spacing between CREs and CRE distance to the minimal promoter by moving two consensus CREs across the 150 bp backgrounds at six defined spacings (0, 5, 10, 15, 20 and 70 bp) between the two sites. In the CRE Number and Affinity Library, we explore the effect of both CRE number and affinity upon expression by designating 6 equally-spaced locations across all backgrounds in which each location is replaced with either the weak CRE, consensus CRE, or no CRE.

We used Agilent OLS synthesis to construct our designed libraries, added random 20 nt barcodes to the 3' end, and mapped these barcode-variant associations. For MPRA analysis we only considered barcodes corresponding to perfect matches to our designs, identified at this stage via next-gen sequencing. We cloned these libraries into a reporter construct we engineered to maximize signal to noise when integrated into the genome (Figure 2.6 C) and then cloned a minimal promoter and luciferase gene between variant and barcode, placing the barcode in the 3' UTR of the luciferase gene. The assays were conducted at varying induction conditions, either episomally by transient transfection (Episomal MPRA), or singly-integrated into the intergenic H11 safe-harbor locus (Zhu et al., 2014) using BxBI-mediated recombination (Genomic MPRA) (Duportet et al., 2014; Jones et al., 2019; Matreyek et al., 2017; Xu et al., 2013) (Figure 2.1 B). The episomal MPRAs were run in biological duplicate across 14 different concentrations of forskolin (Figure 2.1 D, Figure 2.6 E, and Figure 2.7), which stimulates phosphorylation and activation of the CREB protein by activating adenylyl cyclase (Gonzalez and Montminy, 1989). The genomic MPRA was run in biological duplicate at full induction (Figure 2.6 D). After forskolin stimulations, we isolated barcoded transcripts from cells and used next-gen sequencing to determine barcode prevalence per RNA samples and plasmid (episomal MPRA) or genomic (genomic MPRA) DNA samples. Since each variant was mapped to multiple barcodes, we first determined the expression of each barcode via the ratio of normalized reads in the RNA over the DNA sample. We then determined variant expression from the median expression of all barcodes mapped to each variant. Both episomal and genomic MPRAs indicated high reproducibility between separately stimulated replicates (Figure 2.1 E, episomal Pearson's *r* = 0.99, genomic *r* = 0.91, and Figure 2.6 E). In both assays, the difference between backgrounds alone and a positive CRE control adapted from a commercially-available reporter plasmid (Fan and Wood, 2007) spanned the majority of expression variation amongst variants.

The role of CRE spacing and distance on expression

We explored the extent to which positioning of CRE within a regulatory element quantitatively affects its transcriptional activity. We initially assayed the relationship between CRE distance relative to a downstream promoter and variant expression using 1 consensus CRE in a separate library, but found it drove minimal expression in the episomal MPRA after CREB activation (Figure 2.8). We then examined the expression driven by the CRE Spacing and Distance Library. This library varied the relative positioning of two consensus CREs with respect to the minimal promoter (referred to as CRE *distance*), and altered the number of nucleotides between the two sites (referred to as CRE *spacing*) (Figure 2.2 A). We tested *spacings* of 0, 5, 10, 15, 20 or 70 bp between the two CREs, and then tested *distance* by moving these sites with each of the *spacings* across the backgrounds one base at a time, spanning the 150 bp backgrounds. Activation occurs in a dose-dependent manner in the episomal MPRA and we observe a ~10 bp expression periodicity that is more apparent at higher concentrations of

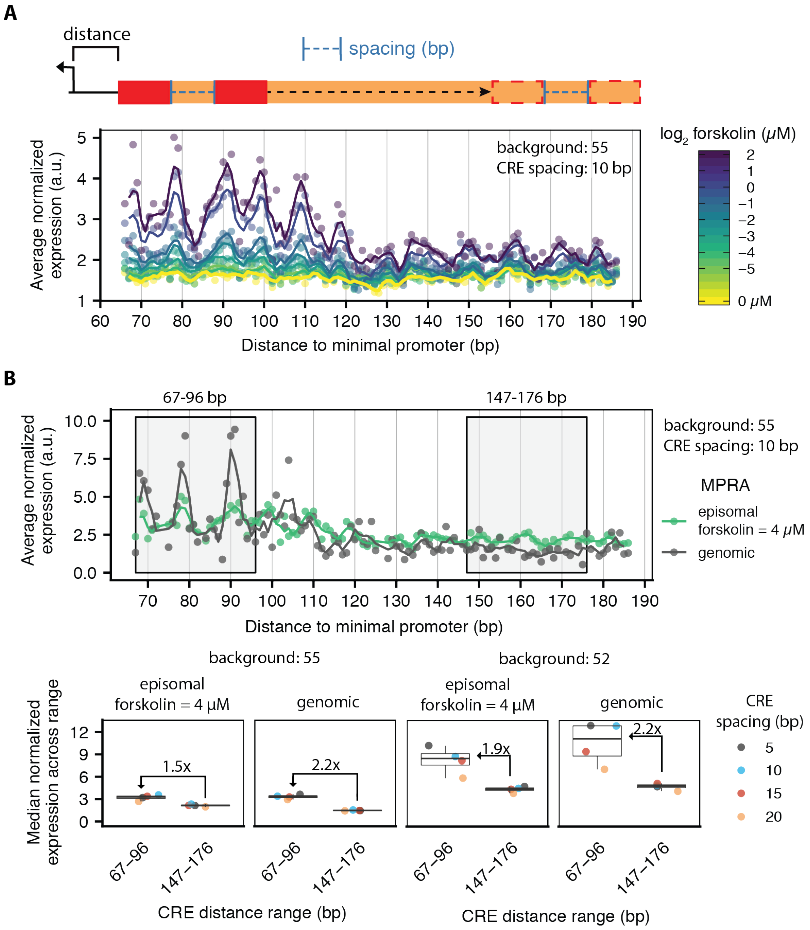


Figure 2.2 CRE proximity to promoter elements is associated with higher expression. (A) The expression profiles for two CREs that are 10 bp apart display a periodic signal as they are moved away from the minimal promoter for Background 55 (with forskolin concentration shown in color). The lines are 3 bp moving averages of the points. (B) Expression profiles for variants with Background 55 and 10 bp CRE *spacing* at maximally-inducing concentrations for both episomal and transient MPRAs (top panel). Expression decreases (lower panel) as the distance of the CREs from the proximal promoters increases across the backgrounds, spacings, and MPRA formats (as measured by the median expression across *distance* ranges 67-96 bp and 147-176 bp).

forskolin. This 10 bp periodicity was consistent across CRE *spacings* and backgrounds, displayed similar patterns between the genomic and episomal assays, but differed between backgrounds (Figure 2.9). Such periodicity has been observed before for single TFBSs in a variety of model systems (Kim et al., 2013; Sharon et al., 2012; Takahashi et al., 1986). In addition, we also observed a general decrease in expression as CRE *distance* increased. Across backgrounds, MPRA formats, and CRE *spacings*, the change in median expression between the CRE *distance* range of 67-96 bp and 147-176 bp resulted in a median 1.5 to 2.2-fold decrease in expression, with larger effects observed in the genomic MPRA (Figure 2.2 B).

In addition, across CRE *spacing,* we noticed different phasings of expression periodicity. This was most pronounced in variants with background 41 in the episomal MPRA (Figure 2.3

A). In particular, 5 and 15 bp CRE *spacings* exhibited similar local expression maxima at CRE *distances* 78 and 88 bp, whereas 10 and 20 bp *spacings* had maxima at 83 and 92 bp. This ~5 bp shift was also observed in background 55, albeit at different *distances* (Figure 2.10); it is of note we did not observe as clear changes in periodicity phasing in a genomic context. In these instances, we indicate *distance* from the minimal promoter to the start of the first CRE, such that the proximal CRE is in the same position across all expression profiles. Thus, the only differences between variants that may be causing this periodicity shift is the altered placement of the distal CRE following CRE *spacings*. Using the CREB bZIP structure bound to CRE (Schumacher et al., 2000), we modeled 2 dimers bound to CREs with 5 and 10 bp *spacings* (discussed in Methods) by aligning protein-DNA density to DNA backbone (Figure 2.3 B). It is of note, this simplified model does not incorporate density from full-length CREB proteins or both proteins' effects on local DNA bending. A 5 bp shift in *distances* driving expression maxima between 5 and 10 bp *spacings* corresponds to about half a helical turn of B-form DNA (10.4 bp/turn). Our model positions the proximal CREB dimer (1) on the opposite face of the DNA helix between 5 and 10 bp *spacings* when they are both at their expression maxima. On the other hand, the distal CREB dimer (2) would be similarly oriented between the 5 and 10 bp *spacings*, but at a full helical turn distance from one another on the DNA.

To explore this phenomenon further, we designed a follow-up CRE Spacing and Distance Library varying *spacings* from 1-13 bp to encompass a full helical turn of relative CREB orientations. According to these *spacings* and the 8 bp CRE length, we expect CREB proteins to be co-aligned along the helix at both 2-3 bp and 12-13 bp *spacings* and we expect CREB proteins to lie on opposite sides of the helix at 7-8 bp *spacings*. The same backgrounds were used as before and CREs were similarly placed along these backgrounds at 1 bp intervals conserving

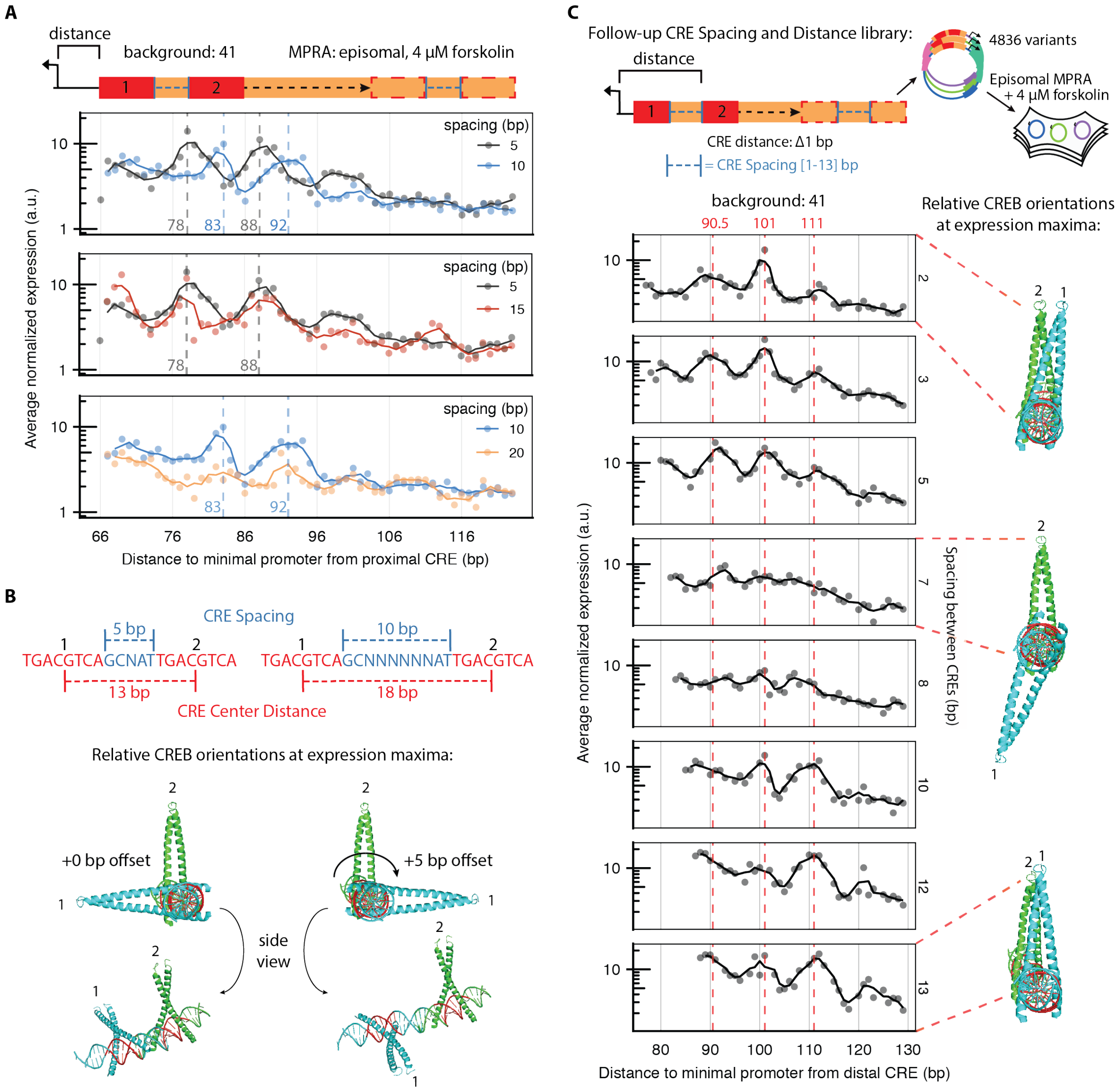


Figure 2.3 CRE *spacing* modulates expression periodicity as CRE *distance* is varied. (A) Average background-normalized expression of variants with background 41 in the episomal MPRA plotted as a function of *distance* between the minimal promoter and proximal CRE. Solid lines correspond to the 3bp moving average estimate and dashed lines indicate local expression maxima determined from 5 and 10 bp CRE *spacings*. Overlays indicate the offset of expression periodicities between 5 and 10 bp *spacings* and alignment between both 5 and 15 and 10 and 20 bp *spacings*. (B) Using the published structure of a CREB::bZIP dimer bound to CRE (PDB: 1DH3), we modeled the expected positioning of CREB dimers bound to the CRE proximal (1) and distal (2) to the promoter for both the 5 and 10 bp CRE *spacings* at the *distances* at their respective local maxima in A. Modeling approximates the distal dimer (2) in similar orientations at the two local expression maxima between *spacings*, with the proximal dimer on opposite faces of the DNA. (C) An additional CRE Spacing and Distance library was synthesized with CREs placed at all locations along the backgrounds with constant 1-13 bp *spacings*. Average background-normalized expression of variants with background 41 in a similarly-performed episomal MPRA are plotted as a function of *distance* between the minimal promoter and distal CRE. Across *spacings*, the placement of the distal CRE determines expression periodicity. Coupled with CREB dimer modeling across CREs with spacings, expression is minimal when both CREB proteins lie on opposite sides of the helix at 7-8 bp *spacings*.

indicated *spacings* (Figure 2.3 C). When plotting the episomal MPRA expression of variants now based on the *distance* from the distal CRE (2) to the promoter, we noticed a conserved pattern of ~10 bp expression periodicity across *spacings* (Figure 2.11). This alignment of expression periodicities across *spacings* was also observed across the different backgrounds. In contrast to the pronounced patterns observed with 12-13 bp *spacings*, in which CREB proteins are modeled to bind on the same face of the helix, expression was dampened at 7-8 bp *spacings* along the backgrounds, where CREB proteins are modeled to lie on opposite faces of the helix. We also noticed diminished expression for the short 1 bp *spacing*, in which we reasoned there may be binding site competition between CREs. Thus, it seems that the placement of the distal CRE (2), and presumably the orientation of the distal CREB protein, drives expression periodicity in these regions, while the placement of the proximal CRE (1) determines the amplitude of this response.

The role of CRE number and affinity upon expression

While CRE *distance* and *spacing* in regulatory elements help shape CRE's activity, the overall number and affinity of CREs likely plays a larger role in determining expression. The design of the CRE Number and Affinity Library assays these effects while taking into consideration the contribution of CRE position. Each variant in this library contained unique combinations of consensus and weaker-affinity CREs spanning an assortment of 6 positions along the backgrounds (Figure 2.1 A). A constant 17 bp CRE *spacing* was implemented in this library design in order to sample a range of predicted CREB protein orientations along the DNA across the 6 positions (Figure 2.12). Even so, the number of consensus CREs alone largely determined variant expression and this relationship followed a non-linear increase and eventual plateauing of expression in both MPRA formats (Figure 2.4 A and Figure 2.12).

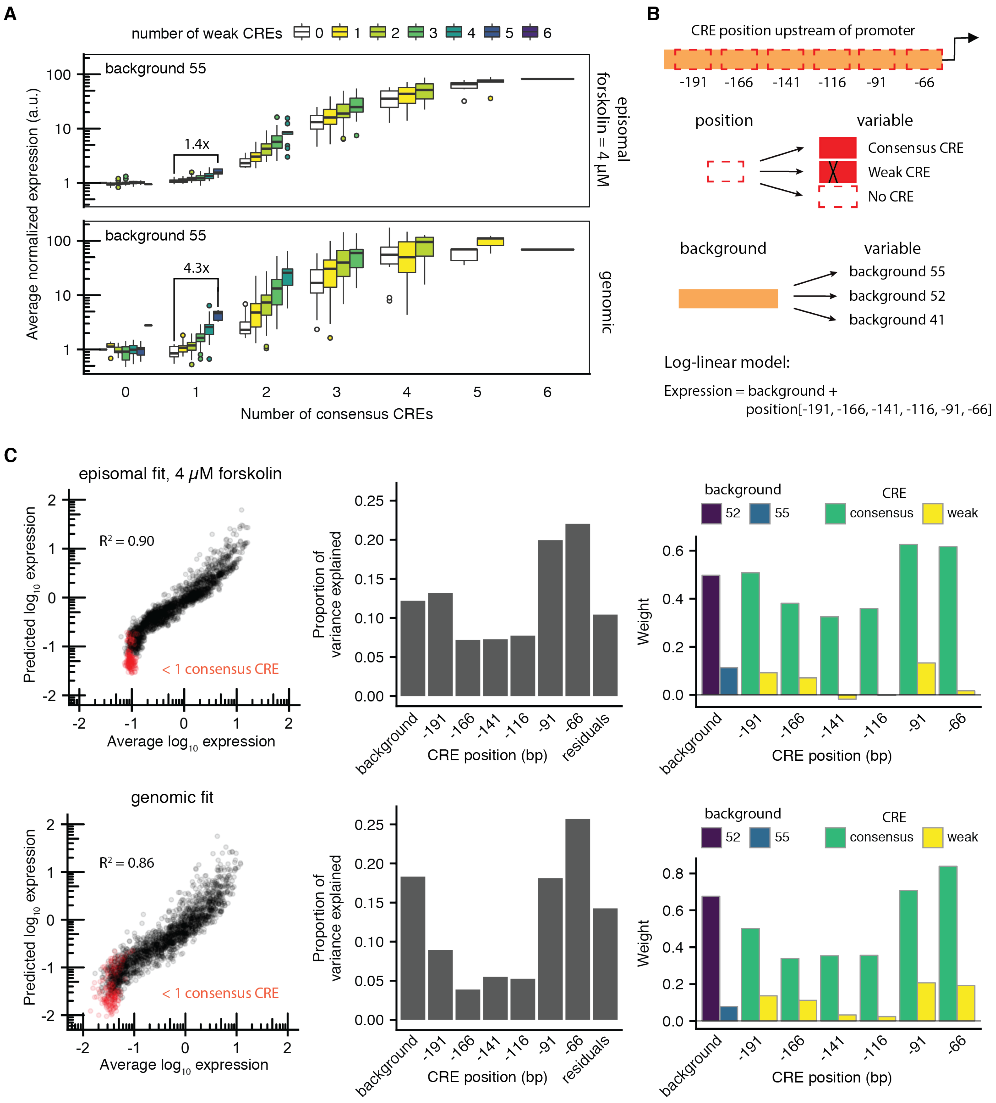


Figure 2.4 CRE number and affinity largely determines variant expression with variation explained by CRE position and background. (A) The expression of variants with background 55 according to their number of consensus (x-axis) and weak (colored subsets) CREs in the integrated and episomal MPRAs. The change in median expression between variants with 1 consensus CRE and those with 1 consensus and 5 weak CREs is indicated. (B) A simple linear model was fit to log-transformed expression using the identities of the background and TFBSs at each position as inputs. (C) This fit model correlates well with measured expression for episomal ($R^2 = 0.90$) and genomic ($R^2 = 0.86$) MPRAs and deviates from measurements of variants with no consensus CREs (left panel). Analysis of variance indicates 90% episomal and 86% genomic variance in expression is explained by the model. CREs occupying the two closest positions to the promoter had the strongest effects. The weights of categorical variables show the relative effects of strong and weak CREs and the effect of each background relative to variants with no CREs and with background 41.

We observed a similar increase with the number of weak CREs if at least one consensus CRE was also present within the variant, although this effect varied per background and between episomal and genomic MPRAs. While the number of consensus CREs largely determined variant expression, there was a large amount of expression variability per arrangement of similar numbers of consensus and weak CREs, perhaps due to combinations of CREB protein orientations. To explore how the different arrangements of CREs across the six positions shaped expression per CRE combination, we fit a log-linear model of expression to the independent contributions of CREs at each position (Figure 2.4 B). We allowed different weights to be fit per CRE affinity per position and also included an independent background term to account for expression differences between backgrounds.

We found this independent, position-specific model explained a majority of expression (Figure 2.4 C, left panels) in both the episomal (r = 0.95) and genomic MPRAs (r = 0.93). Although the model was inaccurate at predicting activity from low-expressing variants, this was largely due to variants with weak CREs and no consensus CREs driving little variation in activity in our assay. Accordingly, we found that CREs closest to the promoter (66 and 91 bp upstream of the promoter) explained 42.6% of the variance in the episomal MPRA and 43.8% in the genomic MPRA (Figure 2.4 C). This is expected as both positions fell within the ~110 bp of higher expression observed with the CRE Spacing and Distance library. None of the weights fit to CRE positions followed a trend in predicted CREB protein orientations (Figure 2.12), thus CRE's distance to promoter elements may mask the subtle effects of CREB protein orientation previously observed with 2 CREs. Background alone explained 12.1% of the variance episomally and 18.4% of genomic expression variance. Overall, while we find that the number of consensus CREs per variant largely determined expression, the combination of CRE positions along the backgrounds and the backgrounds themselves explained a majority of expression in our assays.

Differences between episomal and genomic MPRAs

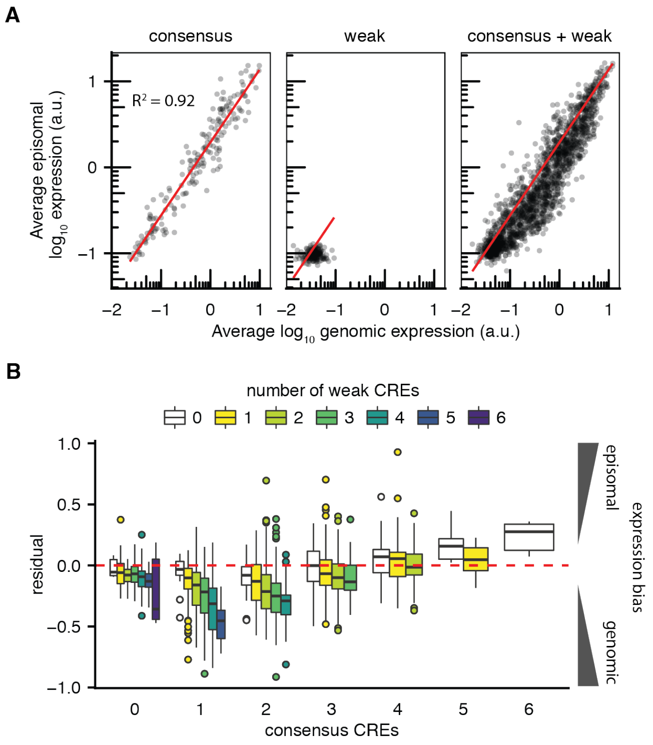


Figure 2.5. Variants with weak CREs exhibit higher relative expression in a genomic context. (A) Variants across all backgrounds subset according to site affinity composition. Subsets include variants with 1-6 consensus CREs (left panel), 1-6 weak CREs (middle) and combinations of 1-6 weak and consensus CREs (right). Variants with only consensus CREs drive similar relative expression (*R^2 = 0.92*, red line) between genomic and episomal MPRAs. Most variants with weak CREs drive higher relative expression in the genomic MPRA compared to variants without weak CREs. (B) Variant expression according to their number of both consensus (x-axis) and weak CREs (colored subsets). Using the expression correlation line between variants with 1-6 consensus CREs in both MPRAs (red line) as a reference, the residual of each variant to this line is plotted along the y-axis. Variants containing higher numbers of weak CREs drive higher relative expression in the genomic MPRA while those with higher numbers of consensus CREs drive higher relative expression in the episomal MPRA.

There were a number of differences between the genomic and episomal expression trends resulting from increasing CRE numbers per variant. First, variants with weak CREs drove greater expression in the genomic MPRA, especially within the context of few consensus CREs per variant. For example, the change in median expression between variants with one consensus CRE and those with one consensus CRE and five weak CREs within background 55 was 1.4-fold in the episomal MPRA and 4.3-fold in the genomic MPRA (Figure 2.4 A). Second, expression plateaued at about six consensus CREs for most backgrounds and across forskolin concentrations in the episomal MPRA (Figure 2.13), while in the genomic MPRA, this occurred at about four consensus CREs for all backgrounds (Figure 2.4 A and Figure 2.13). To explore differences in expression trends between the two MPRAs, we compared expression between the two assays according to different combinations of CRE affinities within variants (Figure 2.5 A). There was a strong linear relationship between genomic and episomal expression for variants containing one to six consensus CREs (Figure 2.5 A left panel, $R^2 = 0.92$). Using this linear relationship as a reference (red line), we then compared expression between the two assays for variants containing one to six weak CREs (Figure 2.5 B middle panel) and combinations of one to six consensus and weak CREs (Figure 2.5 B right panel). Expression deviated from this linear relationship for variants containing weak CREs, with 79% of the variants exhibiting higher relative expression in the genomic MPRA.

When we broke down this comparison according to numbers of consensus and weak CREs per variant, we observed higher relative genomic expression of weak CREs in the context of few (0-3) consensus CREs and higher relative episomal expression with combinations of at least 4 consensus CREs (Figure 2.5 B). This effect may be due to the differences in CRE variant copy numbers between assays. Lipofectamine transfections can result in 10^1-10^4 plasmids/nucleus (Cohen et al., 2009). In this context, high numbers of variants with many consensus CREs may out-compete variants with weak CREs for CREB-protein binding. In similar systems, the titration of plasmids containing the same TFBS as a single chromosomal reporter has altered the expression of the chromosomal reporter (Lee and Maheshri, 2012) in a manner dependent upon the strength of the plasmid "competitor" sites and cellular TF concentrations (Brewster et al., 2014). Although multi-copy MPRAs are ideal for assaying large libraries of variants by reducing the number of cells required to cover library diversity, expression may be interpreted in the context of variant competition when variants contain similar TFBSs. Despite this, we show MPRAs that assay many variants per cell can largely recapitulate other TFBS features shaping genomic expression driven from single variants.

Discussion

By performing synthetic manipulations of a single TFBS, the CRE, we present here a characterization of various regulatory rules governing TF activity. In particular, we assay the effects of CRE number, affinity, distance to promoter, and the spacing between multiple CREs within regulatory elements. Additionally, we show how a subset of these features shape expression when tested in combination. The limited complexity of natural cis-regulatory elements would not have allowed us to explore these features at bp-resolution and in such a controlled manner. Thus, we chose to isolate these features using synthesized regulatory elements in the format of an MPRA. Furthermore, we integrated variants at a single copy per cell within the same genomic environment to better approximate genomic expression. We present here improvements in a single-copy, defined-locus genomic MPRA performed in a human cell line, indicating the feasibility of reproducible expression measurements of lowly-expressing transcripts.

Although the variants assayed here are synthetic, we expect the regulatory trends observed to define expression from natural cis-regulatory elements as well. We observed a drop in transcription with placements of CRE beyond 120 bp *distance* to our minimal promoter, following similar findings with manipulation of native CREs (Tinti et al., 1997). Although it is of note that diminished expression periodicity was still observed at *distances* up to ~180 bp (Figure 2.2 and Figure 2.9). CREB protein recruits RNA polymerase II through interactions with TFIID, while it's phosphorylated form drives polymerase isomerization and transcription (Kim et al., 2000). Thus CRE's proximity to promoter elements may be integral to polymerase recruitment and transcription, perhaps explaining CRE's enriched localization within 200 bp of TSSs in the human genome (Mayr and Montminy, 2001; Zhang et al., 2005). TF's that also directly recruit and activate RNA Pol II may also follow similar trends in activity according to TFBS promoter proximity.

In addition to overall distance effects, CRE's precise positioning likely plays a role in its activity in natural regulatory elements via the periodicity observed. In line with CRE's localization around TSSs, conserved TFBSs that exhibit location-specificity in the genome are mostly found between 200 bp upstream and 100 bp downstream of a TSS (Tabach et al., 2007). For instance, manipulations that place the highly conserved 8 TFBSs in the IFN-ß enhanceosome at half a helical turn from the original 47 bp upstream of the TSS (Panne, 2008) reduce transcription (Kim and Maniatis, 1997; Thanos and Maniatis, 1995). These altered enhanceosome orientations hinder TF recruitment of TFIIB and coactivator CREB protein binding protein (CBP), which binds RNA Pol II (Kim et al., 1998). While CREB protein similarly drives transcription through interactions with CBP (Zhang et al., 2005), it is unclear how such interactions across 2 CREB proteins in local proximity would drive the observed expression periodicities. Additionally, it is surprising that the more distal CRE drove expression periodicity in variants with 2 CREs when CRE promoter proximity drives overall expression. Along with other bZip proteins, CREB protein indicates a positional binding preference around nucleosomes (Zhu et al., 2018) such that binding events are more often observed when their motifs are in certain orientations proximal to nucleosomes. While preferred orientations of the most distal CREB protein with respect to an upstream nucleosome may explain the expression periodicity observed here, further characterizations of TF-nucleosome interactions are necessary to prove this effect. This phenomenon is even more uncertain in an episomal context, where we observed large changes to periodicity phasing, due to altered histone ratios (Hebbar and Archer, 2008) and placement of nucleosomes (Jeong and Stein, 1994) on transient plasmid DNA constructs in comparison to their genomic counterparts. Since we did not observe such changes in the genomic context, further work is needed to resolve the differences TFBS positioning plays in driving expression between episomal and chromatin contexts.

With combinations of TFBSs within a cis-regulatory element the effect of *spacing* between sites plays a role on both the recruitment of TFs, via TF-induced DNA bending and TF-DNA interactions, in addition to aligning TF interactions with other regulatory partners. While CREB protein-induced DNA bending is minimal, estimated at 10° from its crystal structure (Schumacher et al., 2000), this nevertheless may play a role on CREB protein binding of two CREs in close proximity. Even minor protein-induced DNA bending indicates changes to major groove widths, and this change plays a role in TF-binding affinities, resulting in a similar ~10 bp periodicity in *in vitro* studies (Kim et al., 2013). Indeed, these experiments indicated a preference for protein binding on similar faces of the helix as opposed to opposite faces between proteins that similarly modify major groove width. Although it is difficult with our assay format to separate the effect of CRE *spacing* on TF-binding from TF transcriptional activity, such observations may explain the dampening in expression periodicity observed with 2 CREB proteins on opposite faces of the helix.

The number of consensus CREs largely determined expression in our assays, following similar trends as other homotypic clusters of TFBSs assayed (van Dijk et al., 2017; Gertz et al., 2009; Sharon et al., 2012; Weingarten-Gabbay et al., 2019). In contrast, a recently published MPRA in human cell lines found increasing number of CREB binding sites did not increase expression, although these sites were assayed in the absence of forskolin (Weingarten-Gabbay et al., 2019). Although the number of sites generally increase expression in both MPRAs tested here, there are instances in which variants drive less expression following increasing consensus

CRE number. In the genomic MPRA for example, many variants of a particular background drive higher expression with 4 consensus CREs than that of variants with 5 or 6 (Figure 2.4 A and Figure 2.13). In some of these examples, higher expression is observed with the addition of a weak CRE to a variant as opposed to a consensus CRE. Closely-spaced TFBSs can restrict the diffusion of TFs along DNA if one is already bound (Hammar et al., 2012), leading to binding competition between binding sites. This competition has been implicated in decreasing expression in a similar TFBS MPRA (van Dijk et al., 2017) and may explain our observations in the genomic MPRA. This effect is not as apparent in the episomal MPRA, a feature that may be explained by predominantly measuring competition between plasmids over competition between CREs in a single variant.

Lastly, we provide further evidence MPRA design and regulatory context must be considered in characterizations of regulatory features shaping expression. In both MPRAs, the contribution of background to variant expression is more predictive of variant activity than the presence of CRE at many positions along these backgrounds. The surrounding sequence content may play a similarly significant role for many other TFBSs, especially those that exhibit a bias in binding events based on the GC content similarity of the surrounding sequence to that of the TFBS itself (Dror et al., 2015). Therefore, we recommend incorporating multiple sequences as backgrounds in similar synthetic regulatory element designs especially since the use of a single background, as has been employed in many MPRAs, can influence TFBS trends observed (Figure 2.13). Additionally, we indicate here the ability of episomal assays to approximate genomic regulatory rules, yet also warn of the potential pitfalls of transient assays. Overall, we observe strong correlation between our episomal and genomic MPRAs (Figure 2.14, Pearson's r = 0.91). Yet the activity of variants with weaker-affinity CREs varies considerably between assays, which may be explained by differences in variant, and hence CRE, copy numbers in a cell. Thus we would not expect this effect to skew expression measurements of libraries assaying a high diversity of TFBSs. Alternatively, this could also be attributed to more consistent genomic structure in chromatin.

While we characterize various regulatory rules shaping the activity of a single TFBS, the CRE, we use this as a model to estimate a small fraction of the complexity of expression attained by combinations of TFBSs in natural cis-regulatory elements. Exploring how these rules scale with other TFBSs is integral to our understanding of cis-regulatory logic. Similar high-throughput approaches can build from the constraints explored here to develop more complex dissections of TFBS architectures. Transcriptional activation is thought to occur via phase-separated TF-coactivator-Pol II hubs, with local concentrations of these factors driving expression non-specifically (Boehning et al., 2018; Chong et al., 2018; Reiter et al., 2017). The interplay between transcriptional activity in these phase-separated systems and TFBS grammars needs further exploration. Further characterizations using similar synthetic systems will further our comprehension of cis-regulatory elements and our ability to confidently compose new ones with predictable activities.

Acknowledgements:

We thank the Kosuri Lab for providing useful feedback on earlier versions of this manuscript. We thank the UCLA Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research Sequencing Cores and the Technology Center for Genomics and Bioinformatics for providing next-generation sequencing, as well as the UCLA Eli & Edythe Broad Center of Regenerative Medicine Flow Cytometry Core for assistance in flow cytometry. We also thank Laura Day from the Kruglyak lab at UCLA for use of and assistance on their MiSeq. We also thank Dr. Michael R. Sawaya for his guidance in CREB-CRE structural modeling.

Funding: The National Institutes of Health (DP2GM114829 to S.K.), National Science Foundation Brain Initiative (1556207 to S.K), Searle Scholars Program (to S.K.), U.S. Department of Energy (DE-FC02-02ER63421 to S.K.), Ruth L.Kirschstein National Research Service Award (GM007185 to J.E.D.), the USPHS National Research Service Award (5T32GM008496 to E.M.J.), UCLA, and L. Wudl and F. Wudl.

Data and materials availability: All custom scripts and code for figure generation are available at the Kosuri Lab Github and can be accessed at the following link: https://github.com/KosuriLab/CRE_library_code. Raw data is available, and will be made available after peer-review. Plasmids and cell lines are available upon request.

Author Contributions:

J.E.D. and S.K. conceptualized the experiments. K.D.I. computationally designed the oligo library. J.E.D. performed all experiments and most computational analysis with the help of K.D.I. E.M.J. helped with the generation and validation of the H11 landing pad integration vector and cell line generation and validation. Q.H. helped with library generation for Figure 2.8. S.K. and J.E.D. wrote the manuscript with input from all authors.

Declaration of Interests:

S.K. consults for and holds equity in Octant Inc. where ongoing related work continues to be conducted, though no materials nor intellectual property related to this work are being used.

Materials and Methods

CRE regulatory library design

The CRE regulatory library was designed using three 150 bp *backgrounds* as templates and either a consensus CRE, taken from the CREB1 sequence logo in the JASPAR database, or weaker-affinity CRE, in which one of the central dinucleotides important for binding of both CREB protein monomers was mutated (Mayr and Montminy, 2001; Melnikov et al., 2012). Two of the backgrounds were adapted from previous MPRAs (background 55 (Melnikov et al., 2012) and background 41 (Smith et al., 2013)) and a third was isolated from a human genomic region indicating minimal activity in the developing eye in the VISTA enhancer database (Visel et al., 2007) (background 52). Both background 41 and 52 were obtained from the human genome, with 41 corresponding to Chr9: 81,097,684-81,097,833 and 52 to Chr5: 89,377,854-89,378,003 from GRCh38. Background 55 corresponds to the CRE response element of a commercial reporter plasmid (Fan and Wood, 2007) with a portion duplicated to reach 150 nt and with all CREs scrambled, maintaining their GC content. Variants were generated by replacing background sequence with CREs along with a constant 2 bp flanking nucleotides to ameliorate local sequence effects due to CRE placement in the backgrounds (Levo et al., 2015). A MluI restriction enzyme site (ACGCGT) was placed upstream of each variant and KpnI restriction enzyme site (GGTACC) was placed downstream, each for library cloning. Lastly, a pair of 19 nt amplification primers (Eroshenko et al., 2012) specific to each of the 2 library designs were added to each design producing 200 nucleotide libraries of the format: (5' -> 3') subpool primer 1 - MluI - variant - KpnI - subpool primer 2. The resulting 4185 variants corresponding to the 2 designs in Figure 2.1, 417 variants corresponding to the single-CRE library, and one CRE positive control (Fan and Wood, 2007) were all synthesized on Agilent Microarrays.

Library cloning

OLS libraries (Agilent Technologies, Santa Clara, CA) were resuspended to a final volume of 200 nM in TE pH 8.0 (Sigma-Aldrich, Saint Louis, MO). Libraries were amplified using 1 µL of a 10-fold dilution of the library, the respective subpool primer pairs (Subpool_#_F and Subpool_#_R with # representing the sub-libraries present, 2 refers to the single CRE Distance library, 3 refers to the CRE Spacing and Distance library, and 5 refers to the CRE Number and Affinity library) and with KAPA HiFi HotStart Real-time PCR Master Mix (2X) (Kapa Biosystems, Wilmington, MA) following the recommended cycling protocol at 14 cycles. Random barcodes were added to variants in a second PCR in which the primer downstream of variants contained 20 nucleotides of random sequence, synthesized with the machine-mixed setting (Integrated DNA Technologies, Coralville, IA). One ng product was used in this barcoding qPCR using biotinylated primers (SP#_Biotin and SP#_Biotin_BC_R with subpool numbers corresponding to library designs as before). The qPCR was performed with KAPA HiFi HotStart Real-time PCR Master Mix (2X) (Kapa Biosystems) for 11 cycles following the recommended cycling protocol. Barcoded libraries were digested with MluI-HF (New England Biolabs, Ipswich, MA) and SpeI-HF (New England Biolabs) in 1X cut-smart buffer (New England Biolabs). The biotinylated primers and undigested library members were removed using Dynabeads M-270 Streptavidin (Thermo Fisher Scientific, Waltham, MA), using the recommended “Immobilize nucleic acids” protocol, collecting the supernatant after adding the library mixture to the beads.

The barcoded and digested library members were cloned into the integration vector (pJDrcEPP) that had previously been digested using MluI-HF and SpeI-HF in 1X cut-smart buffer (New England Biolabs). Ligation was performed at a 1:3 ratio of pJDrcEPP:library using T4 DNA ligase (New England Biolabs). Ligation product was cleaned-up using a Clean and Concentrator Kit (Zymo Research, Irvine, CA) followed by drop-dialysis for 15 minutes with UltraPure™ DNase/RNase-Free Distilled Water (Thermo Fisher Scientific) before transforming 1 µL into NEB 5-alpha Electrocompetent E. coli (New England Biolabs) following the recommended protocol. Dilutions of transformants were plated on 50 µg/mL Kanamycin (VWR, Radnor, PA) LB plates at 10-fold dilutions to 1/10,000 and grown overnight at 37°C and the remainder of the cells were left in SOC (New England Biolabs) at 4°C overnight. The next day, dilutions were counted, estimating 695,000 and 950,000 original transformants for the placement and spacing library and the number and affinity library, respectively. The remainder transformants kept overnight at 4°C were pelleted and placed in fresh LB for 3 hours at 30°C, then diluted into 100x volume LB + 50 µg/mL Kanamycin (VWR) and grown at 30°C for 18 hours before isolating library vectors using QIAprep (Qiagen, Hilden, Germany) spin miniprep kits.

Library vectors (pJDrcEPP_lib) were then digested in 2 steps, in order to isolate plasmids correctly cut at the synthesized KpnI recognition site. Vectors were first digested with KpnI-HF along with rSAP (New England Biolabs) in 1X CutSmart buffer. Products were run on a 0.8% TAE agarose gel and linearized plasmids were isolated with Zymoclean Gel DNA Recovery Kit (Zymo Research). Vectors were then digested in a similar fashion with XbaI (New England Biolabs) without gel isolation. The minimal promoter and luciferase insert was prepared using biotinylated PCR primers (Amp_minPLuc2_Biotin_For and Amp_minPLuc2_Biotin_Rev) corresponding to pMPRAdonor2 (Addgene plasmid #49353) and Kapa HiFi HotStart ReadyMix (Kapa Biosystems). The insert was digested with both KpnI-HF and XbaI (New England Biolabs). Biotinylated primers and undigested inserts were removed as before using Dynabeads M-270 Streptavidin (Thermo Fisher Scientific). Ligation of pJDrcEPP_lib and minP-Luc2 inserts was performed at a 1:3 ratio as before but with T7 DNA ligase (New England Biolabs). Ligation product was cleaned-up and transformed as before, plating similar dilutions but instead growing the remainder transformants overnight in 100x LB + Kanamycin (50 µg/mL) (VWR) at 30°C. The next day, dilutions were counted, estimating 15,877,000 and 17,845,000 original transformants for the placement and spacing library and the number and affinity library, respectively. Library vectors with insert (pJDrcEPP_lib_minPLuc2) were isolated from the remainder transformants using Qiagen Plasmid Plus Maxi Kit (Qiagen).

Barcode mapping

Barcodes were associated with each library member by sequencing amplicons isolated from pJDrcEPP_lib. 0.5 ng of plasmid was amplified using primers with P5 and P7 Illumina flow cell adapter sequences (Libseq_P7_For and Libseq_P5_Rev) and KAPA HiFi HotStart Real-time PCR Master Mix (2X) (Kapa Biosystems) for 17 cycles. Amplicons were isolated on a 2% TAE agarose gel and bands were confirmed using Agilent's D1000 ScreenTape and reagents (Agilent Technologies) on a 2200 Tapestation system. Libraries were sequenced on an Illumina MiSeq with a v3 600-cycle reagent kit (Illumina, San Diego, CA) using custom read 1 primer LibSeq_R1Seq_Rev and custom read 2 primer LibSeq_R2Seq_For loaded into the cartridges read 1 and read 2 primer wells, respectively. 35,355,712 reads passed filter and 31,486,576 reads were merged with BBMerge version 9.00. A custom python script was used to map unique barcodes to variants lacking synthesis errors. Briefly, this script searched the last 150 bp of merged reads for sequences perfectly matching the variants designed. The first 20 bp of each merged read was determined to be a barcode and each barcode was then mapped to the most common sequence associated with it, only retaining barcodes that appeared more than twice in merged reads. In order to differentiate between mapped barcodes that are associated with variants with sequencing errors and another variant in the library, we used a Levenshtein distance cut-off of 13 between variants that share a common barcode. This cut-off represented 1% of the total bootstrapped distances between perfect variants in the library. Barcodes mapped to perfect variants were kept if all other variants associated with a barcode fell below this cut-off, retaining 724668 barcodes.

Genomic MPRA Library integration

2.6 x 10^6 Hek293T H11 landing pad cells were plated per T75 flask, 6 flasks in total, and grown in DMEM with 1% Penicillin-streptomycin and 10% FBS (Thermo Fisher Scientific); this is the cell media used in all tissue culture work unless otherwise stated. The next day, the cells were transfected with a total of 187.5 µL Lipofectamine 3000 (Thermo Fisher Scientific), 6.252 mL Opti-MEM (Thermo Fisher Scientific), 13.86 µg BxB1 expression vector (Duportet et al., 2014), 153.36 µg spacing and distance library vector, 180 µg number and affinity library vector and 250 µL P3000 (Thermo Fisher Scientific) following the recommended protocol. BxB1 was added to the DNA mixture at a 1:8 ratio, while both libraries were added at 3x to increase efficiency. Cells were passaged after 3 days onto a T875, in which the media was changed to 1 µg/mL puromycin (Life Technologies, Carlsbad, CA) selection media. Unless otherwise stated, cells were passaged in all tissue culture work according to: trypsinization with Trypsin-EDTA

0.25% (Thermo Fisher Scientific) followed by inactivation with 2x volume of cell media, pelleting at 1000 x g for 5 minutes and resuspension in fresh media. 1/160 of the cells were removed before selection and grown without puromycin to analyze overall integration efficiency. The selection cells were passaged at 1:10, 1:20 or at 1:1 as needed during selection every 1-4 days, with 6 passages in total over 16 days of selection. Cells plated for integration efficiency analysis were passaged at 1:10 or 1:20 every 3 or 4 days for a total of 6 passages. Both cells were analyzed using flow cytometry 20 days after transfection, shown in Figure 2.6 B. Samples were prepared in PBS pH 7.4 (Thermo Fisher Scientific) using the LSRII at the UCLA Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research Flow Cytometry Core. Cytometer settings were adjusted to: FSC – 157 V, SSC – 233 V, Alexa Fluor 488 – 400 V. Selected cells were frozen at 5 x 10^6 cells/mL in 5% DMSO (Thermo Fisher Scientific) and aliquots were used in the genomic MPRA.

Luminescence assays

For the landing pad orientation luminescence assay (Figure 2.6 C), 22 x 10^3 cells containing integrated control sequences were plated in triplicate across a 96-well plate. 100x forskolin stocks were made via serial dilution in DMSO (Thermo Fisher Scientific) , and 1x forskolin solutions were made in CD 293 media (Thermo Fisher Scientific) supplemented with 4 mM L-Glutamine (Thermo Fisher Scientific). The next day, media was removed from all 96-wells and replaced with 25 µL of media with forskolin (0, 0.5, 1, 5, 10, 50, 100, and 120 µM). After 4 hours, fluorescence was measured using the Dual-Glo Luciferase Assay Kit (Promega, Madison, WI), in which 10 µL of Dual-Glo Luciferase Reagent was added, cells were shook for 10 minutes and luminescence was measured on a plate reader.

For the MPRA library luminescence assays (Figure 2.6 D), 880,000 H11 landing pad cells and the genomic MPRA cells were resuspended in 12 mL media and 100 µL was distributed per well across a 96-well plate. The next day, the H11 landing pad cells were transfected with a total of 6.6 µL Lipofectamine 3000, 220 µL Opti-MEM , 0.44 µg Renilla luciferase expression vector, 2.15 µg spacing and distance library vector, 2.79 µg number and affinity library vector and 8.8 µL P3000 (Thermo Fisher Scientific) following the recommended protocol, using 10 µL of this mixture per 96-well. 100x forskolin stocks were made via serial dilution in DMSO (Thermo Fisher Scientific), and 1x forskolin solutions were made in cell media. Media was removed from all 96-wells and replaced with 25 µL of media with forskolin (0, 1, 2, 4, 8, 16, and 25 µM). After 4 hours, fluorescence was measured using the Dual-Glo Luciferase Assay Kit (Promega), in which 10 µL of Dual-Glo Luciferase Reagent was added, cells were shook for 10 minutes and luminescence was measured on a plate reader. Renilla luminescence from the transfected cells was measured following Stop & Glo Reagent addition.

Episomal MPRA

Two mL of a 1.026 x 10^5 cells/mL stock of H11 Landing Pad cells were plated per 6-well per biological replicate. The next day, cells were transfected with a total of 64.5 µL Lipofectamine 3000, 4.3 mL Opti-MEM, 18.1 µg CRE Spacing and Distance library vector, 21.9 µg CRE Number and Affinity library vector and 86 µL P3000 (Thermo Fisher Scientific) following the recommended protocol, using 250 µL of this mixture per 6-well. Both library vectors were concentrated using a Wizard SV Gel and PCR Clean-up (Thermo Fisher Scientific) prior to transfection. 100x forskolin stocks were made via serial dilution in DMSO (Thermo Fisher Scientific), and 2x forskolin solutions were made in cell media. The next day, 2 mL of media with 2x forskolin was added to the 2 mL media within each 6-well (final concentrations: 0, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0} and 2^{2} µM forskolin). After 3 hours, RNA was collected using Qiagen RNeasy Mini Kits with Qiashredder and on-column DNase I digestions with Qiagen RNase-free DNase Set (Qiagen).

Per sample, RNA was reverse-transcribed using 1.5x the recommended materials for Superscript IV Reverse Transcriptase (Thermo Fisher Scientific) with 7.5 µg total RNA and the library-specific primer Creb_Hand_RT, which anneals downstream of barcoded transcripts. The recommended protocol was followed with changes including reverse transcription at 55°C for 1 hour and RNase H (Thermo Fisher Scientific) removal of RNA in RNA:DNA hybrids. To ensure the same amount of barcoded cDNA was used per PCR across forskolin concentrations and that this amount covered library complexity, a preliminary qPCR of total RNA samples was performed alongside serial dilutions of a known amount of barcoded cDNA previously amplified. Sample Cq's were referenced to Cq's of the serial dilutions to determine approximate concentrations of barcoded transcripts per total RNA loaded. Volumes of samples that approximated 6000-fold coverage of the number of variants (not including barcode complexity) were determined and used in the following PCR.

cDNA was amplified with NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) using input amounts determined from qPCR, distributed across 4 replicates each so as to not exceed 10% of the total PCR volume, and with primers specific to luciferase (Creb_Seq_Luc_R) and a 20 nt annealing site added during reverse-transcription (Creb_Hand). PCR conditions were followed as recommended with 61°C annealing for 20s, extension at 72°C for 20s and a final extension at 72°C for 2 minutes for a total of 10 cycles. Meanwhile, library plasmids mixed at the same ratio as for transfection were used for DNA normalization in the episomal MPRA. 128 ng of this mixture was used per PCR with NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) along with the reverse-transcriptase primer (Creb_Hand_RT) and the primer specific to luciferase (Creb_Seq_Luc_R) using the same PCR conditions for cDNA for 9 cycles. Amplicons were isolated on a 2% TAE gel, after which they were cleaned-up again.

A second PCR was performed on both the cDNA and plasmid DNA amplicons in order to add P5 and P7 Illumina flow cell adapter sequences and indices. NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) was used with 0.5 ng input and the primers P5_Seq_Luc_F and P7_Ind_#_Han or P7_In_####_Han, with # corresponding to the index code per sample indicated in Table 2.1. PCR conditions were followed as recommended with 63°C annealing for 20s, extension at 72°C for 25s and a final extension at 72°C for 2 minutes for a total of 7 cycles. Bands were confirmed using Agilent's D1000 ScreenTape and reagents (Agilent Technologies) on a 2200 Tapestation system. Samples were mixed equally and sequenced on a NextSeq500 at the UCLA Technology Center for Genomics & Bioinformatics (TCGB) using the 1 x 75 v2 kit (Illumina) with only 30 cycles. Before sequencing, the custom read 1 sequencing primer Creb_R1_Seq_P and indexing primer Creb_Ind_Seq_P were loaded into the read 1 and index primer positions in the NextSeq cartridge. 365.67 x 10^6 reads passing filter were obtained across the 8 dilutions (2 replicates each) and plasmid DNA sample with reads per index ranging between 12 x 10^6 and 20 x 10^6.

The episomal MPRA performed at concentrations beyond 1 μM forskolin indicated in Figure 2.7 was similarly transfected and prepped according to the above protocol but with incubations at 0, 2^0, 2^1, 2^2, 2^3, 2^4, 25, 2^5 and 2^6 μM forskolin. Samples were sequenced on the NextSeq500 using the 1 x 75 v2 kit (Illumina) with 75 cycles and 305.06 x 10^6 reads passing filter were obtained across the 9 dilutions (2 replicates each) and plasmid DNA sample with reads per index ranging between 10 x 10^6 and 14 x 10^6. The single CRE library indicated in Figure 2.8 was similarly transfected and prepped according to the above protocol but with incubations at 0 and 25 μM forskolin. Additionally, volumes of cDNA samples that approximated 1500-fold coverage of the number of variants (not including barcode complexity) were determined and used in the following PCR. 101.55 x 10^6 HiSeq reads passing filter were obtained across all cDNA and plasmid DNA samples.

Genomic MPRA

2.5 x 10^5 Genomic MPRA library-integrated and selected cells were plated on 2 separate 6-wells, forming the two biological replicates used in the Genomic MPRA. These cells were passaged twice in their expansion, after which cells were frozen at 5 x 10^6 cells/mL in 5% DMSO (Thermo Fisher Scientific). For the Genomic MPRA, 5 aliquots of each replicate, 2.5 x 10^7 cells total, were thawed and grown to cover the initial bottleneck amount 100-fold. 2 days later these cells were trypsinized and plated for stimulation at 3.47 x 10^6 cells per 150 cm plate with 20 mL of media, eight plates total per replicate. Two days later, both replicates were stimulated by adding 20 mL of media with 16 μM forskolin to the 20 mL media already on plates. After 3 hours, replicates were trypsinized, combined, spun down at 1000 x g for 5 minutes, resuspended in media and split evenly into 2 tubes, one RNA extraction and one for genomic DNA extraction.

Cells aliquoted for RNA processing were spun down and resuspended in 3.2 mL of RLT (1% ß-Mercaptoethanol) from a Qiagen RNeasy Midi kit (Qiagen). Cells were homogenized by passing the lysate through an 18-gauge needle 10 times and were stored at -80°C. Two days later, lysates were thawed and processed according to the RNeasy Midi Protocol for Isolation of Total RNA from Animal Cells from Qiagen RNeasy Midi/Maxi Handbook (09/2010) (Qiagen) starting at the addition of 1x volume of 70% ethanol to thawed lysates. On-column DNase I digestions were performed with the Qiagen RNase-free DNase Set (Qiagen). RNA was eluted with 200 µL RNAse-free water (Qiagen) and subsequently concentrated using an Amicon Ultra-0.5 mL Centrifugal Filter with a 10 kDa cut-off. Total RNA was stored at -20°C.

Cells aliquoted for genomic DNA were spun down and resuspended in PBS pH 7.4 (Thermo Fisher Scientific) twice to give a final concentration of 1 x 10^7 cells/mL. Samples were processed according to the Sample Preparation and Lysis Protocol for Cell Cultures from QIAGEN Genomic DNA Handbook (08/2001) using the settings for the Qiagen Blood and Cell Culture DNA Maxi Kit (Qiagen). Pelleted nuclei were frozen at -20°C before G2 buffer was added. Two days later, nuclei were thawed and the remainder of the protocol was followed using Qiagen Protease digestion at 50°C for 60 minutes, and precipitating DNA according to the recommended protocol for vortexing and centrifugation after isopropanol addition followed by washing with cold 70% ethanol. Genomic DNA was resuspended in 800 µL Qiagen Elution Buffer (Qiagen) and left at room temperature overnight. The next day, gDNA was shook at 600 rpm for 3 hours at 55°C. RNase A (DNase and protease-free, Thermo Fisher Scientific) was added to a final concentration of 99 ng/µL. Over the next 3 days, 600 µL Qiagen Elution Buffer (Qiagen) was added incrementally, with additional shaking at 55-60°C after each addition for a total of 28 hours; this was largely due to resuspension issus. Resuspended DNA was stored at 4°C.

Per replicate, 130 µg total RNA was reverse transcribed using the recommended materials for Superscript IV Reverse Transcriptase (Thermo Fisher Scientific) but with 10 µg total RNA instead of the recommended 5 µg per 20 µL reaction. The library-specific primer Creb_RT_Hand_3, which anneals downstream of barcoded transcripts, was added to reactions and the recommended protocol was followed with changes including reverse transcription at 55°C for 1 hour and RNase H (Thermo Fisher Scientific) removal of RNA in RNA:DNA hybrids. RNAse A (DNase and protease-free, Thermo Fisher Scientific) was added to each reaction at a final concentration of 100 ng/µL and incubated at 37°C for 30 minutes. Reactions were combined and concentrated using an Amicon Ultra-0.5 mL Centrifugal Filter with a 10 kDa cut-off (Sigma-Aldrich). To ensure the same amount of barcoded cDNA was used per PCR across replicates and that this amount covered library complexity, a preliminary qPCR of total RNA samples was performed alongside serial dilutions of a known amount of barcoded cDNA previously amplified. Sample Cq's were referenced to Cq's of the serial dilutions to determine approximate concentrations of barcoded transcripts per total RNA loaded per replicate. Of the 30 µL volume of cDNA remaining in each replicate, total barcoded molecules were estimated. 30 µL of the replicate with the lower concentration of barcoded molecules was used in the following PCR while a portion of the replicate with the higher concentration was used to approximate similar cDNA input into the following PCR. Both amounts loaded covered the original 2.5 x 10^5 cell bottleneck amount 24.5-fold.

cDNA was amplified with NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) using volumes determined from qPCR distributed across 5 replicates each, so as not to exceed 10% of the total PCR volume, and with primers specific to luciferase (Creb_Luc_Seq_R) and a 20 nt annealing site added during reverse-transcription (Creb_Hand). PCR conditions were followed as recommended with 61°C annealing for 20s, extension at 72°C for 20s and a final extension at 72°C for 2 minutes for a total of 16 cycles. A second PCR was performed in order to add P5 and P7 illumina flow cell adapter sequences and indices. NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) was used with 0.5 ng cDNA and the primers P5_Seq_Luc_F and P7_Ind_##_Han, with ## corresponding to the index code per sample indicated in Table 2.1. PCR conditions were followed as recommended with 63°C annealing for 20s, extension at 72°C for 25s and a final extension at 72°C for 2 minutes for a total of 7 cycles.

Meanwhile, gDNA was aliquoted into 2 tubes evenly before PCR to establish 2 technical replicates per biological replicate. Per technical replicate, gDNA was amplified with NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) with a biotinylated reverse-transcriptase primer Creb_Hand_RT_3 and a biotinylated primer specific to luciferase (Creb_Luc_Seq_R). 5 µg gDNA was loaded per 50 µL in a 96-well PCR plate, with 57 total reactions per technical replicate for one biological replicate and only 51 for the other, due to sample loss. PCR conditions were followed as recommended with 61°C annealing for 20s, extension at 72°C for 20s and a final extension at 72°C for 2 minutes. After 7 cycles, wells were combined and cleaned-up using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA). 0.4x volume of beads were used per sample and after magnetic separation, the supernatant was collected, isolating the amplicons from genomic DNA. Similar as in (Matreyek et al., 2017), 40% of eluted volume was used in a second PCR to add P5 and P7 illumina flow cell adapter sequences and indices. This volume was distributed across 14 replicates per technical replicate so as to not exceed 10% of the total PCR volume and amplified using NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) with the primers P5_Seq_Luc_F and P7_Ind_##_Han, with ## corresponding to the index code per sample. PCR conditions were followed as recommended with 63°C annealing for 20s, extension at 72°C for 25s and a final extension at 72°C for 2 minutes for a total of 14 cycles.

Both the cDNA and gDNA amplicons were isolated on a 2% TAE gel and bands were confirmed using Agilent's D1000 ScreenTape and reagents (Agilent) on a 2200 Tapestation system. Samples were mixed equally and sequenced on a HiSeq2500 at UCLA's Broad Stem Cell Research Center using a 1 x 50 kit. Before sequencing, the custom read 1 sequencing primer Creb_R1_Seq_P and indexing primer Creb_Ind_Seq_P were loaded into the read 1 and index primer positions in the HiSeq cartridge. Indexed samples were de-multiplexed in-house using a custom python script that matched indices with those submitted and all 1 bp mutations, accounting for sequencing errors. 143,759,096 reads passing filter were obtained across the 2 biological replicates, each consisting of 2 genomic DNA technical replicates and 1 RNA sample. Reads per index ranged between 9 x 10^6 - 22 x 10^6.

Processing MPRA Sequencing Data

A custom python script was used to isolate barcode sequences from reads and determine their total number of reads. Briefly, the first 20 sequences were extracted from each read, reverse complemented to match the barcode format from barcode mapping and then the occurence of each of these barcodes was summed as their total number of reads per indexed sample. RStudio (R version 3.5.3 and the packages: tidyverse 1.2.1, lemon 0.4.3, viridisLite 0.3.0, cowplot 0.9.4, caTools 1.17.1.2, broom, 0.5.1, and modelr 0.1.4) was used for the remainder of data processing. Barcodes were normalized to sequencing depth per sample and represented as normalized reads per million. Barcodes were retained and used in variant expression determination only if they also appeared in the barcode-variant mapping table. Barcodes that were present in the barcode-variant mapping table that were not present in a sample were given the value of 0 normalized reads per million amongst retained barcodes.

Determining MPRA Variant Expression

In the episomal MPRA, barcodes were retained across all samples with > 6 reads in the DNA sample. Barcode expression was determined by dividing normalized barcode reads per million in RNA samples to their normalized reads in the plasmid DNA sample. Variants were retained across all samples if they had > 7 barcodes retained in the plasmid DNA sample. Median expression per variant was determined by taking the median expression of all barcodes associated with a single variant, maintaining variants with > 0 expression in all samples. In total, 4162 of the original 4185 variants designed along with the CRE control were retained after processing. Median variant expression in the single CRE library indicated in Figure 2.8 was similarly determined except retaining barcodes across all samples with > 5 reads in the DNA sample.

In the genomic MPRA, expression was calculated similarly as for the episomal MPRA, with changes accounting for 2 DNA technical replicates. Barcodes were retained across all samples per biological replicate if they had > 6 reads in both DNA technical replicates. Barcode expression was determined by dividing normalized barcode reads per million in RNA samples to their average normalized reads across the genomic DNA samples. Per biological replicate, variants were similarly retained in the RNA sample if they were associated with > 7 barcodes retained in the combined DNA sample. Median expression per variant was determined by taking the median expression of all barcodes associated with a single variant. Variants were retained for further analysis if they had >0 expression in both biological replicates. Overall, 3479 of the original 4185 variants designed were retained in analysis in addition to the CRE control. Variants not retained consisted of: 128 from the number and affinity library and 578 from the spacing and distance library (488 of this was with background 41, of which was dropped from analysis). For both MPRAs, the average expression between biological replicates was used for all variant analyses.

CREB::ßzip structure superpositions along DNA

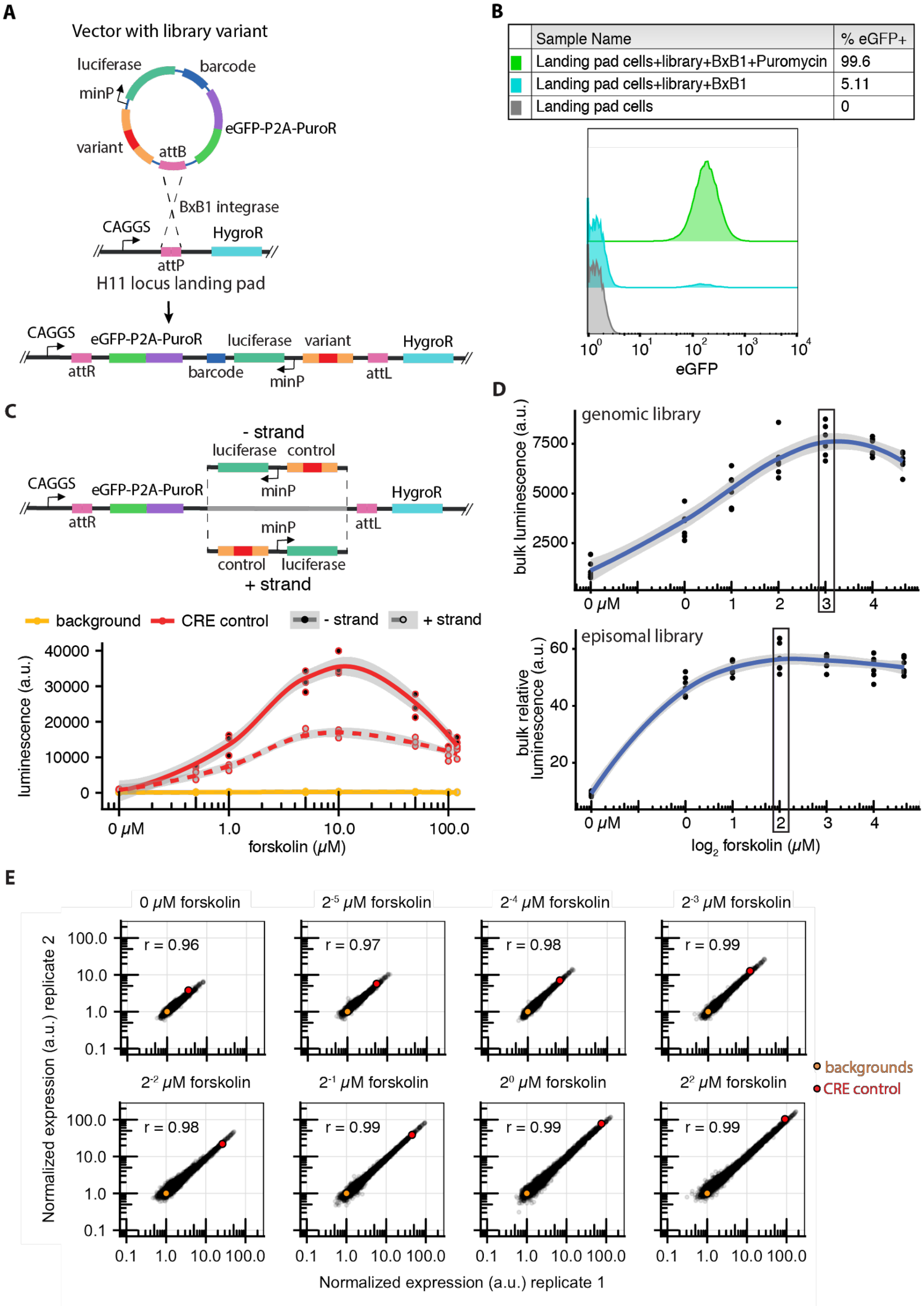
The structure coordinates of a CREB::Bzip dimer bound to the somatostatin CRE (Schumacher et al., 2000) was downloaded from PDB (code: 1DH3) and loaded into Coot (version 0.8.9.2). Models of CREB protein spacing were made by using the LSQ Superpose function, superpositioning a copy of the protein:DNA complex onto the original structure using least squares fit to the mainchain of Chain B, corresponding to one DNA strand. Briefly, a model of 5 bp CREB protein spacing was established by taking residues -10:-4 on Chain B of the reference structure and moving them to residues 4:10 on Chain B. This matched 63 atoms with a rms deviation of 1.22. A model of 10 bp CREB protein spacing was established by taking residues -11:-9 on Chain B of the reference structure and moving them to residues 8:10 on Chain B. This matched 26 atoms with an rms deviation of 1.49.

A model of six CREB proteins bound to six CREs in the Number and Affinity library was constructed similarly. CREB protein bound to the first CRE was created by taking residues 5:9 on Chain B of the reference structure and moving them to residues -9:5 on Chain B. CREB protein bound to the second CRE was created by taking residues -9:5 on Chain B of the reference structure and moving them to residues 4:8 on Chain B. The third instance of CREB protein-CRE was created by taking residues -9:-5 on Chain B of the reference structure of the second CREB protein-CRE and moving them to residues 5:9 on Chain B then taking residues -9:5 on Chain B of this new reference structure and moving them to residues 4:8 on Chain B. This was repeated sequentially using the previous CREB protein-CRE structure as a reference until superpositioning the sixth instance of CREB protein bound to a CRE. A total of 10 LSQ superpositions were performed matching 45 atoms each time with a rms deviation of either 1.15 or 1.09 depending upon the reference and moving residues.

Log-linear expression modelling

A model was fit using lm()in R *stats* package to predict average expression from the independent contribution of background and the 6 CRE positions in the site number and affinity library (expression ~ background + site1 + site2 + site3 + site4 + site5 + site6). Background and each position was represented as categorical variables according to the 3 backgrounds used and 3 possible affinities per position (consensus CRE, weak CRE, and no CRE). The percent of variance explained per model term was obtained using the sum of squares from anova().

Supplemental Materials



	Sample Name	% eGFP+
🟩	Landing pad cells+library+BxB1+Puromycin	99.6
🟦	Landing pad cells+library+BxB1	5.11
⬛	Landing pad cells	0

Figure 2.6. Establishment of the episomal and genomic MPRA. (A) Library-containing vectors are genomically-integrated by co-transfecting with a BxB1 expression vector into a HEK293T cell line containing a single-copy landing pad at the H11 locus. BxB1-mediated integration occurs through the genomic recombination site, attP, and the vector recombination site, attB. Successful integration at H11 switches cell antibiotic resistance from hygromycin to puromycin, via the genomic CAGGS promoter, in addition to driving expression of eGFP. Vector and landing pad components not shown to scale. (B) Integration of library-containing vectors in the genomic MPRA was monitored using eGFP activation upon genomic integration. 5.11% of transfected cells expressed eGFP from an integrated construct (cyan). Successful integrants were isolated after outgrowth in media containing puromycin (green). (C) Integration orientation of library controls at the landing pad resulted in different levels of induced expression. The negative strand orientation placed the attL sequence immediately upstream of the CRE control while in the positive strand orientation, this was replaced by bacterial backbone.The negative strand orientation was chosen for the genomic CRE MPRA. Lines indicate a loess fit with shaded regions indicated standard error. (D, top graph) Bulk genomically-integrated library luciferase expression measured across forskolin dilutions, 6 technical replicates each. The genomic MPRA was performed using 8 μM forskolin for comparisons to the episomal MPRA. (D, bottom graph) Similar bulk luciferase expression measurements but after transfection of the episomal library. Luciferase luminescence normalized to luminescence from a Renilla transfection control. The episomal MPRA analysis was performed at 4 μM forskolin for comparisons to the genomic MPRA. In both graphs, lines indicate a loess fit with shaded regions indicated standard error. (E) Replicability plots following the episomal MPRA titration curve in figure 2.1 D. Expression from each variant was normalized to the expression of their corresponding background per biological replicate to visualize induction following increasing forskolin. Replicability ranged from r = 0.96 to r = 0.99 across concentrations.

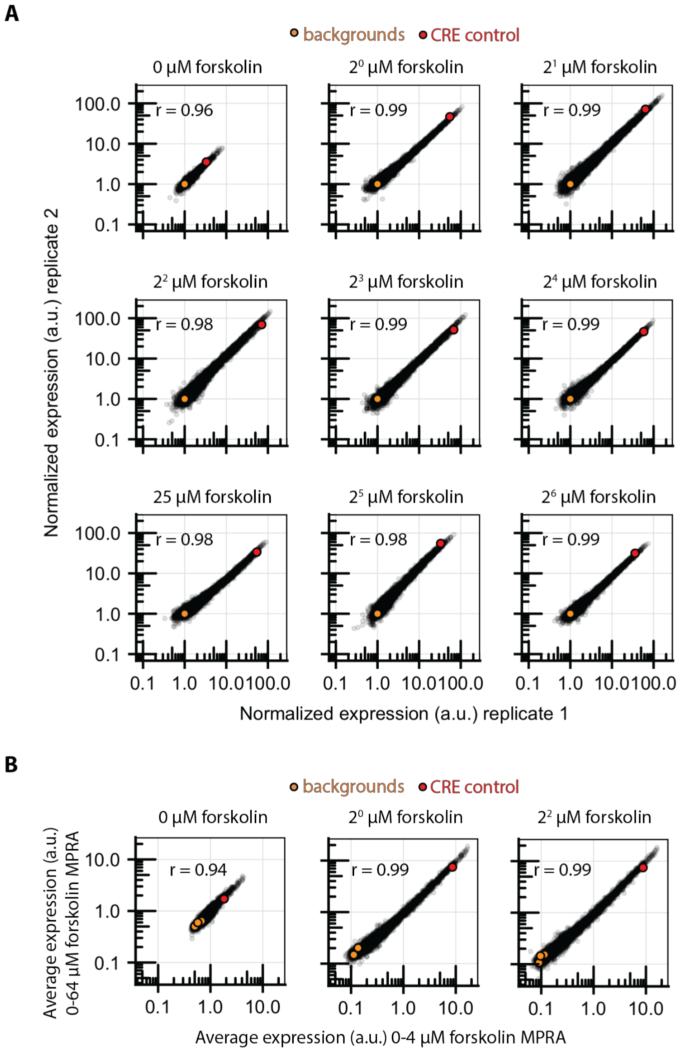


Figure 2.7 Episomal MPRA tested at concentrations beyond maximal induction indicate little change in expression range. (A) The episomal MPRA was performed at concentrations spanning beyond those used in the main analysis (Figure 2.1 D and Figure 2.6 E) to confirm the full induction range in episomal conditions. (B) Repeated concentrations between episomal MPRAs indicate high reproducibility (r = 0.94-0.99).

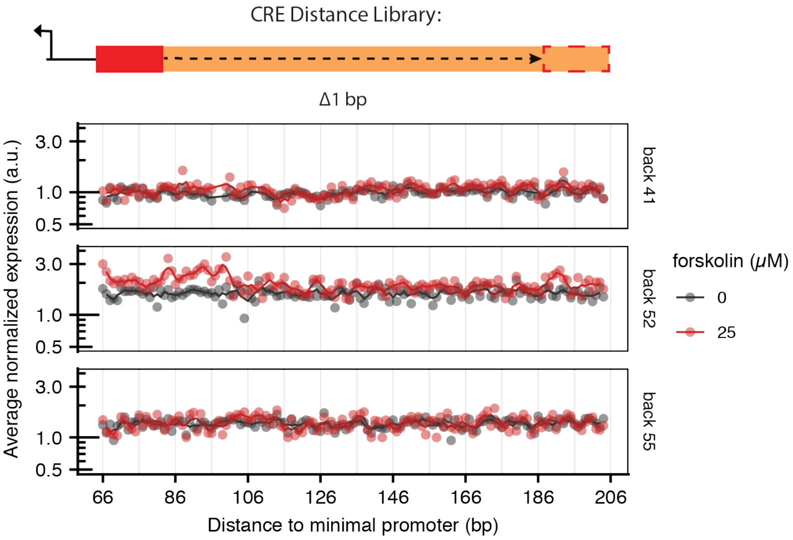


Figure 2.8 Variants with one CRE drive minimally-induced expression and do not indicate large expression variation as CRE distance to the promoter is varied in the episomal MPRA. Variants containing a single CRE at every position along the three backgrounds assayed in an separate MPRA at uninduced and fully-induced forskolin concentrations. Variant expression is normalized to the expression of their backgrounds, averaged across replicates and compared between forskolin concentrations. Line shown is the 3 bp moving average expression estimate. As the distance of one CRE to the promoter varies there is little expression variation in all but a portion of distances in variants with background 52.

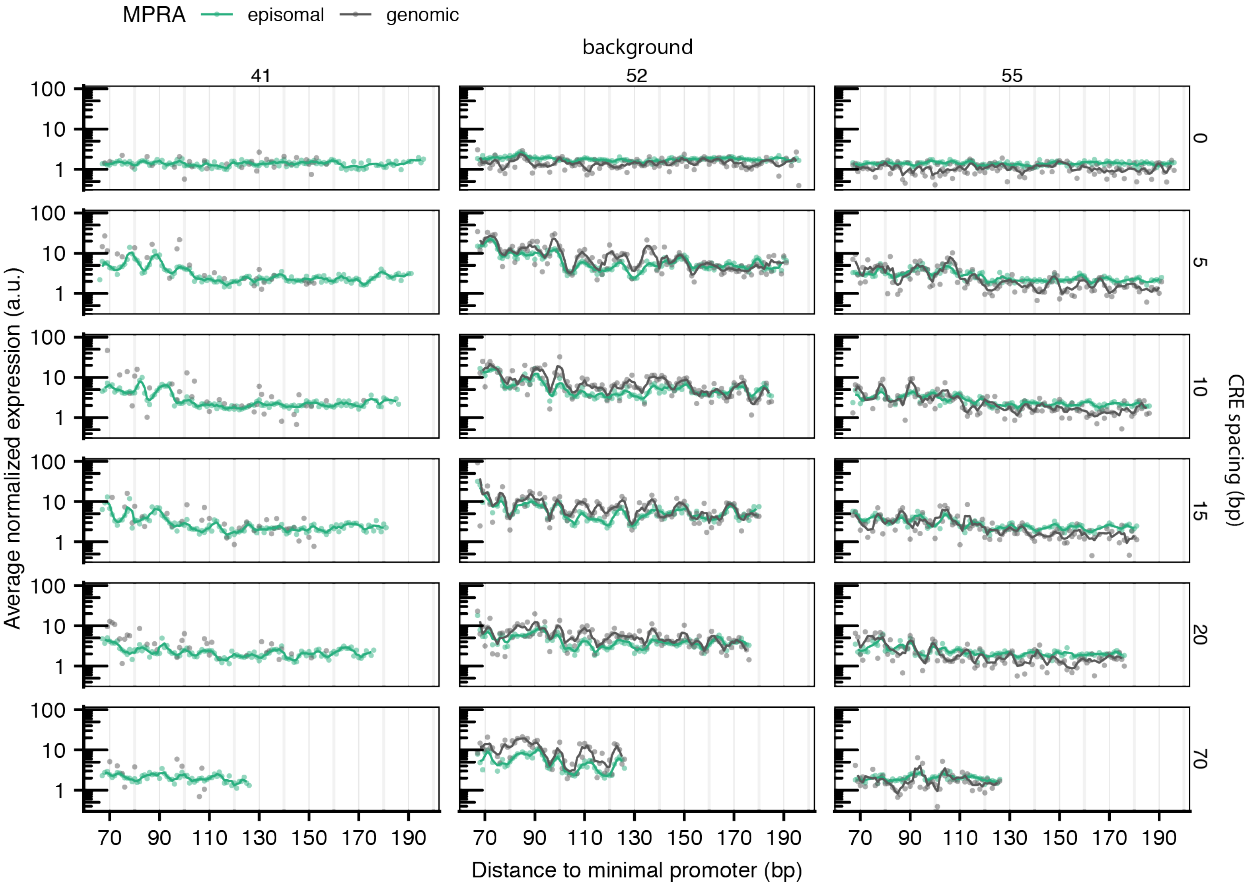


Figure 2.9. Expression measurements for all variants in the CRE Spacing and Distance library retained in analysis. All MPRA expression measurements used for analysis of variants in the Spacing and Distance library. Comparisons to the episomal MPRA were performed with expression obtained at 4 μM forskolin. Data quality filters in the genomic MPRA remove 74% of variants with background 41, thus only episomal MPRA expression is used for periodicity analysis in variants with this background in Figure 3. Similar normalized expression profiles are observed between variants in the episomal and genomic context. Variants with 0 bp CRE *spacing* were predicted to occlude the binding of CREB protein to both CREs; here we observe minimal expression driven by these variants.

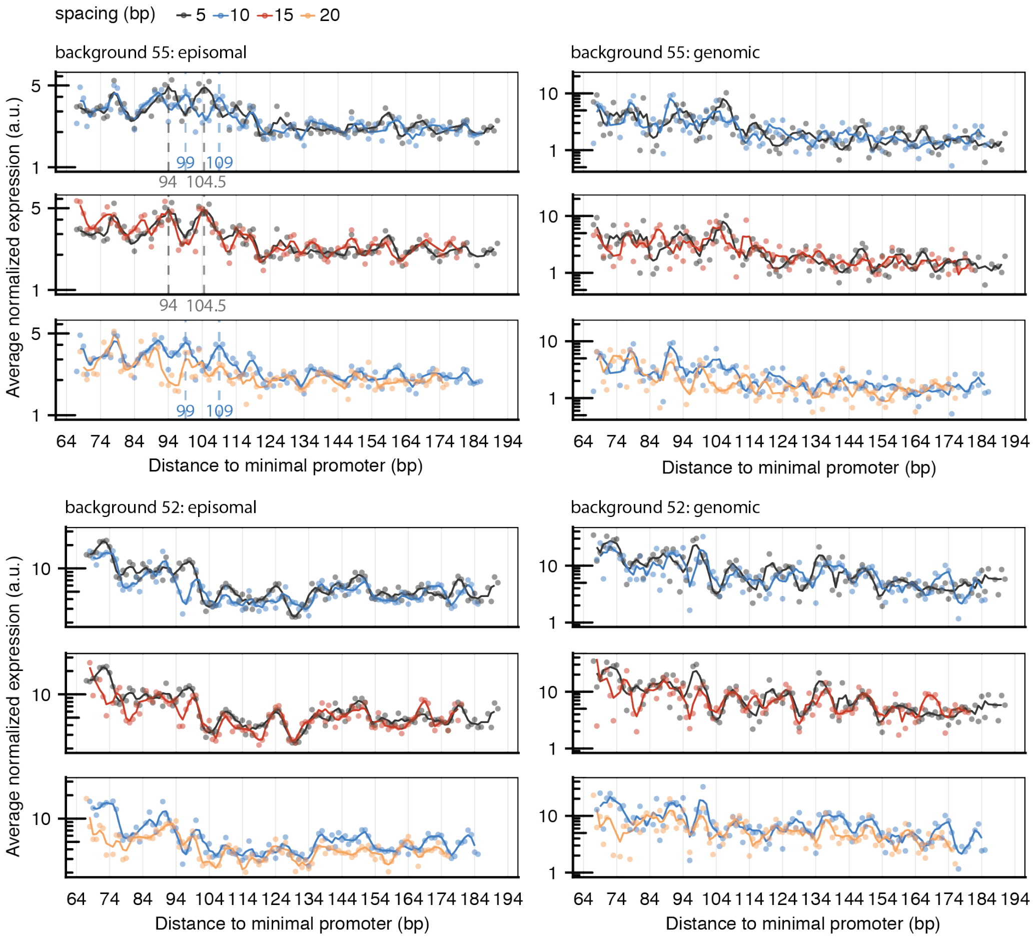


Figure 2.10 Variant periodicity offset graphs as in Fig. 3A with backgrounds 55 and 52 in both MPRAs. Average variant expression across replicates according to proximal CRE *distance* to the promoter (x-axis) is subset according to background and MPRA and colored by the *spacing* between CREs. The line corresponds to a 3 bp moving average estimate. Variants with background 55 in the episomal MPRA display a similar offset between 5 and 10 bp *spacings* and alignment between both 5 and 15 bp and 10 and 20 bp *spacings*. Dashed lines correspond points at local expression maxima across CRE *distances* or the midpoint between points if a maxima was not apparent in variants with 5 and 10 bp CRE *spacings*. Similar periodicity offsets and alignments are not as pronounced along this background in the genomic MPRA and along background 52 in both MPRAs.

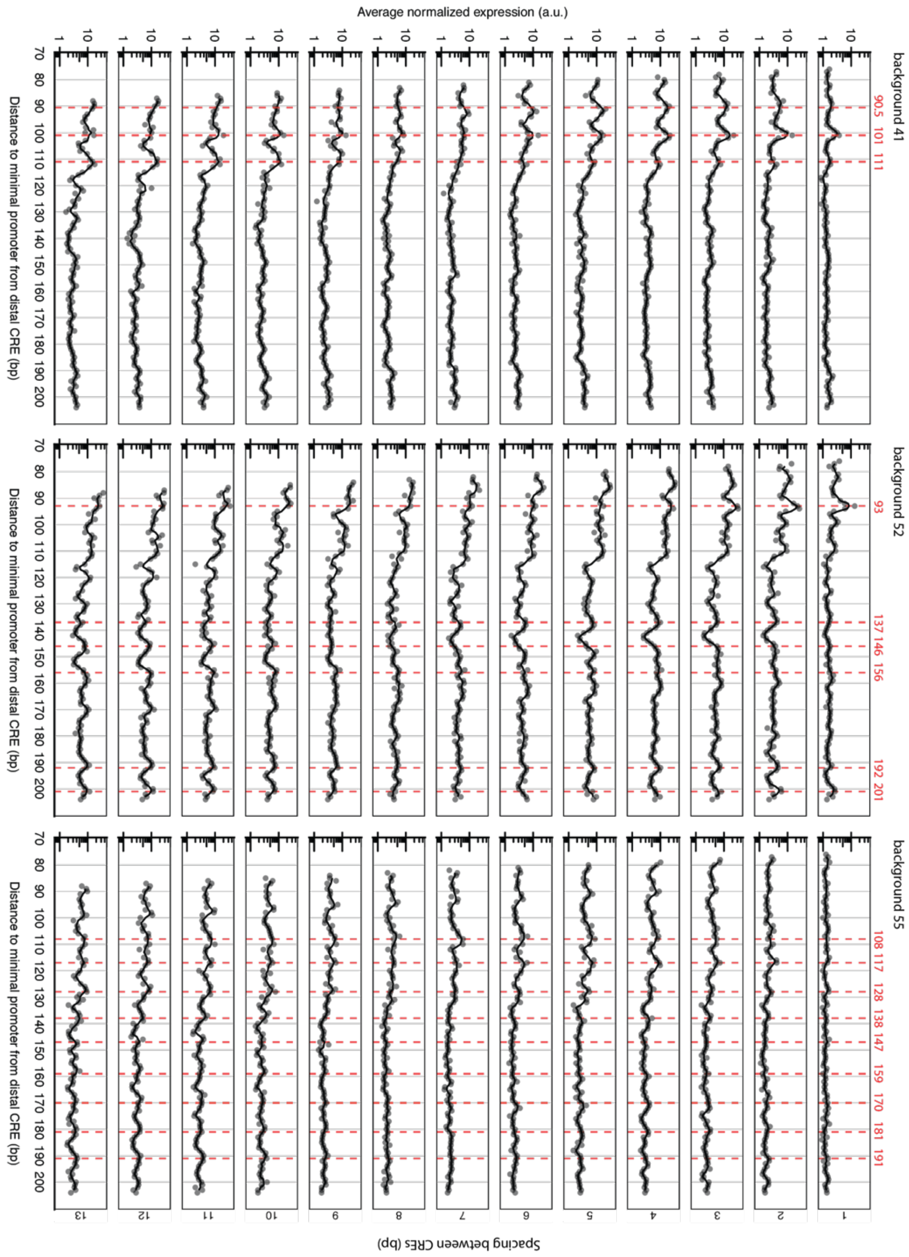


Figure 2.11 Expression measurements for all variants in the follow-up CRE Spacing and Distance library retained in analysis. Average normalized variant expression across replicates in an episomal MPRA is indicated according to distal CRE *distance* to the minimal promoter. Variants are subset by background and *spacing* between CREs. The line corresponds to a 3 bp moving average estimate while dashed red lines correspond to clear points of local expression maxima, determined as averages across CRE *spacings.* While ~10 bp periodicity is not constant across all *distances* along backgrounds (background 52), when present, expression periodicity follows similar phasing across *spacings* when plotted according to the distal CRE *distance* to the minimal promoter.

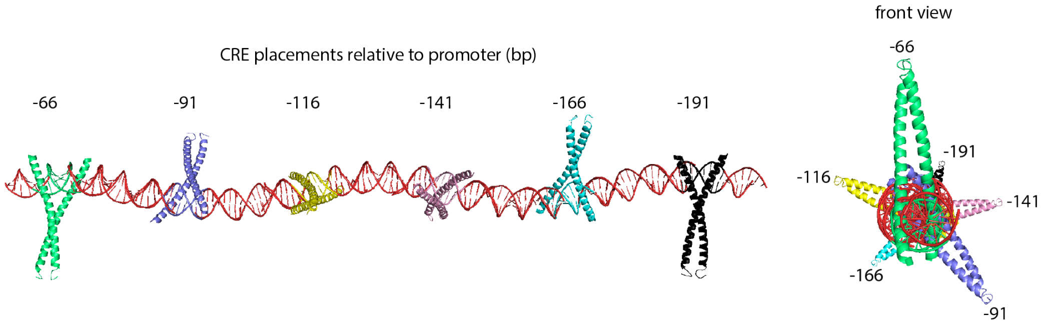


Figure 2.12 CREs in the Number and Affinity library were placed at positions which were expected to sample multiple CREB protein orientations along DNA. Using the published structure of a CREB::bZIP dimer bound to CRE (PDB: 1DH3), we modeled the expected placement of CREB dimers bound to the 6 CRE sites and constant 17 bp CRE *spacing* used in the CRE Number and Affinity library. Orientation of each CREB dimer is indicated relative to the first dimer following CRE *distances* relative to the minimal promoter.

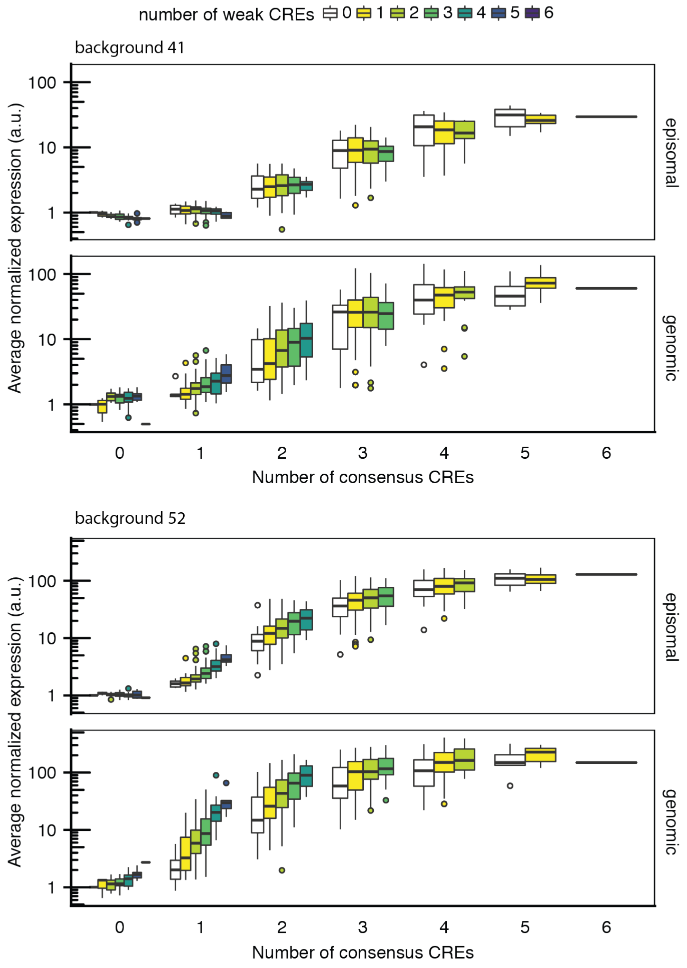


Figure 2.13 The number of consensus CREs in variants largely determines expression while the number of weak CREs has a variable effect between episomal and genomic MPRAs. Similar as in Fig. 4A, variants with background 41 and 52 in both MPRAs grouped according to their total number of both consensus (x-axis) and weak CREs (colored subsets) and average expression plotted per variant per MPRA (y-axis). The number of consensus CREs largely determines the expression per variant and drives a non-linear trend in expression. For reasons that are not apparent, the number of weak CREs per variant drives a similarly non-linear increase in expression across all variants but those with background 41 in the episomal MPRA.

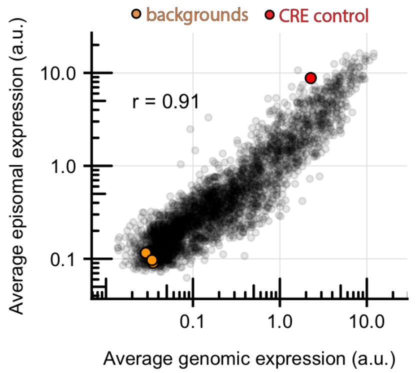


Figure 2.14 Library member expression largely correlates between MPRA formats. Despite differences, library variant expression correlates well between the episomal and genomic MPRA (r = 0.91).

Table 2.1. List of primers and sequences used throughout this study.

Subpool_2_F	GCTCTCCGCTATCAGTAACA
Subpool_3_F	CCGATAGGAGGGGAGAGTTA
Subpool_5_F	ATTACCATGTTATCGGGCGA
Subpool_2_R	CCAAATAGGATGTGTGCTCG
Subpool_3_R	CTGGTATAGTCTCCTCAGCG
Subpool_5_R	ATCTAAACCACGACCTCAGG
SP2_Biotin	/5Biosg/GCTCTCCGCTATCAGTAACA
SP3_Biotin	/5Biosg/CCGATAGGAGGGGAGAGTTA
SP5_Biotin	/5Biosg/ATTACCATGTTATCGGGCGA
SP2_Biotin_BC_R	/5Biosg/AAGTCGACTAGTNNNNNNNNNNNNNNNNNNNBTCTAGACCAAATAGGATGTGTGCTCG
SP3_Biotin_BC_R	/5Biosg/AAGTCGACTAGTNNNNNNNNNNNNNNNNNNNBTCTAGACTGGTATAGTCTCCTCAGCG
SP5_Biotin_BC_R	/5Biosg/AAGTCGACTAGTNNNNNNNNNNNNNNNNNNNBTCTAGAATCTAAACCACGACCTCAGG

Amp_minPLuc2_Biotin_Rev	/5Biosg/CACAGGAAACAGCTATGACC
Amp_minPLuc2_Biotin_For	/5Biosg/ACGACGTTGTAAAACGACGG
LibSeq_P5_Rev	AATGATACGGCGACCACCGAGATCTACACGTAACCACCCTGATCGACGG
LibSeq_P7_For	CAAGCAGAAGACGGCATACGAGATTCGGCAGTTGGGAAGAGCATAGTCG
LibSeq_R1Seq_Rev	GTAACCACCCTGATCGACGGGGAGTGTACTAGT
LibSeq_R2Seq_For	TCGGCAGTTGGGAAGAGCATAGTCGTAGAGCACGCGT
Creb_Hand_RT	ATGCTCTTCCCAACTGCCGACGACGGGGAGTGTACTAGT
Creb_Hand	ATGCTCTTCCCAACTGCCGA
Creb_Seq_Luc_R	TACAACCGCCAAGAAGCTGC
P5_Seq_Luc_F	AATGATACGGCGACCACCGAGATCTACACTACAACCGCCAAGAAGCTGC
P7_Ind_11_Han	CAAGCAGAAGACGGCATACGAGATCGAGGCTGGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_5_Han	CAAGCAGAAGACGGCATACGAGATACCCAGCAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA

P7_Ind_2_Han	CAAGCAGAAGACGGCATACGAGATGTGTGGTGGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_3_Han	CAAGCAGAAGACGGCATACGAGATTGGGTTTCGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_4_Han	CAAGCAGAAGACGGCATACGAGATCATGCCTAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_D3_Han	CAAGCAGAAGACGGCATACGAGATAACCCCTCGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R0A_Han	CAAGCAGAAGACGGCATACGAGATATCACGACGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R0B_Han	CAAGCAGAAGACGGCATACGAGATTAAGGCGAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R1A_Han	CAAGCAGAAGACGGCATACGAGATAAGAGGCAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R1B_Han	CAAGCAGAAGACGGCATACGAGATGCTACGCTGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R2A_Han	CAAGCAGAAGACGGCATACGAGATCAGAGAGGGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R2B_Han	CAAGCAGAAGACGGCATACGAGATAGGAGTGGGCG

	TGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R4A_Han	CAAGCAGAAGACGGCATACGAGATCAGATCCAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R4B_Han	CAAGCAGAAGACGGCATACGAGATGGACTCCTGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R8A_Han	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_Ind_R8B_Han	CAAGCAGAAGACGGCATACGAGATACAAACGGGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_In_R16A_Han	CAAGCAGAAGACGGCATACGAGATAGGCAGAAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_In_R16B_Han	CAAGCAGAAGACGGCATACGAGATCGTACTAGGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_In_R25A_Han	CAAGCAGAAGACGGCATACGAGATCTCTCTACGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_In_R25B_Han	CAAGCAGAAGACGGCATACGAGATAGGGTCAAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_In_R32A_Han	CAAGCAGAAGACGGCATACGAGATGTAGAGGAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA

P7_In_R32B_Han	CAAGCAGAAGACGGCATACGAGATTGTGACCAGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_In_R64A_Han	CAAGCAGAAGACGGCATACGAGATACAGTGGTGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
P7_In_R64B_Han	CAAGCAGAAGACGGCATACGAGATCCCAACCTGCGTGCTCTACGACTATGCTCTTCCCAACTGCCGA
Creb_Ind_Seq_P	TCGGCAGTTGGGAAGAGCATAGTCGTAGAGCACGC
Creb_R1_Seq_P	CCAAGAAGGGCGGCAAGATCGCCGTGTAATAATTCTAGA
Creb_RT_Hand_3	ATGCTCTTCCCAACTGCCGAAACCACCCTGATCGACGGGG

References

Belliveau, N.M., Barnes, S.L., Ireland, W.T., Jones, D.L., Sweredoski, M.J., Moradian, A., Hess, S., Kinney, J.B., and Phillips, R. (2018). Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. Proc. Natl. Acad. Sci. U. S. A. *115*, E4796–E4805.

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A.S., Yu, T., Marie-Nelly, H., McSwiggen, D.T., Kokic, G., Dailey, G.M., Cramer, P., et al. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. Nat. Struct. Mol. Biol. *25*, 833–

840.

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

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*.

Cohen, R.N., van der Aa, M.A.E.M., Macaraeg, N., Lee, A.P., Szoka, F.C., and Jr. (2009). Quantification of plasmid DNA copies in the nucleus after lipoplex and polyplex transfection. J. Control. Release *135*, 166–174.

van Dijk, D., Sharon, E., Lotan-Pompan, M., Weinberger, A., Segal, E., and Carey, L.B. (2017). Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators. Genome Res. *27*, 87–94.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment on transcription factor binding across diverse protein families. Genome Res. *25*, 1268–1280.

Duportet, X., Wroblewska, L., Guye, P., Li, Y., Eyquem, J., Rieders, J., Rimchala, T., Batt, G., and Weiss, R. (2014). A platform for rapid prototyping of synthetic gene networks in mammalian cells. Nucleic Acids Res. *42*, 13440–13451.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S., and Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat. Biotechnol. *34*, 1180–1190.

Eroshenko, N., Kosuri, S., Marblestone, A.H., Conway, N., and Church, G.M. (2012). Gene Assembly from Chip-Synthesized Oligonucleotides. Curr. Protoc. Chem. Biol. *2012*.

Fan, F., and Wood, K.V. (2007). Bioluminescent Assays for High-Throughput Screening. Assay Drug Dev. Technol. *5*, 127–136.

Fiore, C., and Cohen, B.A. (2016). Interactions between pluripotency factors specify *cis* - regulation in embryonic stem cells. Genome Res. *26*, 778–786.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.

Giniger, E., and Ptashne, M. (1988). Cooperative DNA binding of the yeast transcriptional activator GAL4. Proc. Natl. Acad. Sci. U. S. A. *85*, 382–386.

Gonzalez, G.A., and Montminy, M.R. (1989). Cyclic AMP stimulates somatostatin gene transcription by phosphorylation of CREB at serine 133. Cell *59*, 675–680.

Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E., et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. U. S. A. 201621150.

Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E.G., Berg, O.G., and Elf, J. (2012). The lac

Repressor Displays Facilitated Diffusion in Living Cells. Science *336*, 1595–1598.

Hebbar, P.B., and Archer, T.K. (2008). Altered histone H1 stoichiometry and an absence of nucleosome positioning on transfected DNA. J. Biol. Chem. *283*, 4595–4601.

Huang, Q., Gong, C., Li, J., Zhuo, Z., Chen, Y., Wang, J., and Hua, Z.-C. (2012). Distance and helical phase dependence of synergistic transcription activation in cis-regulatory module. PLoS One *7*, e31198.

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., and Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. *27*, 38–52.

Jeong, S., and Stein, A. (1994). Micrococcal nuclease digestion of nuclei reveals extended nucleosome ladders having anomalous DNA lengths for chromatin assembled on non-replicating plasmids in transfected cells. Nucleic Acids Res. *22*, 370–375.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *advance on*.

Jones, E.M., Lubock, N.B., Venkatakrishnan, A.J., Wang, J., Tseng, A.M., Paggi, J.M., Latorraca, N.R., Cancilla, D., Satyadi, M., Davis, J., et al. (2019). Structural and Functional Characterization of G Protein-Coupled Receptors with Deep Mutational Scanning.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811.

Kim, T.K., and Maniatis, T. (1997). The mechanism of transcriptional synergy of an in vitro assembled interferon-beta enhanceosome. Mol. Cell *1*, 119–129.

Kim, J., Lu, J., and Quinn, P.G. (2000). Distinct cAMP response element-binding protein (CREB) domains stimulate different steps in a concerted mechanism of transcription activation. Proc. Natl. Acad. Sci. U. S. A. *97*, 11292–11296.

Kim, S., Broströmer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y.Q., et al. (2013). Probing allostery through DNA. Science *339*, 816–819.

Kim, T.K., Kim, T.H., and Maniatis, T. (1998). Efficient recruitment of TFIIB and CBP-RNA polymerase II holoenzyme by an interferon-beta enhanceosome in vitro. Proc. Natl. Acad. Sci. U. S. A. *95*, 12191–12196.

Klein, J., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2019). A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays.

Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc. Natl. Acad. Sci. U. S. A. *109*, 19498–19503.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–

665.

Lee, T.-H., and Maheshri, N. (2012). A regulatory role for repeated decoy transcription factor binding sites in target gene expression. Mol. Syst. Biol. *8*, 576.

Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A.C., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. Genome Res. *25*, 1018–1029.

Levo, M., Avnit-Sagi, T., Lotan-Pompan, M., Kalma, Y., Weinberger, A., Yakhini, Z., and Segal, E. (2017). Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. Mol. Cell *65*, 604–617.e6.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22.

Lizio, M., Harshbarger, J., Abugessaisa, I., Noguchi, S., Kondo, A., Severin, J., Mungall, C., Arenillas, D., Mathelier, A., Medvedeva, Y.A., et al. (2017). Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. Nucleic Acids Res. *45*, D737–D743.

Matreyek, K.A., Stephany, J.J., and Fowler, D.M. (2017). A platform for functional assessment of large variant libraries in mammalian cells. Nucleic Acids Res. *45*, e102.

Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor CREB. Nat. Rev. Mol. Cell Biol. *2*, 599–609.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277.

Montminy, M.R., Sevarino, K.A., Wagner, J.A., Mandel, G., and Goodman, R.H. (1986). Identification of a cyclic-AMP-responsive element within the rat somatostatin gene. Proc. Natl. Acad. Sci. U. S. A. *83*, 6682–6686.

Panne, D. (2008). The enhanceosome. Curr. Opin. Struct. Biol. *18*, 236–242.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. *43*, 73–81.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Schumacher, M.A., Goodman, R.H., and Brennan, R.G. (2000). The structure of a CREB bZIP.somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. J. Biol. Chem. *275*, 35242–35247.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *advance on*.

Tabach, Y., Brosh, R., Buganim, Y., Reiner, A., Zuk, O., Yitzhaky, A., Koudritsky, M., Rotter, V., and Domany, E. (2007). Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. PLoS One *2*, e807.

Takahashi, K., Vigneron, M., Matthes, H., Wildeman, A., Zenke, M., and Chambon, P. (1986). Requirement of stereospecific alignments for initiation from the simian virus 40 early promoter. Nature *319*, 121–126.

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell 83, 1091–1100.

Tinti, C., Yang, C., Seo, H., Conti, B., Kim, C., Joh, T.H., and Kim, K.S. (1997). Structure/function relationship of the cAMP response element in tyrosine hydroxylase gene transcription. J. Biol. Chem. *272*, 19158–19164.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. *35*, D88–D92.

Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A., and Segal,

E. (2019). Systematic interrogation of human promoters. Genome Res. *29*, 171–183.

White, M.A. (2015). Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. Genomics *106*, 165–170.

White, M.A., Kwasnieski, J.C., Myers, C.A., Shen, S.Q., Corbo, J.C., and Cohen, B.A. (2016). A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors.

Xu, Z., Thomas, L., Davies, B., Chalmers, R., Smith, M., and Brown, W. (2013). Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. BMC Biotechnol. *13*, 87.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*.

Zhang, X., Odom, D.T., Koo, S.-H., Conkright, M.D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., et al. (2005). Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. Proc. Natl. Acad. Sci. U. S. A. *102*, 4459–4464.

Zhu, F., Gamboa, M., Farruggio, A.P., Hippenmeyer, S., Tasic, B., Schüle, B., Chen-Tsai, Y., and Calos, M.P. (2014). DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. Nucleic Acids Res. *42*, e34.

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S.O., Nitta, K.R., Morgunova, E., Taipale, M., et al. (2018). The interaction landscape between transcription factors and the nucleosome. Nature *562*, 76–81.

CHAPTER THREE

Conclusions and Future Directions

Comparison between episomal and single-copy, genomic MPRA formats

We present the first systematic comparison between the more typical, episomal MPRA format and a single-copy, genomically-integrated MPRA in mammalian cells exploring sequence-function relationships of TFBS architecture. Genomic MPRAs are becoming more popular due to their expected approximation of genomic transcription by being placed in a more natural, chromatin context (Inoue et al., 2017; Klein et al., 2019; Weingarten-Gabbay et al., 2019, Maricque et al., 2018). Yet these assays require more time, resources, experimental steps, and in certain cases special cell lines, as used here. Additionally, they place restrictions on the number of constructs feasibly tested and can suffer from loss of library coverage as observed in this study. Therefore, single-copy genomic MPRAs limit the high-throughput nature of the MPRA itself. Nevertheless, these limitations may be accommodated if episomal assays indeed fail to characterize genomic trends in expression.

Here we compare MPRA formats between library members varying cis-regulatory architecture of a model TFBS and find a number of differences. 1) Although variations in CRE *distance* drive similar expression periodicities between MPRA formats, the amplitude of this effect is heightened in the genomic MPRA. 2) In addition to this, the role of *spacing* on expression periodicity varied between MPRA formats. Although variants containing the background with the clearest change in expression periodicity in the episomal MPRA were dropped from the genomic MPRA analysis due to sequencing coverage, comparisons between the other backgrounds did not indicate similar changes in periodicity phasing from CRE *spacings* in the genomic MPRA. This may imply the precise positioning of CREs upstream of a promoter may have different effects on expression in a genomic context. In line with the finding that CREB protein orientation near nucleosomes modulates CRE binding occupancy (Zhu et al., 2018), further MPRA characterizations of nucleosomal deposition along variants with varied CRE *spacings* and *distances*, similar to (Levo et al., 2017), may explain this disconnect. 3) Lastly, as indicated in Chapter 2, the role of CRE affinity in driving expression differed greatly between assay formats. Yet this may have been an artifact of testing sequences containing variations of the same TFBS. Thus TFBS affinity may not play as large of a role in episomal MPRAs performed on library members with diverse TFBSs.

Overall, the comparison between single-copy, genomic and episomal MPRAs implies episomal assays can approximate overall trends but may miss out on some of the finer details obtained in genomic assays. While lentiviral assays present their own issues, increasing copy numbers of genomic assays even slightly using other means may simplify experimental methods. Using established and well-characterized cell lines with multiple landing pads, similar to the single-copy version here, would simplify genomic MPRAs by decreasing the cells and RNA/DNA required for processing while still avoiding TF competition with 10^1-10^4 copy numbers in transfected MPRAs. While library member expression can differ across landing pads due to local chromatin effects (Maricque et al., 2018), cell lines can be generated containing landing pads in loci with similar effects on constructs.

Current understandings of sequence-function relationships governing CRE-directed gene expression

By manipulating regulatory architecture in a more controlled manner rather than relying on genomic sequences, we have isolated the effects of a number of variables influencing expression from CREs. First and foremost, CRE number and affinity are the largest determinants of expression, as would be expected. Similar to other findings (Gertz et al., 2009; Sharon et al., 2012; Weingarten-Gabbay et al., 2019), CRE number followed a non-linear trend in expression, with combined CRE affinity along variants scattered throughout the overall trend. We anticipate within this trend, which is shaped by active TF levels (van Dijk et al., 2017), CRE *distance*, *spacing*, and the surrounding sequence content further modulates expression. The overall *distance* between CRE and the minimal promoter tested here largely influenced CRE activity, both observed in the placement of multiple CREs along variants and when iteratively tested in the CRE Spacing and Distance library. Yet, assay designs constricted the distances tested to within 67 bp of the minimal promoter, curtailing the full characterization of CRE *distance* effects. Further characterizations minimizing this distance may reveal greater increases in CRE activity and may also provide more insight on CRE's expression periodicity and *spacing* effects, especially on the role of the distal CRE in driving periodicity phasing. Additionally, CRE *spacing* may have played a role in the trends observed from the CRE Number and Affinity library expression. More sequences manipulating CRE *spacings* and *distances* in cis-regulatory elements with greater than 2 CREs is needed to explore the full impact non-optimal *spacing* has on overall expression in these cases. Lastly, the role of surrounding sequence content was not fully explored here as only 3 backgrounds were tested, but further work across more backgrounds may reveal an overall trend or motif co-occurrence following CRE activity. In conjunction with more sequences, it is likely necessary to assay combinations of features tested here to fully capture the variation in expression from CREs in cis-regulatory elements. For instance, increasing GC content between binding sites for other proteins has ameliorated the ~10 bp binding periodicity observed from varied site *spacings* (Kim et al., 2013). Construction of a more predictive model, following additional characterizations of larger libraries of CRE manipulations in addition to molecular modeling, may accurately capture all these features and likely would rely on a combination of the overall CRE distance to promoter motifs, CRE *distance* within ~10 bp intervals, local sequence content and its interaction with CRE *spacing* and *distance* effects, and overall CRE number and affinity.

Extrapolation of findings to other TFBSs

I expect many of the features assayed here to also play a role in expression from cis-regulatory elements with non-CRE TFBSs. Overall TFBS number and affinities will likely also fill out an overall non-linear trend in expression, although the combination of active TF levels per cell state would influence the magnitude of this trend. Of course, the combinatorial logic of TF pairs and their conserved TFBS arrangements will guide the effects of specific TFBS combinations. This was not assayed here and presents an added layer of difficulty in predicting cis-regulatory element activity from sequence alone. Some TFs do not activate on their own and require obligatory partners in driving transcription (Stampfel et al., 2015), yet confirming which combinations of TFs drives expression presents itself as its own heroic endeavor alone. Although obligatory TFBS *spacings* and orientations have been mapped for many combinations of TFs *in vitro* (Jolma et al., 2013, 2015), these relationships are further complicated by local sequence content and overall distances to non-interacting TFs or members of the pre-initiation complex within cis-regulatory elements. Additionally, expanding the trends observed here to other TFBSs that localize around TSSs may reveal similar or different distance-dependencies, perhaps based on specific interactions to general transcription factors and coactivators. In the end, I believe the strength of sequence-based prediction models will lie in approximating expression from cis-regulatory elements, with precise estimations only for those with TFBSs that have been well-characterized. This precision will be guided by measurements of active TF levels per cell type and state, obligatory TFBS *spacings* and partners, TF-induced local changes to the DNA (as mentioned with potential *spacing* effects in Chapter Two), overall *distance*-dependence to promoter elements, specific distance effects following TF helical phasing, and lastly local sequence content. I anticipate a combination of biophysical-based models along with high-throughput sequence characterizations similar to that which is presented here will aid such predictions of sequence-function relationship driving gene expression.

REFERENCES

van Dijk, D., Sharon, E., Lotan-Pompan, M., Weinberger, A., Segal, E., and Carey, L.B. (2017). Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators. Genome Res. *27*, 87–94.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., and Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. *27*, 38–52.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *advance on*.

Kim, S., Broströmer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y.Q., et al. (2013). Probing allostery through DNA. Science *339*, 816–819.

Klein, J., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2019). A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays.

Levo, M., Avnit-Sagi, T., Lotan-Pompan, M., Kalma, Y., Weinberger, A., Yakhini, Z., and Segal, E. (2017). Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. Mol. Cell *65*, 604–617.e6.

Maricque, B.B., Chaudhari, H.G., and Cohen, B.A. (2018). A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. Nat. Biotechnol.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *advance on*.

Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A., and Segal, E. (2019). Systematic interrogation of human promoters. Genome Res. *29*, 171–183.

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S.O., Nitta, K.R., Morgunova, E., Taipale, M., et al. (2018). The interaction landscape between transcription factors and the nucleosome. Nature *562*, 76–81.