Spring 3-4-2014

cis-Regulation in the Mammalian Rod Photoreceptor

Jamie C. Kwasnieski
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology & Biomedical Sciences
Molecular Genetics & Genomics

Dissertation Examination Committee:
Barak A. Cohen, Chair
Joseph C. Corbo
Joseph D. Dougherty
James J. Havranek
Robi D. Mitra
Tim Schedl

cis-Regulation in the Mammalian Rod Photoreceptor

By

Jamie C Kwasnieski

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2014
St. Louis, Missouri

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My thesis work would not have been possible without the support of many people. First, I would like to thank my advisor, Barak Cohen, for his guidance and encouragement. Without his mentorship, this work would not have been possible. I am grateful to have a mentor who sees the big picture, asks hard questions, and challenges me to take risks. Since joining the lab, Barak has taught me how to identify important questions, think about problems quantitatively and convey my thoughts clearly. Barak has cultivated a spirit of rigor and curiosity in the lab and it is truly a great place to learn.

I would like to thank my thesis committee—Joe Corbo, Joe Dougherty, Jim Havranek, Rob Mitra, and Tim Schedl—for their advice and critical feedback, which has been crucial throughout my project. I am grateful to Melanie Relich and Margie Andersohn for their help with the administrative aspects of my graduate work.

In particular, I would like to thank Joe Corbo and his lab for their close collaboration on this project. I am thankful for the opportunity to work in the mammalian retina; it has been a complex and interesting system for the study of transcriptional regulation. Joe's passion and excitement has been critical in propelling this work forward. I would like to thank Connie Myers for helping me with the retinal experiments. Her hard work, patience, and attention to detail have made this study possible. I would also like to thank Susan Shen for her guidance and technical advice regarding retinal ChIP.

I have enjoyed my time in the Cohen lab, and I cannot imagine training in a better environment. To Ilaria Mogno, thank you for being my partner throughout the development of CRE-seq. I appreciate your support, ideas, and persistence throughout the project. To Chris Fiore and Hemangi Chaudhari, it has been a pleasure working with you on the ENCODE project—your enthusiasm and excitement made the study possible. To Chris Fiore, thank you for your technical and emotional support. It has been wonderful to have such a close friend in the lab. To Kim Lorenz, thank you for the emotional support, company during errands, proofreading help, and constant supply of chocolate. To Robert Zeigler, thank you for the R tutorials and the conversations about science. Your quantitative perspective has always challenged me and your passion for science is contagious. To Aaron Spivak, thank you for the technical advice regarding transcription factor binding. To Brett Maricque, thank you for being a patient and fun bay mate. To the other members of the Cohen lab: Priya Sudarsanam, Josh Witten, Linda Riles, Mike White,

Marc Sherman, Dana King, Devi Swain, Amy Li, and Darcy Nayler, thank you for your scientific feedback and friendly conversations.

I would like to thank the members of the CGS epigenetics journal club: Francesco Vallania, Maxim Schillebeeckx, Brett Maricque, Kate Chiappinelli, and Chris Fiore. The time we spent reading and discussing papers was valuable to my development as a scientist and challenged the way I thought about problems.

I would like to thank all of my friends, those near and far, because without them, the hard parts of graduate school would have been even more difficult. Specifically, thanks to Marie Strand for helping me

ccare for my cat Simon and for her company.

I cannot thank my parents enough for their support. Throughout my time in school they have made trips to St. Louis, sent me surprise packages, and helped me in ways I cannot even remember. I would especially like to thank Megan Falke, my sister and best friend. She has been an enthusiastic cheerleader and a reliable friend, even taking calls in the middle of the night to support me. I would like to thank my boyfriend's family, for welcoming me into their life and being interested in my work.

Most of all, I would like to thank my boyfriend, Ben Schriesheim, for his love and support. Ben has moved all the way from Boston to live with me in St. Louis. Having Ben in St Louis has made every day better because I have a partner for adventures and another cook in the kitchen, but especially because I can spend time with my best friend.

This work was funded by NIH grants: R01 RGM092910A, R21 HG006346-01, R01 HG006790-01, and T32 HG000045-15.

ABSTRACT OF THE DISSERTATION

cis-Regulation in the Mammalian Rod Photoreceptor

By

Jamie C Kwasnieski

Doctor of Philosophy in Molecular Genetics and Genomics

Washington University in St. Louis, 2014

Professor Barak A. Cohen, Chair

Transcription factors regulate the expression level of target genes by binding to *cis*-regulatory elements (CREs) present in gene promoters. The goal of my thesis research is to define the sequence components of CREs that determine transcriptional output. In order to accomplish this goal, I developed a method to measure the regulatory activity of thousands of CREs in a single experiment. In this method I insert unique barcodes in the 3'UTR of a reporter gene and multiplex expression measurements with RNA sequencing. Using this technique in explanted retinas, I determined the impact of single nucleotide variants in a mammalian promoter by measuring expression controlled by all single nucleotide variants of the Rhodopsin proximal promoter. I found that nearly all (86%) sequence variants drive significantly different activity than the wild-type promoter and that the mechanism of most variants can be interpreted as altered transcription factor binding. In addition, we found that the largest changes in expression resulted from variants located in characterized transcription factor binding site sequences. Next, I explored how combinations of binding sites drive particular levels of gene expression by utilizing a synthetic biology approach. I generated synthetic CREs composed of various combinations of binding sites found in the Rhodopsin promoter and measured the expression driven by these sequences. In this study I found that synthetic CREs containing binding sites for transcriptional activators yielded diverse expression outputs, including both activation and repression of a minimal promoter. Together, these experiments demonstrate that interactions between binding sites and dual regulation of a single binding site can produce diverse gene expression patterns. I conclude that simple *cis*-regulatory elements can produce complex expression outputs due to interactions between transcriptional activators and detailed quantitative models will be necessary to predict expression from these sequences.

CHAPTER 1: INTRODUCTION

Gene Regulation Underlies Development and Cellular Response to Environment

Changes in gene expression control the process of cell differentiation. The development of an organism begins when an egg is fertilized to become a zygote. These cells grow, divide, and differentiate from a population of identical cells to an adult organism composed of tissue and organ systems with distinct cell identities. Directions describing the appropriate expression levels and patterns for each gene are encoded in the four base pair code of the organism's genome and cells with distinct identities interpret these instructions differently.

Changes in gene expression also allow cells to respond to environmental stimuli, as demonstrated in the work by Francois Jacob and Jacques Monod. Jacob and Monod showed that at any time, not all of an organism's genes are expressed, and the regulation of gene levels is critical for a cell's ability to adapt to a changing environment (Jacob and Monod 1961). In *Escherichia coli*, cells altered expression of genes when there was a change in sugar supply or bacteriophage infection and this cellular response could be de-coupled from the regulation of the response. This decoupling is due to the actions of a class of genes that regulate the expression of other genes. Since this discovery, the field of transcriptional regulation has studied how regulatory proteins manage expression levels.

The expression level of a given gene can be controlled at many steps in the process: transcription initiation, RNA Polymerase II (RNA Pol II) elongation, mRNA transcript stability, translation efficiency, protein modification, and protein degradation. Of these steps, the regulation of transcription is most capable of generating diverse expression levels. Transcriptional regulation is critical for key biological processes. Understanding how this regulation is coordinated is an important step toward understanding cellular behavior, particularly during development and in response to environment.

Regulation of Transcription

Transcription factors (TFs) are proteins that regulate gene expression by binding to short specific DNA sequences called transcription factor binding sites (TFBS). Combinations of TFBS make up DNA sequences that modulate transcriptional regulation, like proximal promoter sequences or distal enhancer

elements. In particular, *cis*-regulatory elements (CREs) are sequences composed of TFBS that are located *cis* to gene bodies and control expression of the nearby gene. Transcriptional activators are TFs that promote recruitment of the RNA Pol II complex, whereas transcriptional repressors inhibit recruitment. TFs bind TFBS located within CREs, TFs recruit RNA Pol II machinery and set in motion a series of steps that begin transcription (Ptashne and Gann 1997; Ptashne 2005). The assembly of the RNA Pol II complex and subsequent transcription is well-described (Lee and Young 2000), but there are many outstanding questions about how and when transcription factors can recruit the RNA Pol II complex. Sequence features such as the number or affinity of a TFBS, can impact the control of gene expression (Yuh et al. 2001; Arnosti 2003; Rastegar et al. 2008) but it is not clear which patterns are important for gene activation. The major goal of this study is to understand the characteristics and sequence features that allow a CRE to control gene expression

Properties of *cis*-Regulation

There are several well-understood examples of *cis*-regulatory elements that illustrate how eukaryotic sequences are able to generate condition-specific gene expression levels. These examples often utilize combinatorial regulation by multiple TFs, cooperative TF binding, and competition between TFs to achieve condition or cell specific expression patterns. Three such examples are described.

Combinations of TFs are important to integrate cellular signals. As an example, we can examine the regulation of the *Saccharomyces cerevisae* galactokinase gene *GAL1* (Ptashne and Gann 2002), which is necessary for the metabolism of galactose. Yeast prefers glucose to galactose, so Gal1 is only expressed when galactose is the sole sugar in the media. Transcriptional activator Gal4 is constitutively bound to the upstream activating sequence (UAS) of the *GAL1* gene. In the presence of glucose, the transcriptional repressor Gal80 inhibits RNA Pol II and represses the *GAL1* gene. Upon sensing galactose, Gal80 repression is relieved and Gal4 is able to activate transcription of *GAL1*. If both glucose and galactose are present, the transcriptional repressor Mig1 represses transcription of *GAL1*. Thus, the *GAL1* gene is expressed only in the presence of galactose through the activity of condition-specific transcriptional repressors. This example demonstrates that several TFs are necessary to integrate different signals, a common feature of eukaryotic regulatory elements (Remenyi et al. 2004).

Cooperative TF binding is often utilized to achieve switch-like behavior for a gene that needs to be quickly activated (Carey 1998; Vashee et al. 1998). The human Interferon-beta (IFN-beta) gene is usually silenced but is activated in the presence of viral infection. Regulation of this gene utilizes TF cooperativity to achieve this dynamic behavior (Merika and Thanos 2001). The inputs of three signaling pathways activate TFs that bind to three distinct enhancer sequences. Upon TF binding, the three enhancers loop to the promoter of the IFN-beta gene to activate transcription. All three enhancers are necessary for activation of the gene and TF binding within these enhancers is highly cooperative (Carey 1998), allowing for robust activation of the gene.

Competitive TF binding results when TFs with overlapping specificity are expressed in the same cell at the same time. In the case of regulation of the battery of genes necessary to maintain pluripotency, paralogous TFs POU5F1 and POU2F1 are co-expressed in the same cells and bind the same DNA sequences (Phillips and Luisi 2000). Genome-wide studies have found that POU5F1 achieves specificity by binding near TFBS for the co-regulator NANOG, whereas POU2F1 achieves specificity by binding nearby SOX2 TFBS (Ferraris et al. 2011). Competitive binding of paralogous TFs is a side effect of the co-expression of TFs with overlapping specificities and binding site specificity is achieved by combinatorial binding with co-regulators.

Taken together, we learn that eukaryotic transcriptional regulation involves the integration of multiple TF binding events, cooperative TF binding, and competitive TF binding in order to generate expression patterns that have spatial and temporal specificity.

Developing Quantitative Models of Gene Expression

As these examples demonstrate, there is a general understanding of how transcription is regulated but the field lacks a quantitative model to describe how regulatory sequence controls gene expression level. It would be useful to predict the level of gene expression controlled by a particular sequence for two main reasons. First, this model would predict how gene expression levels are altered by sequence mutations, aiding in the interpretation of human genetic variation. Most of the variation in the human genome is located in non-coding regions of the genome (Abecasis et al. 2010). In fact, a large fraction of GWAS hits are found in non-coding sequence, suggesting that these polymorphisms have

phenotypic consequences (Kasowski et al. 2010; Maurano et al. 2012). In addition, we have little intuition for how these variants might impact gene expression of the individual. Second, this model would predict the expression controlled by novel regulatory sequences, allowing for engineering of synthetic *cis*-regulatory sequences that drive a gene product at desired levels and patterns. The ability to engineer targeted gene expression levels would be useful in a variety of applications, including the development of gene therapies, the modification of agricultural crops, and the production of biofuels.

In order to formulate such predictive models, a more detailed understanding of *cis*-regulatory sequences must be established. To start, it would be useful to understand how common sequence properties impact gene expression. Properties including binding site affinity (Tanay 2006), binding site combinations (Zinzen et al. 2009; Slattery et al. 2011; Rembold et al. 2014), and DNA shape (Parker et al. 2009) can effect gene expression control, in addition to genomic features like nucleosome occupancy (Polach and Widom 1996; Lam et al. 2008; Kaplan et al. 2009). Developing a quantitative model for a particular sequence would require manipulating the sequence a large number of times, cataloging the resulting expression, and developing a framework that fits this behavior.

There exist several commonly used quantitative frameworks to predict expression from DNA sequence. Methods utilizing thermodynamic principles (Shea and Ackers 1985; Buchler et al. 2003; Segal et al. 2008; Gertz et al. 2009; Raveh-Sadka et al. 2009; Wasson and Hartemink 2009) generate parameter sets that represent biophysical interactions. Other approaches use linear regression (Bussemaker et al. 2001), dynamical equations (Goutsias and Kim 2004), Bayesian networks (Beer and Tavazoie 2004), stochastic models (Paulsson 2005) and hybrid methods (Setty et al. 2003). The choice of a particular model framework is made based on the available type of gene expression data.

Regardless of the specified model framework, there generally does not exist enough to fit these models. The Encyclopedia of DNA Elements Consortium (ENCODE) was established to catalog all active regulatory elements in the human genome (Birney et al. 2007; Dunham et al. 2012). As part of the ENCODE dataset, several labs have measured RNA-seq for many cell types, but connecting these expression levels to the regulatory elements that are responsible for the expression remains a difficult task for scientists. Thus, these efforts have gathered large datasets but they are not systematic enough to be useful in training many of the previously mentioned quantitative models. This study focuses on the

development of high throughput assay development in order to make progress toward a more detailed quantitative model of gene expression.

High Throughput Reporter Assay Development

At the time I began my thesis work there were a variety of high-throughput techniques that could be used to identify mammalian sequences with regulatory potential. These techniques include methods to identify open chromatin with DNase Hypersensitivity sequencing (DNase-seq) or Formaldehyde-Assisted Isolation of Regulatory Element sequencing (FAIRE-seq) (Giresi et al. 2007; Boyle et al. 2008), methods to identify genomic regions where regulating TFs are bound by chromatin immunoprecipitation sequencing (ChIP-seq) (Solomon et al. 1988; Harbison et al. 2004; Johnson et al. 2007; Visel et al. 2009), RNA sequencing (RNA-seq) methods to identify transcribed portions of the genome (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Kim et al. 2010), and chromatin capture methods to identify distal sequences that are looped into core promoters (Dekker et al. 2002; Dostie et al. 2006). These technologies are widely used in the field to identify putative enhancers. Many groups have looked for patterns in this compendium of enhancers, but any identified TFBS patterns are based on correlation between sequence and expression, rather than a result of direct manipulation. Thus, there was a gap between our ability to generate putative enhancers and our ability to functionally test sequences for activity.

Traditionally, reporter assay techniques have been used to determine if a DNA sequence has the ability to control expression of a downstream gene. In this approach, a putative *cis*-regulatory sequence is cloned upstream of a minimal promoter driving expression of a visualizable reporter gene, such as Green Fluorescent Protein gene or Firefly Luciferase gene. This DNA construct is transfected or electroporated into the tissue of interest and activity is determined by how much the CRE modifies expression of the minimal promoter. Reporter assay analysis allows direct functional testing of *cis*-regulatory function.

Reporter assays are widely used, although the need for serial construction and testing makes this a low throughput technique that requires large amounts of cell material. Much time and effort is required to develop even a modest set of reporter genes, which are necessary for a simple experiment. For example, if one wanted to understand how sequence mutations impact the control of gene expression,

one might create a series of constructs, each with a single change in the wild type sequence. Even for a short 10 bp sequence, examining the three possible mutations at each position would require 30 DNA constructs, which is time intensive when utilizing a low throughput construction approach. In addition, the traditional assay would require at least 30 separate cell samples for serial assay, not including those necessary for biological replicates.

Despite the amount of time it takes, several groups have used this strategy to develop large numbers of reporter genes (Chiang et al. 2006; Ligr et al. 2006; Cox et al. 2007; Kinney et al. 2007; Kinkhabwala and Guet 2008; Gertz and Cohen 2009; Gertz et al. 2009; Landolin et al. 2010; Schlabach et al. ; Yun et al. 2012). Some groups have attempted to improve throughput by employing randomized libraries, but because of the random nature of the construction, these techniques do not provide a systematic set of constructs. In order to gain a detailed understanding of how sequence characteristics affect gene expression control, expression comparisons of all sequence derivatives are necessary. Therefore a higher throughput reporter assay technique was needed.

cis-Regulation in the Mammalian Retina

There are three main reasons why the mammalian rod photoreceptor cell is useful system for the development of a high throughput reporter assay. First, the mammalian retina is an accessible tissue that allows for nearly homogeneous cell analyses. The mouse retina is readily dissectible and it is possible to electroporate transgenic DNA directly into the live or explanted retinas (Matsuda and Cepko 2004; Hsiau et al. 2007). The retina is composed of several distinct cell types, but 70% (Carter-Dawson and LaVail 1979; Roorda and Williams 1999) of the cells in the mammalian retina are photoreceptors and roughly 73% of photoreceptors are rod cells. We are able to measure gene expression from rod cells specifically because we electroporate transgenic DNA after other common cell types have completed mitosis, but before rod precursor cells finish dividing (Swaroop et al. 2010). By electroporating DNA into explanted retinas on postnatal day 0 and harvesting RNA at postnatal day 8, we are able to accomplish this.

Second, the transcription factors that regulate rod cell differentiation and homeostasis are well characterized. Rod cell gene expression is controlled by three main factors: CRX, NRL, and Nr2e3. The cone/rod homeobox TF, CRX, specifies photoreceptor cell fate in the retina (Freund et al. 1997;

Furukawa et al. 1997; Furukawa et al. 1999; Hennig et al. 2008). CRX is a transcriptional activator (Chen et al. 1997), that may act by recruiting chromatin remodeling enzymes (Peng and Chen 2007). The neural leucine zipper transcriptional activator, NRL, is both necessary and sufficient for rod cell differentiation (Mears et al. 2001; Oh et al. 2007). The orphan nuclear receptor transcription factor Nr2e3 is a dual transcriptional regulator that is important for the repression of cone genes in rod cells (Kobayashi et al. 1999; Chen et al. 2005; Peng et al. 2005; Webber et al. 2008).

In addition, binding specificity (Kataoka et al. 1994; Chen et al. 2005; Montana et al. 2011b) and genomic binding preferences (Corbo et al. 2010; Hao et al. 2012) have been characterized for these factors, so we have a complete understanding of *in vitro* and *in vivo* binding preferences for these TFs. CRX, NRL and Nr2e3 have been shown to co-regulate at rod specific genes (Rehemtulla et al. 1996; Cheng et al. 2004; Peng et al. 2005), signifying that they are important for transcriptional control in the rod cell. In addition, other experiments have been done to characterize rod cell gene expression at the transcriptome level (Livesey et al. 2000; Blackshaw et al. 2001; Akimoto et al. 2006; Hsiau et al. 2007). These studies are useful because they provide a comprehensive view of expressed genes and other potential regulators.

Third, we know that gene regulation in the mammalian retina has implications for human health. The Rhodopsin (Rho) gene is the most highly expressed gene in the mammalian retina; it allows rod photoreceptors to coordinate black/white and dim light vision. Rho dysregulation can result in retinal degeneration and subsequent vision loss (Humphries et al. 1997). In addition, mutations in known regulators CRX (Freund et al. 1997; Freund et al. 1998; Sohocki et al. 1998; Rivolta et al. 2001), NRL(Bessant et al. 1999), and Nr2e3 (Haider et al. 2006) have been associated with retinal diseases such as autosomal dominant Retinitis Pigmentosa, Leber's congenital amaurosis, and enhanced S-cone syndrome. Thus, study of the rod photoreceptor cell has clinical importance.

Taken together, the mammalian retina allows for the study of *cis*-regulation in a well-characterized developmental environment where the tissue is highly accessible. In addition, the findings of this study and the development of future quantitative models of gene expression will have important implications for the treatment of human diseases.

Scope of Thesis Work

The central goal of my thesis is to develop a high-throughput multiplexed reporter assay that can be used to understand the rules governing *cis*-regulation in the mammalian retina. To address reporter assay problems related to low throughput construction and serial analysis, I set out to develop a technique where I could clone and measure thousands of reporter genes in a single experiment. Given the increasing accessibility of next-generation sequencing, I chose to read out the expression of the reporter gene constructs with RNA sequencing. Because every construct contains identical reporter gene sequences, I cloned a short barcode sequence into the 3'UTR of the reporter gene so I could differentiate between reporter genes by sequencing the barcoded transcript. I cloned a unique CRE upstream of the barcoded reporter gene to generate a library of constructs, so that each CRE controls the level of a particular barcoded reporter gene. I developed this technique in collaboration with Dr. Joseph Corbo's lab (Washington University in St. Louis School of Medicine) for use in explanted mammalian retinal cells, but this technique can be generalized to study the function of any regulatory elements in most cell types or tissues of interest.

In Chapter 2, I describe CRE-seq, the method we have developed for multiplexed reporter analysis and I use the technique to understand how every nucleotide in central portion of the Rhodopsin promoter contributes to the control of gene expression. In total, we measured the impact of 156 single nucleotide substitutions and found that the majority of these substitutions (86%) had a significant effect on the control of gene expression. We observed that the largest changes in expression were caused by those mutations located in annotated TFBS. In addition, we were able to test a subset of double combinations of variants (819 of 11,934). Comparison of expression driven by CREs with one and two variants allowed us to conclude that most (82%) combinations of variants generate complex expression patterns that are non-additive. These data suggest that sequence variants alter expression through binding site creation and nucleotide changes have the most effect in TFBS, as a result of cooperative and competitive interactions between regulating TFs. Taken together, we find that most substitutions have a significant impact on the control of expression and combinations of substitutions often control expression that is not predicted from single substitution expression levels.

In Chapter 3, I discuss my ongoing work to understand how combinations of TFBS control gene expression. To answer this question we constructed synthetic CREs from combinations of CRX and NRL TFBS and measured expression controlled by 1300 of these synthetic elements. This set includes all CREs composed of one, two, or three TFBS and a subset of the possible CREs composed of four TFBS. We find that synthetic elements composed of binding sites for both TFs drive the highest expression levels in this library, suggesting that CRX and NRL TFs may interact cooperatively. Surprisingly, we find that combinations of CRX binding sites control low levels of expression and can even repress expression below the promoter alone. From expression measurements made in CRX null retinas, we conclude that this repression is caused by CRX itself, rather than another TF that can recognize the CRX binding site. We also find that repression by CRX is dependent on the absence of NRL binding sites. Taken together, we observe complex patterns of gene expression driven by simple synthetic CREs and we hypothesize that TF cooperativity and dual regulation underlie these patterns.

In addition to the main focus of my thesis, Appendices 1 and 2 describe collaborative work to develop and apply this technique in new biological systems. In Appendix 1, I describe our work to develop CRE-seq in *S. cerevisiae* in order to interrogate the impact of TFBS affinity on transcriptional regulation. To accomplish this, we constructed synthetic promoters composed of TFBS with various affinities, used CRE-seq to measure the expression driven by these synthetic elements in the yeast genome, and modeled the resulting expression with a statistical thermodynamic model of gene expression. We found that a thermodynamic framework can accurately estimate the *in vivo* affinity of TFBS and that TFBS affinity can have unpredicted effects on control of gene expression due to altered affinity for cognate and non-cognate TF binding.

In Appendix 2, I describe my work to use CRE-seq to test ENCODE Consortium enhancer predictions (Dunham et al. 2012; Hoffman et al. 2013). As part of the ENCODE project the human genome was annotated with functional classes, including Enhancer, Weak Enhancer and Repressed predicted activities. We chose a subset of these predictions and tested for the ability of these sequences to drive gene expression in the K562 cell line. We found that ENCODE Enhancer and Weak Enhancer predictions drive expression that is different from negative controls, whereas Repressed and H1-hESC Enhancer sequences controlled expression that was indistinguishable from controls. Based on linear

regression models we found that the presence of TFBS motifs, rather than levels of histone modifications, were most predictive of active sequences. We concluded that ENCODE enhancer predictions identify active enhancers, but the TF binding preferences are most predictive of *cis*-regulatory activity.

This body of work makes several contributions to the field of transcriptional regulation. First, the development of CRE-seq allows for high throughput reporter gene analysis. This technique makes the study of *cis*-regulatory sequences considerably faster and easier than previous methods. Second, this study provides the first detailed examination of how sequence variants impact the activity of a mammalian promoter. In particular, the Rho promoter is relevant to human health and this study serves as a first step toward the interpretation of variation in this sequence. Finally, this study adds to our knowledge of how TFBS regulate expression, particularly in the mammalian rod photoreceptor cell.

CHAPTER 2: COMPLEX EFFECTS OF NUCLEOTIDE VARIANTS IN A MAMMALIAN *CIS*-REGULATORY ELEMENT

Jamie C. Kwasnieski,[1*] Ilaria Mogno,[1*] Connie A. Myers,[2] Joseph C. Corbo,[2] and Barak A. Cohen[1]

*These authors contributed equally to this work

[1]Department of Genetics, Washington University School of Medicine in St. Louis, 4444 Forest Park Parkway, St. Louis, MO 63108

[2]Department of Pathology and Immunology, Washington University School of Medicine in St. Louis, 660 South Euclid Avenue, St. Louis, MO 63110

This work was done in collaboration with Ilaria Mogno, Connie Myers, Joe Corbo, and Barak Cohen. Barak Cohen and I conceived this project. Barak Cohen, Joe Corbo, Ilaria Mogno, Connie Myers and I designed the experiments. Connie Myers carried out the retinal electroporation. Ilaria Mogno did the analysis of transcription factor binding sites. I carried out all other experiments and analysis. Barak Cohen, Ilaria Mogno, Joe Corbo and I wrote the paper. This chapter is a manuscript published in 2012 in the journal *PNAS*.

ABSTRACT

Cis-regulatory elements (CREs) control gene expression by recruiting transcription factors and other DNA binding proteins. We aim to understand how individual nucleotides contribute to the function of CREs. Here we introduce CRE-seq, a high-throughput method for producing and testing large numbers of reporter genes in mammalian cells. We used CRE-seq to assay over 1000 single and double nucleotide mutations in a 52-bp CRE in the Rhodopsin promoter that drives strong and specific expression in mammalian photoreceptors. We find that this particular CRE is remarkably complex. The majority (86%) of single nucleotide substitutions in this sequence exert significant effects on regulatory activity. While changes in the affinity of known transcription factor binding sites (TFBS) explain some of these expression changes, we present evidence for complex phenomena including binding site turnover and TF competition. Analysis of double mutants revealed complex, nucleotide-specific interactions between residues in different TFBS. We conclude that some mammalian CREs are finely tuned by evolution, and function through complex, non-additive interactions between bound TFs. CRE-seq will be an important tool to uncover the rules that govern these interactions.

INTRODUCTION

Mutations in *cis*-regulatory elements (CREs) often have unexpected effects on gene regulation. We lack models with the predictive power to accurately interpret the functional consequences of non-coding polymorphisms. More generally, we do not understand the nucleotide-level architecture that distinguishes true CREs from non-functional groupings of transcription factor binding sites. While consortium-driven efforts continue to predict that large numbers of mammalian sequences are CREs (Birney et al. 2007; Bernstein et al. 2010), we lack a corresponding high-throughput method for functionally analyzing the consequences of variants in these elements. Addressing these problems requires fine structure mutational analysis of mammalian CREs on a large scale, experiments which are difficult to perform using traditional assays. To facilitate such experiments we developed CRE-seq (***c**is*-**r**egulatory **e**lement analysis by **seq**uencing) a high-throughput reporter gene assay for mammalian cells.

CRE-seq leverages recent advances in oligonucleotide (oligo) synthesis (LeProust et al. 2010) and high-throughput sequencing (Mardis 2008). Using array based oligo synthesis we construct large numbers of reporter genes with unique sequence barcodes in their 3' UTRs. These libraries of barcoded reporter genes are then transfected, *en masse*, into mammalian cells and quantified by performing RNA-seq (Mortazavi et al. 2008) on the sequence barcodes. Here we present a study using CRE-seq to dissect a *cis*-regulatory element in mouse *Rhodopsin* (*Rho*), a gene that is expressed strongly and specifically in the mammalian retina.

Tight control of *Rho* expression is critical for the function of mammalian retinas (Olsson et al. 1992; Humphries et al. 1997). *Rho* expression is regulated in mice by multiple CREs located at varying distances from the transcription start site (Hsiau et al. 2007; Corbo et al. 2010). These elements are occupied *in vivo* by CRX, a retinal homeodomain TF (Corbo et al. 2010). One of these CREs, RhoCRE3, is located immediately upstream of the TSS and is sufficient to drive high levels of expression in rod photoreceptors (Lem et al. 1991; Zack et al. 1991). This element contains binding sites for CRX, and for NRL, a retina-specific basic-leucine-zipper protein (Chen and Zack 1996; Rehemtulla et al. 1996). How individual nucleotides within RhoCRE3 contribute to its function is not clear. To elucidate the functional architecture of RhoCRE3 we used CRE-seq to analyze the effects of more than 1000 variants of this element.

RESULTS

We developed CRE-seq, a method that parallelizes the construction and measurement of mammalian reporter genes. Using high-throughput oligo synthesis (LeProust et al. 2010) we created 1,040 variants of RhoCRE3, including all possible single nucleotide substitutions (156 mutations), and a large number of double substitutions (819 mutations). We also synthesized the wild type RhoCRE3 sequence 65 times to include as controls. Each mutant RhoCRE3 was synthesized adjacent to a unique 9 bp sequence barcode. To provide redundancy in the experiment, each single mutant was attached to ten different unique barcodes and each double mutant was attached to five unique barcodes. We then cloned the native *Rho* minimal promoter driving the fluorescent protein DsRed between the RhoCRE3 variants and their identifying barcodes. In the final plasmid library RhoCRE3 variants were located in a position immediately upstream of the TSS, surrounded by the same context sequence as in the endogenous *Rho* gene. The library contains 5,720 distinct reporter genes, each with a unique sequence barcode in the 3' UTR of the DsRed gene (Fig. 2.1A). We electroporated this library into explanted newborn mouse retinas (Hsiau et al. 2007). Using DsRed to monitor the progression of the experiment we confirmed that library expression was specific to rod photoreceptors (Fig. 2.1B), the cell type in which *Rho* is expressed.

We measured the expression of all library members using high throughput sequencing. After growing the electroporated retinal explants in culture we extracted both RNA and DNA, and sequenced the barcodes from both samples. The expression level of each barcoded reporter gene was calculated as the cDNA barcode sequence counts normalized to the DNA barcode counts (Table 2.S1). To test the reliability of CRE-seq, we compared the expression of twelve RhoCRE3 variants measured by both CRE-seq and by a standard fluorescence reporter assay (Lee et al. 2010). We observed strong correlation between both measurements (R^2=0.95, Fig. 2.1C), providing evidence that CRE-seq accurately quantifies gene expression in this system.

In a previously reported *in vitro* multiplex reporter gene assay (Patwardhan et al. 2009), sequence barcodes were placed adjacent to the transcription start site and had large effects on reporter gene expression. To ameliorate this problem we placed our barcodes in the 3' UTR of DsRed, far from the start site of transcription. To demonstrate that the barcodes do not interfere with our assay, we performed an experiment in which we fused more than 100,000 different barcodes to a single promoter, the wild-type

RhoCRE3, and measured the expression of this control library in electroporated retinas. If barcodes exerted a consistent effect on expression, we would expect to observe a correlation between replicates of this control library as barcodes with an activating effect would reproducibly increase expression while barcodes with a repressive effect would reproducibly decrease expression. No such correlation was found (R^2=0.04, Fig. 2.S1A), which suggests that the variation between members of the control library resulted from experimental noise rather than the barcode sequences. In contrast, the correlation between replicates of our library of RhoCRE3 mutants was high (R^2=0.95, Fig. 2.S1B, Table 2.S2). Taken together our results suggest that the 3' UTR barcodes have little effect in our assay, and that there are strong and reproducible differences between RhoCRE3 variants in our library.

Analysis of Single Mutants

We first analyzed expression data from single nucleotide substitutions in RhoCRE3. We found that 86% of single nucleotide substitutions significantly alter expression (Welch's t-test, p<0.05), with many substitutions causing large changes (>30-fold) in expression (Fig. 2.2A). For 98% (51/52) of the positions in RhoCRE3 at least one of the three possible substitutions had a significant effect on reporter gene expression. At 20% of positions different substitutions showed opposite effects on expression depending on the identity (A, G, C, or T) of the substituted base. Our results suggest that RhoCRE3 is an element whose function is finely tuned and that the majority of single nucleotide substitutions alter this function.

This finding differs dramatically from those of Melnikov et al. (Melnikov et al. 2012) and Patwardhan et al. (Patwardhan et al. 2012), two recent studies which use methods similar to CRE-seq to analyze mammalian enhancers. In these studies very few single nucleotide substitutions resulted in significant changes in expression. All single nucleotide changes in Melnikov et al. and Patwardhan et al. showed changes less than 3.5-fold, with the vast majority of effects less than 1.5-fold. In contrast, 49% of the single nucleotide substitutions we measured had effects between 3.5-fold and 30-fold. This sharp difference in results may reflect functional differences between the CREs in each study, or may be due to differences in the technologies employed by the different groups.

Single mutants with significant effects delineate five distinct regions across RhoCRE3 (Fig. 2.2A). Regions A and E contain known CRX sites and region D contains a known NRL site (Chen and Zack 1996; Rehemtulla et al. 1996). Regions B and C do not correspond to previously identified regulatory sequences in RhoCRE3. Regions of RhoCRE3 that show evolutionary conservation in mammalian lineages coincide with regions that have strong effects on expression (Fig. 2.2B). However, the modest correlation between effect size and conservation (R^2=0.31) shows that the degree of conservation does not quantitatively predict the effect size of individual mutations, though it does delineate functional regions.

Binding Site Mutations

We examined the behavior of mutations in regions A and E, the known CRX binding sites, as well as region D, the region that contains the known NRL binding site. We generated activity logos (Shin et al. 2000) for these regions using the single nucleotide substitution data (Figs. 2.S2A-C). Using the average log-likelihood ratio (ALLR) test (Wang and Stormo 2003) we found that the activity logos for regions A and D are similar (p<0.05) to Position Weight Matrices (PWMs) generated for these TFs in previous studies (Kataoka et al. 1994; Lee et al. 2010). Although there is a validated CRX site in region E, the activity logo made from region E does not show statistically significant similarity to the CRX PWM (p=0.22), but does show qualitative similarity. This finding suggests that the effect of mutations on expression do not depend solely on their effects on TF binding affinity, an issue we next addressed in more detail.

We examined the quantitative relationship between the predicted affinity of sequences for CRX or NRL and gene expression levels. Using a Position Weight Matrix (PWM) model of CRX specificity derived from quantitative *in vitro* gel shift assays (Lee et al. 2010) we computed the predicted change in affinity for each mutation in regions A and E for CRX. We then compared the predicted effect of each mutation on CRX affinity to its observed effect on expression (Fig. 2.2C). We performed a similar analysis for region D using a PWM that describes NRL binding specificity (Kataoka et al. 1994). For all three sites the wild-type sequence was predicted to have high relative affinity and drove high expression. However, while mutations that decrease predicted affinity tended to have lower expression, overall the correlation between predicted affinity and observed expression was modest (R^2 = 0.22 CRX(1); R^2 = 0.10 CRX(2); R^2

= 0.41 NRL). In all three regions several mutations that create sequences with predicted affinity as high as the wild-type sequence exhibit markedly decreased expression. Our data suggest that variables besides the affinities of CRX and NRL also help determine the *in vivo* activity of RhoCRE3 variants. Mutations in the binding sites for CRX and NRL may have effects on the binding of other TFs.

Regions A, D and E of RhoCRE3 contain binding sites for known transcriptional activators (CRX, NRL), and accordingly these regions contain dense clusters of single nucleotide substitutions that decrease expression (Fig. 2.2A). Because region C shows a similar pattern of variants with decreased expression, we hypothesize that it also represents a TFBS. Using the single mutant expression data we created an activity logo of this sequence (Fig. 2.S2D). The activity logo generated for region C does not match any of the sequence logos in the Transfac database (Wingender et al. 1996), which suggests that if these mutants effect the binding of a TF, it is a TF whose binding specificity has yet to be determined. However, the activity logo for the region overlapping with region B does show similarity to the CRX PWM ($p<0.05$, ALLR test), which suggests that many mutations in region B exert their effects by creating a CRX site (Fig. 2.S2E).

Region B is qualitatively different from the other regions of RhoCRE3 in that substitutions in region B often increase, rather than decrease expression. This finding suggests that mutations in region B either disrupt a binding site for a repressor or create new binding sites for activators. Because the density of mutations with large effects in this region is less than the density in regions with known TFBS, we favor the hypothesis that these mutations create binding sites for an activator. An obvious candidate for this activator is CRX.

To identify mutations that create CRX binding sites, we searched the entire library of single nucleotide mutants for matches to the CRX PWM. Mutations that create sequences with low predicted affinity to CRX are marked with an asterisk in Figure 2.2A. Surprisingly, mutations that create putative CRX sites, as defined by a PWM score threshold, are not distributed randomly throughout RhoCRE3; rather, these mutations cluster in regions B and D. This finding suggests that both regions B and D contain sequences that are very close to CRX binding sites. We analyzed these sequences in more detail.

Region B corresponds to a region of high evolutionary conservation in vertebrates. We scanned the orthologous regions from multiple vertebrate species with our CRX PWM and found that in most other mammals region B contains a sequence with a low affinity match to CRX PWM (Fig. 2.S3). The presence of this putative low affinity CRX site in most mammals accounts for the strong evolutionary conservation of region B. However, neither the rat nor the mouse genome has a potential CRX site in region B as defined by the same PWM score threshold. Using a neutral rate of 0.233 substitutions/site (Chun and Fay 2009) the three observed differences between mouse and rat in region B are not significantly different from the 2.33 that are expected under neutrality. This finding suggests that these two species are accumulating substitutions in region B at a rate consistent with neutral evolution, and that there has been recent turnover of CRX sites in region B in the rodent lineage. Taken together these results suggest that mutations in region B that create CRX sites increase expression by re-creating sequences that resemble the ancestral form of RhoCRE3.

In contrast to region B, all mutations that create sequences matching CRX sites in region D decrease expression. In region D these new hypothetical CRX sites disrupt the known NRL site. These overlapping CRX and NRL binding sites create complex expression patterns that suggest competition between the two factors. For example, mutations at position 36 suggest competition between NRL and CRX binding. All three substitutions at position 36 increase the predicted affinity of NRL, but only the two substitutions that also decrease the predicted affinity of CRX increase expression (Fig. 2.3). Only some mutations in region D support the competition model suggesting that other factors play a role in regulation of region D (Fig. 2.S4). Further experiments are needed to determine whether CRX and NRL actually compete for binding to the sequence in region D.

Other position specific effects on expression cannot be explained by CRX and NRL site affinity. For example, mutations at position 42 have a dramatic effect on expression but a relatively minor effect on predicted NRL or CRX affinity. Mutations at position 42 may create a new TFBS (Fig. 2.S2F) or alter the DNA helix structure (Fig. 2.S5) (Parker et al. 2009; Bishop et al.).

Analysis of Double Mutant CRE Variants

Our library also contained all possible double mutations between the NRL site in region D and the CRX site in region E. On average, the double mutant RhoCRE3 variants have lower expression than the single mutants (p<0.05, Wilcoxon). In addition, many (58%) double mutants have effect sizes that are larger than either of their component single mutations. We used linear modeling (Methods) to identify significant non-additive interactions between mutations, and if present, to determine the strength and direction of such interactions. Using ANOVA we found that 82% of double mutants have a significant interaction between positions that cannot be explained by an additive model of single mutant expression values (p<0.01, F-test). Our results contrast with those of Melnikov et al. (18) and Patwardhan et al. (19), who found very few interactions between mutations in enhancers. Our data show that some positions have interactions with many other positions, while other positions do not participate in any interactions (Fig. 2.4A). For some positions a particular substitution can have an interaction that is either more or less than additive, depending on the position of the second mutation (Fig. 2.4B).

In addition we observed that interactions between pairs of mutations are usually nucleotide-specific, rather than position specific. Most models are improved with the addition of some, but not all of the nine possible interaction terms (p<0.01, F-test). Mutations at two particular residues can combine to be either more or less than additive, depending on the specific identities of the substituted nucleotides (Fig. 2.4C). Combinations of mutations give rise to expression levels that cannot easily be predicted given the expression levels of single mutants. Physical interactions between NRL and CRX (Mitton et al. 2000) likely underlie some of the complexity we observe, but more detailed physical models will be required to unravel the molecular basis of these non-linear genetic interactions.

DISCUSSION

Using CRE-seq we found that single nucleotide substitutions in RhoCRE3 often result in large increases or decreases in expression, suggesting that this element may be highly constrained as to which mutations it tolerates through evolution, a hypothesis supported by the strong evolutionary conservation of this region. Changes in the predicted affinity of TFBS to known TFs partially explain the effects of these mutations, but cannot explain the full *in vivo* activity of substitutions. More complex phenomena such as

recent binding site turnover as well as competition between CRX and NRL may help explain the effects of some substitutions. The fact that the majority of double mutants have expression levels consistent with a non-additive model between positions suggests that interactions between CRX and NRL underlie some of the complex behaviors of RhoCRE3 mutants. Consistent with the many other studies, our results suggest that the interpretation of non-coding polymorphisms will require consideration of both low affinity sites and the creation of novel sites. Overall our results paint a picture of CREs as complex regulatory elements whose function can easily be altered by subtle changes to their nucleotide sequences.

In contrast, Melnikov et al. (18) and Patwardhan et al. (19) conclude that *cis*-regulatory elements are largely redundant at the nucleotide level, such that few mutations create large changes in expression. Our results showing large effects of single nucleotide substitutions agree with many detailed studies of single reporter genes, for example (Weiher et al. 1983; Myers et al. 1986; Goodbourn and Maniatis 1988; Singh et al. 1989; Smith et al. 1990; Shimell et al. 1994; Gurnett et al. 2007; Lettice et al. 2012; Yun et al. 2012). We also show evidence for extensive non-linear interactions between CRE mutations, while both Melnikov et al. (18) and Patwardhan et al. (19) conclude that mutations in enhancers act independently of each other. What might account for these dramatic differences in conclusions?

Some differences between these studies and our study are likely due to the CREs each group chose to examine. We chose to study an extensively validated CRE (Hsiau et al. 2007; Corbo et al. 2010) that drives the expression of *Rhodopsin,* the most highly expressed gene in the mammalian retina. This element shows strong evolutionary conservation throughout the vertebrate lineage and lies proximal to the start site of transcription. These features may make RhoCRE3 activity in CRE-seq susceptible to substitutions in ways that other *cis*-regulatory elements are not.

Technical differences between protocols probably contribute to some of the differences between these studies. Template switching during the PCR cycles used for library preparation can result in high rates of chimerism (Stemmer 1994), which can disrupt the unique correspondence between CRE variants and barcodes. Chimerism decreases both the accuracy and the dynamic range of the assay. In Melnikov et al. (18) and Patwardhan et al. (19) limited dynamic range due to the extensive use of PCR may have made it appear that enhancers are robust to mutation. In contrast, our method makes limited use of PCR amplification (Methods) to avoid the creation of chimeras. Finally, we chose to focus on only 1,040

mutations in a single experiment, which allowed us to obtain very high sequence coverage of our library. Both Melnikov et al. (18) and Patwardhan et al. (19) assayed much larger numbers of mutant promoters, making low sequence coverage and the resulting loss of statistical power a significant issue in these studies. In Patwardhan et al. (19) low sequence coverage of their library necessitated the use of an indirect measure of barcode counts as the metric of expression.

We have demonstrated that CRE-seq is a robust technology for assaying large numbers of CRE mutations. CRE-seq has a large dynamic range, nucleotide level resolution, and excellent reproducibility. Our analyses revealed a surprising amount of previously unknown complexity in RhoCRE3, an element that has been extensively studied. Our results demonstrate that nucleotides within CREs interact in complex, and often non-intuitive ways to produce regulated patterns of gene expression. We anticipate that CRE-seq will be an important tool for unraveling the rules that govern the function of CREs.

ACKNOWLEDGEMENTS

We thank Heather Lawson and Jim Cheverud for advice on statistical analyses; Sung Chun and Justin Fay for advice on models of molecular evolution; Aaron Spivak for assistance with TF binding analysis; Chris Fiore for discussions of CRE-seq technical issues; and Tom Tullius and Dana Zafiropoulos for help with DNA structure analysis. This work was supported by National Institutes of Health Grants GM092910 (to B.A.C.), EY018826 (to J.C.C.), and HG006346 (to B.A.C. and J.C.C.).

MATERIALS AND METHODS

CRE-seq Library Construction

To create the CRE-seq library plasmid backbone, we replaced the NotI site in plasmid Rho_minprox-DsRed (Hsiau et al. 2007; Corbo et al. 2010) with an EagI-XhoI-ClaI polylinker, creating plasmid pJK01. We then engineered sites for MfeI at position 25 and KasI at position 102 (following the numbering in (Hsiau et al. 2007; Corbo et al. 2010)), creating pJK02.

The wild type RhoCRE3 sequence spans positions 115,881,830-115,881,881 on mouse chromosome 6 (NCBI37/mm9). This region corresponds to the sequence between positions 68 and 120.

A pool of 5720 unique 150-mer oligonucleotides (oligos) was ordered through a limited licensing agreement with Agilent Technologies. Oligos were designed with the following structure: JKP1F, MfeI site, RhoCRE3 variant, KasI site, EcoRI site, EagI site, 9 bp barcode, ClaI site, and JKP1R (Table 2.S1). RhoCRE3 variants were designed between positions 68 and 120.

We amplified the oligo pool using 4 cycles of PCR with Phusion High-Fidelity polymerase (New England Biolabs, Beverley, MA) and primers JKP1F and JKP1R(Table 2.S1). We cloned the amplicon into pJK02 using MfeI and ClaI. We prepared DNA from 48,000 colonies to generate library PL5_1. We then cloned the minimal *Rho* promoter driving DsRed into PL5_1. A cassette containing the *Rho* minimal promoter fused to DsRed was amplified from pJK02 with primers JKP2F and JKP2R (Table 2.S1). The PCR amplicon was cloned into library PL5_1 using KasI and EagI, creating library PL5_2. To select for library members with full-length RhoCRE3 variants, we linearized PL5_2 with HindIII, gel purified the library, and re-circularized to create library PL5_3.

We created library (BL_1) to determine the effect of barcode sequences on reporter gene expression. Barcode sequence inserts, JKP5F and JKP5R (Table 2.S1), were annealed and cloned into pJK01 using EagI and ClaI. We prepared library DNA from 100,000 colonies.

Retinal Electroporation

Electroporations and explant cultures were performed as previously described (Hsiau et al. 2007) using 0.5 µg/ml of library PL5_3 as well as 0.5 µg/ml *Rho* minimal promoter driving GFP for visualization

of electroporation efficiency (Lee et al. 2010) . After 8 days in culture retinas were washed twice in sterile HBSS (Gibco, Life Technologies, Grand Island, NY), and total RNA and DNA were extracted using Trizol according to manufacturer's instructions (Ambion, Life Technologies, Grand Island, NY).

Comparison Between a Fluorescence Assay and CRE-seq

Twelve previously characterized RhoCRE3 variants (Lee et al. 2010; Montana et al. 2011b) were excised from their respective vectors with XbaI and KpnI and cloned into the barcode library BL_1. For each variant we isolated ten uniquely barcoded constructs. We mixed all 120 constructs together for CRE-seq analysis.

Preparing Samples for RNA-Seq

Isolated RNA was treated with DNAseI (Ambion) to eliminate potential genomic DNA contamination. The first strand of cDNA was synthesized with Invitrogen Superscript II reverse transcriptase using oligo-dT primers. After first strand synthesis, the reaction was treated with RNAse H (NEB) to remove RNA. The 3' UTR of the DsRed gene including the barcode sequence, was amplified from both cDNA and isolated plasmid DNA samples. Amplification (98°C for 1min, 21 cycles: 98°C for 10s, 58°C for 30s, 72°C for 30s, and 72°C for 5 min, NEB HF Phusion MM) of the cDNA and plasmid DNA samples isolated by PCR using primers JKP3F and JKP3R (Table 2.S1) yielded barcode sequences that are flanked with EagI and EcoRI restriction enzyme sites. The products were purified with Qiagen QIAquick PCR Purification kit and digested with EagI and EcoRI. Illumina adapter sequences were ligated onto these overhangs. This product was amplified (98°C for 1 min, 21 cycles: 98°C for 30s, 65°C for 30s, 72°C for 30s, and 72°C for 5 min) with HF Phusion MM using primers JKP4F and JKP4R (Table 2.S1) to enrich for molecules that contain both adapter sequences.

To measure expression of the barcoded library, electroporation replicates were multiplexed and run on two lanes of an Illumina HiSeq machine which generated 48.2 million sequence reads corresponding to cDNA and 48.8 million reads corresponding to DNA. Sequencing reads that matched the first 20 nucleotides of designed sequence were counted, regardless of quality score. Only barcodes with more than 10 reads in either cDNA or DNA pools were used for analysis.

Determination of Expression Levels of *Rho* Variants

We determined the expression levels of each RhoCRE3 variant by computing the average cDNA/DNA ratio for all barcodes that marked the same RhoCRE3 variant (Table 2.S2). We also computed the standard error of the mean (SEM) for each average in each replicate. We then averaged the averages from the three replicates and propagated the standard error using the formula:

$$SE\left(\frac{\sum E_n}{n}\right) = \sqrt{\sum\left(\frac{SE(E_n)}{n}\right)^2}$$

The propagation of standard error when computing the ratio of mutant to wild-type expression levels was calculated as:

$$SE\left(\frac{E_{mut}}{E_{wt}}\right) = \sqrt{\frac{{E_{mut}}^2}{{E_{wt}}^2}\left[\left(\frac{SE(E_{mut})^2}{{E_{mut}}^2}\right) + \left(\frac{SE(E_{wt})^2}{{E_{wt}}^2}\right)\right]}$$

Analysis of Binding Site Mutations

The CRX position weight matrix was derived from quantitative binding affinity data (Lee et al. 2010) according to (Stormo and Fields 1998). Similarly, a NRL position weight matrix was derived from sequence data from (Kataoka et al. 1994). The two TF matrices were scored against every position of each CRE variant using *patser (Stormo et al. 1982)*.

Non-additive Model for Variants with Two Substitutions

Linear regression was used to determine the extent of non-additive expression in the variants with two substitutions. First, the following two models were generated for each combination of positions and compared using an ANOVA ($p<0.01$).

$$E_{1,2} = WT + \beta_1 mut_1 + \beta_2 mut_2$$

$$E_{1,2} = WT + \beta_1 mut_1 + \beta_2 mut_2 + \beta_3 mut_{1,2}$$

Models that were significantly improved (82%, 75/91) by the addition of the interaction term were further analyzed for base-specific interactions. After selecting positions with significant non-additive

expression, we used ANOVA ($p<0.01$) to test each of the nine possible base-specific interactions between positions.

All experiments were conducted in accordance with the Guide for the Care and Use of Laboratory Animals and the Animal Welfare Act and were approved by the Washington University in St. Louis Institutional Care and Use Committee.

FIGURE 2.1

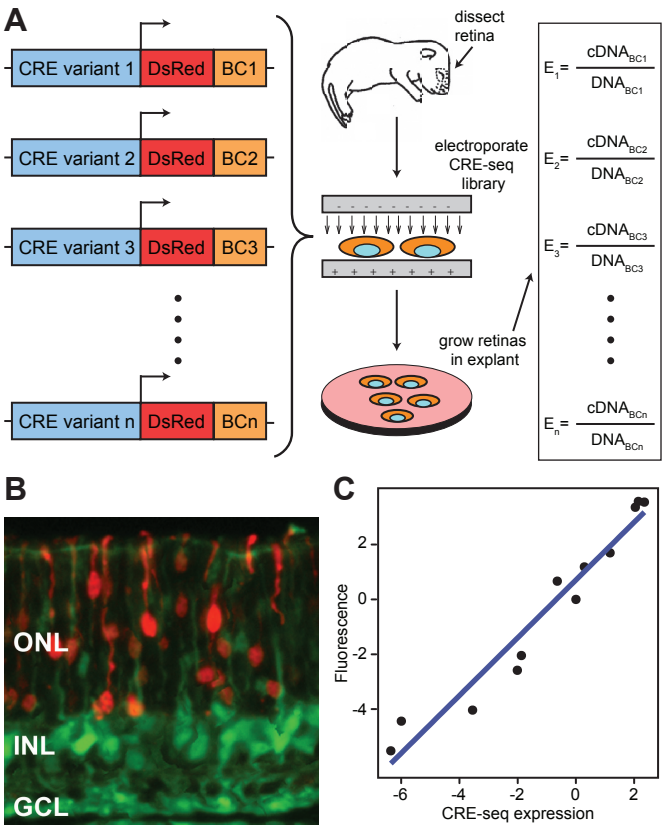


Figure 2.1 (**A**) Schematic of the CRE-seq method. Each CRE variant in the library is fused to a reporter gene marked by a unique DNA barcode in the 3' UTR of DsRed. The library is electroporated into newborn mouse retinas, which are cultured as explants for eight days. Barcodes are then sequenced from harvested mRNA and DNA. The cDNA/DNA ratio of each barcode is a quantitative measure of the expression levels driven by each CRE variant in the library. (**B**) Cell type specificity of the CRE-seq library. A cross section of an electroporated retina shows DsRed expression in rod photoreceptors residing in the outer nuclear layer (ONL). Green fluorescence driven by a ubiquitously expressing CAG-GFP construct is observed in all layers including the inner nuclear layer (INL) and the granular cell layer (GCL). (**C**) Correlation between CRE-seq and fluorescent reporter genes. Twelve RhoCRE3 variants were quantified using a standard dual color fluorescent reporter gene assay (Lee et al. 2010; Montana et al. 2011b) and also using CRE-seq. X-axis, CRE-seq expression, $\log_2$(variant RhoCRE3/wild-type RhoCRE3); Y-axis, fluorescent reporter gene expression, $\log_2$(fluorescence of variant RhoCRE3/fluorescence of wild-type RhoCRE3).

FIGURE 2.2

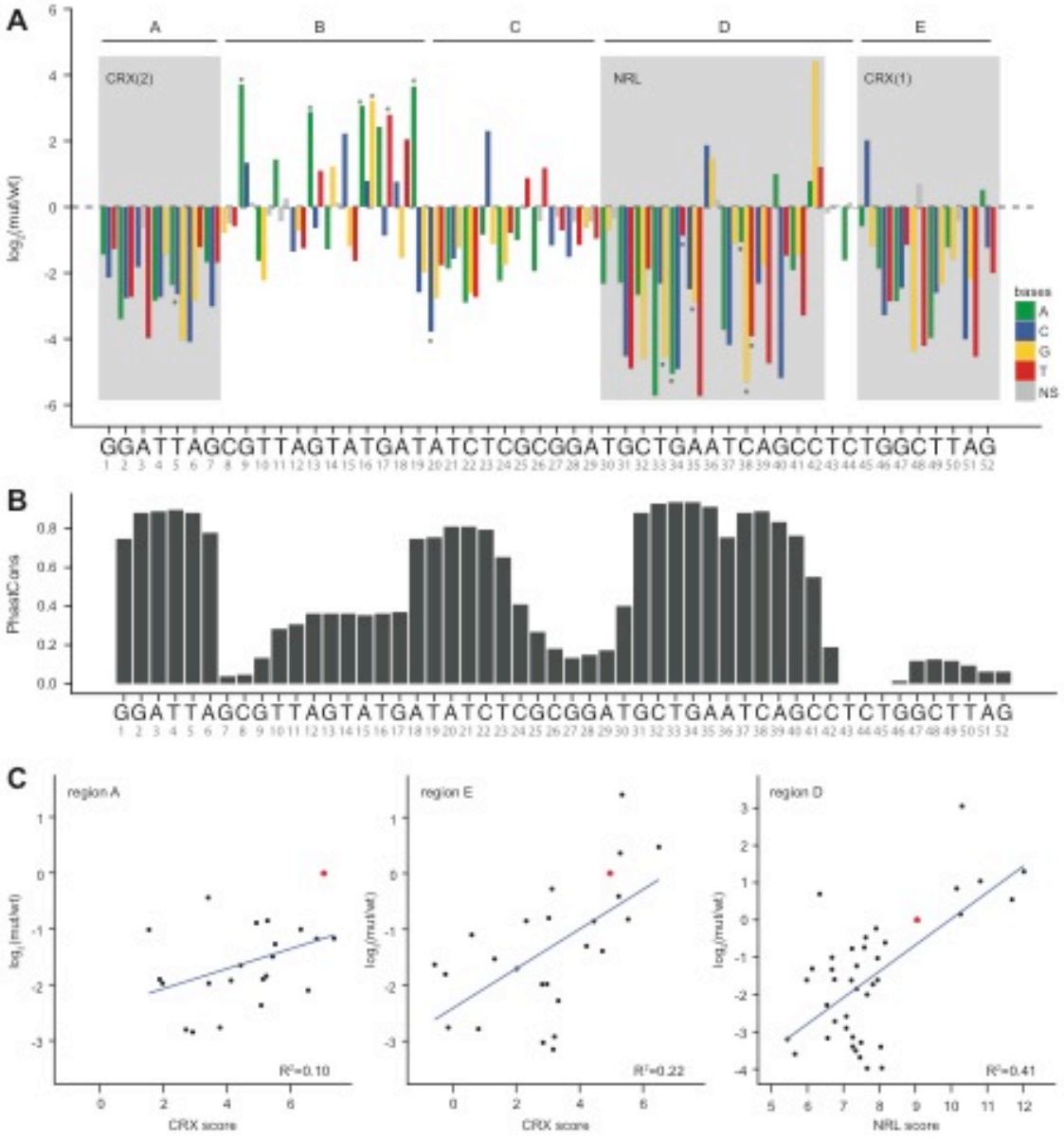


Figure 2.2 (**A**) Effects of single nucleotide mutations on reporter expression. X-axis, nucleotide position in RhoCRE3; Y-axis, relative expression by CRE-seq quantified as the $\log_2$(variant RhoCRE3/wild-type RhoCRE3). Colored bars represent substitutions whose expression was significantly different from wild-type (Welch's t-test, $p<0.05$). Gray bars represent substitutions that are not significantly different from wild-type. Each position has three bars representing the three possible substitutions at that position. For analysis we divided RhoCRE3 into five regions (A-E). Experimentally validated binding sites for CRX and NRL are shaded in gray. Asterisks mark the locations of substitutions that create new CRX binding sites (≥2% predicted affinity relative to consensus). (**B**) Phylogenetic conservation of nucleotides in RhoCRE3. X-axis, position in RhoCRE3; Y-axis PhastCons score derived from a multiple alignment of twenty-eight mammalian sequences (Miller et al. 2007). (**C**) Relationship between the effects of mutations on the predicted binding affinity of TFs versus the mutations' effects on gene expression. X-axis, predicted affinity of RhoCRE3 mutations for CRX or NRL; Y-axis, observed effects of mutations on gene expression. In each graph the red dot represents the wild type sequence of RhoCRE3 and the blue line represents the linear regression model.

FIGURE 2.3

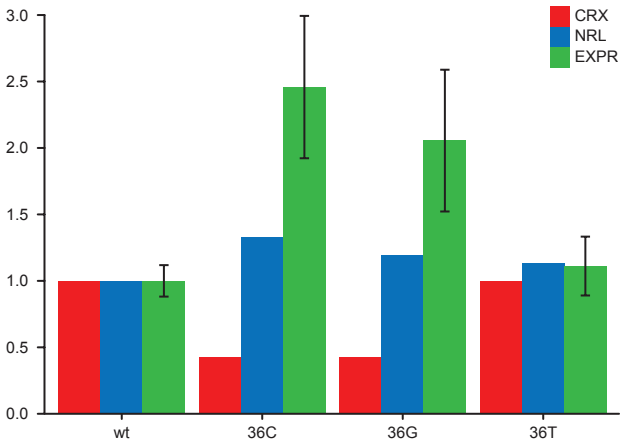


Figure 2.3 Complex interactions between CRX and NRL in region D. The effects of mutations at position 36 support a model of competition between NRL and CRX. The X-axis shows the identity of the base at position 36. Red bars, relative affinity of CRX normalized to the wild type sequence. Blue bars, relative affinity of NRL normalized to the wild type sequence. Green bars, expression normalized to the wild type sequence. Errors bars represent the standard error of the mean.

FIGURE 2.4

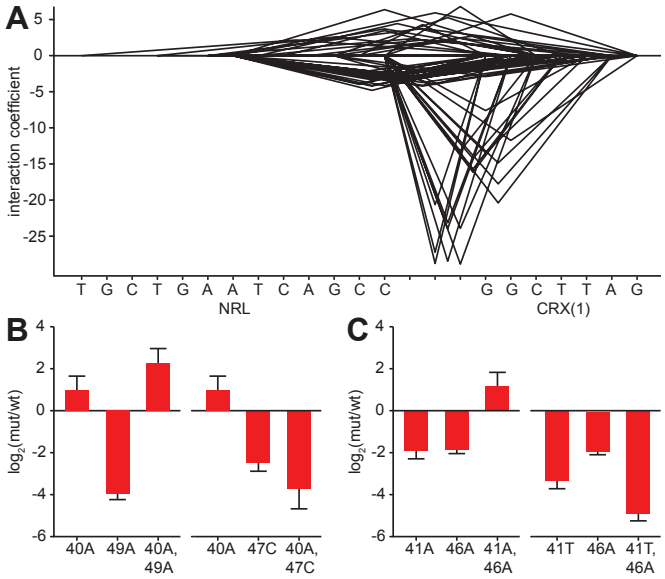


Figure 2.4 Interactions between mutations in TFBS within regions D and E. (**A**) Pattern of interactions. Each arc connects two nucleotides that have a significant interaction term by ANOVA (Methods). The height of the arc represents the magnitude of the interaction term. Only interactions greater than 3 or less than -3 are shown. (**B**) The magnitude and direction of interactions are position specific. The interaction between mutations 40A (CRX) and 49A (NRL) is significantly more than additive and greater than the effect of either single mutant, while the interaction between mutations 40A and 47C (NRL) is less than additive and greater than the effect of either single mutant. (**C**) Interactions between the same positions are base specific. The interaction between 41A (CRX) and 46A (NRL) is significantly more than additive, while the interaction between 41T and 46A is less than additive.

SUPPLEMENTARY FIGURES AND TABLES

TABLE 2.S1

Name	Primer sequence
JKP1F	TGCAAGCCAATTGGGCCCCGGT
JKP1R	GGAGTGCGGCCATCGATGCG
JKP2F	ATATCTCGCGGATGCTGAAT
JKP2R	TATTTGTGAGCCAGGGCATT
JKP3F	TGTTCCTGTAGCGGCCGACG
JKP3R	GAATTCTAGCCAGAAGTCAGATGCTCAAG
JKP4F	AATGATACGGCGACCACCGAG
JKP4R	CAAGCAGAAGACGGCATACGA
JKP5F	GGCCGACGNNNNNNNNNNCGCAT
JKP5R	CGATGCGNNNNNNNNNNCGTC

Table 2.S1 Primer sequences used in this study.

FIGURE 2.S1

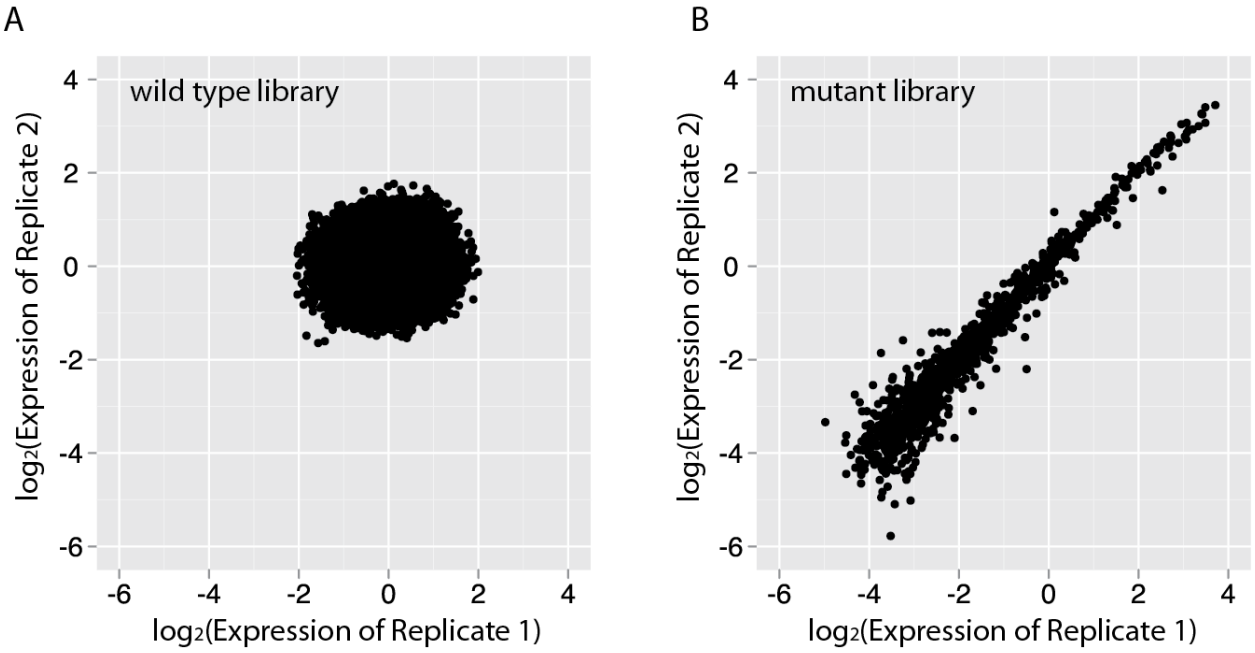


Figure 2.S1 Correlation between replicate experiments. (**A**) Barcodes do not impact expression. More than 100,000 distinct barcodes were fused to the wild-type RhoCRE3 sequence. Measurements in replicate experiments showed no correlation (R^2=0.04) indicating that different barcodes do not have significant effects on reporter gene expression. (**B**) Mutant RhoCRE3 variants have strong effects on expression. The correlation between replicate experiments measuring the 1,040 RhoCRE3 variants is high (R^2=0.95) indicating that the mutations have strong and reproducible effects on expression.

FIGURE 2.S2

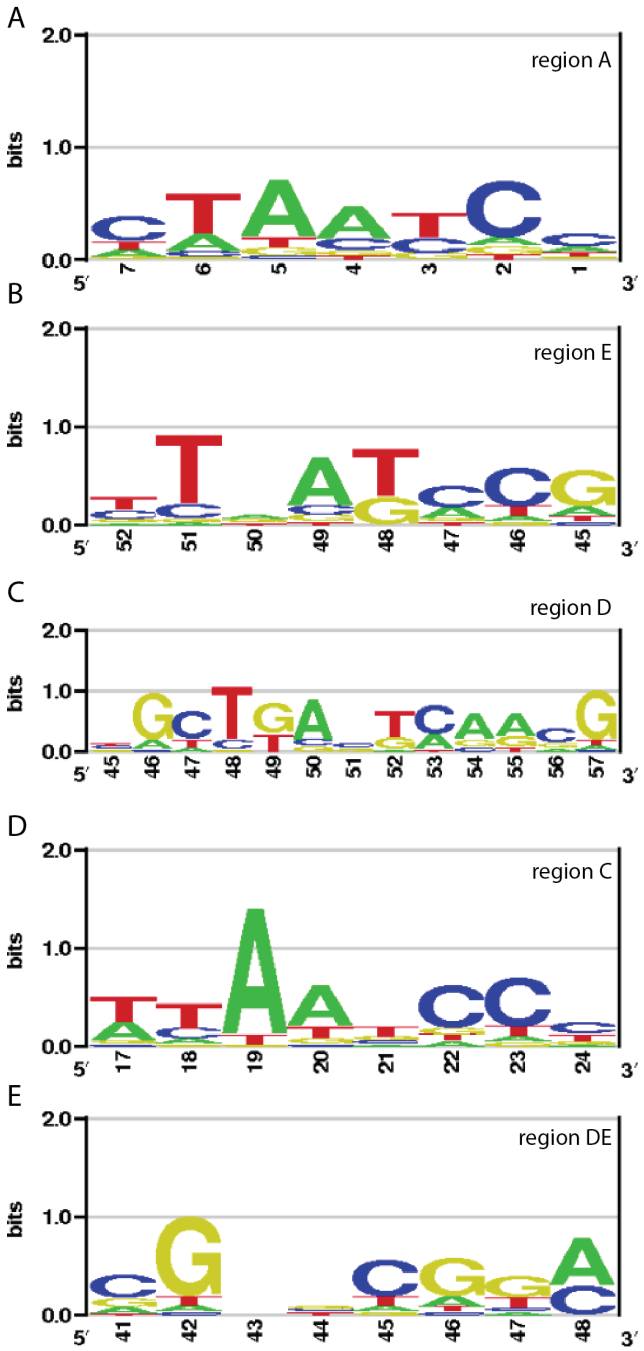


Figure 2.S2 Activity logos of known and hypothesized transcription factor binding sites. The activity logos of (**A**) region A and (**B**) region E correspond to the known sequence logo for CRX (depicted here in 'reverse' orientation). (**C**) The activity logo of region D corresponds to the known sequence logo for NRL (Kataoka et al. 1994). (**D**) Activity logo of region C. (**E**) Activity logo of the sequence joining regions D and E.

FIGURE 2.S3

Figure 2.S3 RhoCRE3 sequence in mammals. The sequence of the RhoCRE3 element in 21 mammals and the platypus is shown. Red, sequences that match the CRX position weight matrix model. Blue, sequences that match the NRL position weight matrix model. Many of the species have a low affinity CRX site directly adjacent to the first known CRX site.

FIGURE 2.S4

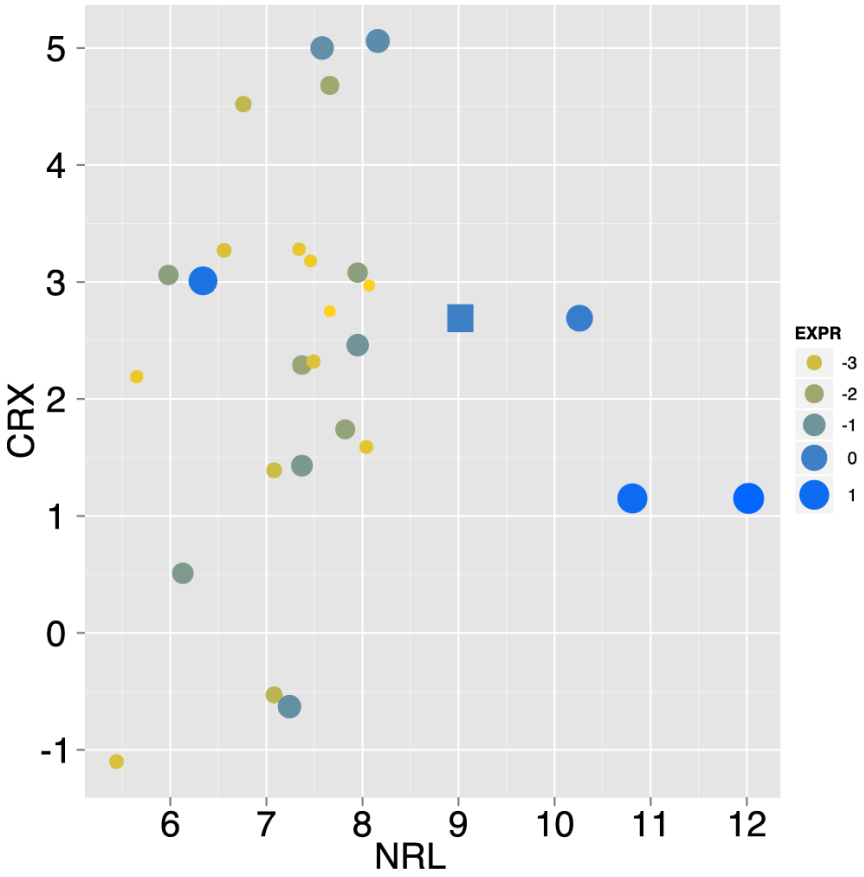


Figure 2.S4 The effects of mutations in region D. For each mutation in region D we plotted its effect on the predicted affinity of the NRL site (X-axis) and the CRX site (Y-axis). Each circle represents a single mutation in region D. The size and color of the circles show the effect on reporter gene expression. The square represents the wild-type sequence.

FIGURE 2.S5

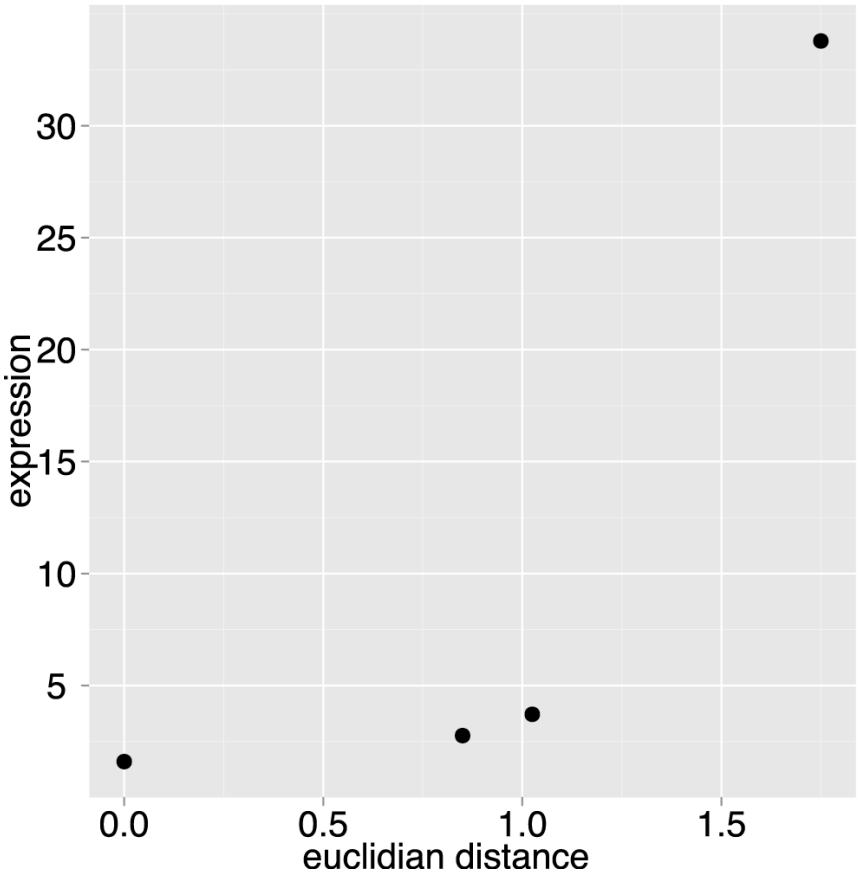


Figure 2.S5 Structural analysis of mutations at position 42. We used the program ORChID2 (Bishop et al.) to generate structural plots of the RhoCRE3 sequence between positions 30 and 52. We generated these plots for the wild-type sequence as well as the sequence containing the three possible substitutions at position 42. We computed the Euclidean distance between each sequence and the wild-type (x-axis) and compared these distances to the expression level driven by each variant (y-axis). Mutant 42G, which has the largest effect on expression, also has the largest predicted effect on DNA structure.

CHAPTER 3: COMBINATIONS OF CRX AND NRL BINDING SITES GENERATE DIVERSE EXPRESSION LEVELS

Jamie C. Kwasnieski,[1] Connie A. Myers,[2] Joseph C. Corbo,[2] and Barak A. Cohen[1]

[1]Department of Genetics, Washington University School of Medicine in St. Louis, 4444 Forest Park Parkway, St. Louis, MO 63108

[2]Department of Pathology and Immunology, Washington University School of Medicine in St. Louis, 660 South Euclid Avenue, St. Louis, MO 63110

This work was done in collaboration with Connie Myers, Joe Corbo, and Barak Cohen. Barak Cohen, Joe Corbo, and I conceived this project. Barak Cohen and I designed the experiments. Connie Myers carried out the retinal electroporation. I carried out all other experiments and analysis and I wrote the paper.

ABSTRACT

The ability to predict gene expression from sequence is an unsolved problem in the field of the genetics. Successful prediction of gene expression levels will aid in the interpretation of human non-coding sequence variation and allow us to engineer novel regulatory elements with desired expression patterns. There are a variety of methods to identify functional transcription factor binding sites (TFBS), but little understanding of how these binding sites combine to control gene expression. In the developing mammalian retina, transcriptional activators CRX and NRL regulate rod photoreceptor cell differentiation and homeostasis. In order to determine the rules that govern combinatorial regulation by CRX and NRL, we have examined the expression controlled by synthetic promoters composed of NRL and CRX TFBSs. We have used *cis*-regulatory element sequencing (CRE-seq), a high throughput reporter assay, to measure the expression controlled by these synthetic *cis*-regulatory elements (CREs) in explanted retinal cells. We find there is a strong synergistic interaction between NRL and CRX binding sites and that CRX binding sites can switch from activating to repressing in different contexts. These results have major implications for our ability to interpret *cis*-regulatory sequence and will improve our ability to predict expression from genomic sequence.

INTRODUCTION

cis-regulatory elements (CREs) are DNA sequences that regulate the expression of a nearby gene. CREs are composed of combinations of short transcription factor binding sites (TFBS), which are recognized by regulating transcription factors (TFs). TFs bind these sites and recruit RNA Polymerase II (RNA Pol II) machinery to initiate transcription. Transcriptional activators promote the recruitment of the polymerase machinery, whereas transcriptional repressors inhibit this recruitment. A major goal in the field of transcriptional regulation is to develop a quantitative understanding of how functional combinations of binding sites drive gene expression.

There are a variety of methods to identify functional TFBS in a genome: directly measuring TF binding with chromatin immunoprecipitation (ChIP) (Harbison et al. 2004; Johnson et al. 2007; Visel et al. 2009); identifying clusters of TFBS in co-expressed sequences (Beer and Tavazoie 2004; Elemento et al. 2007); and transcriptional profiling of cells mutant for the TF of interest (Hu et al. 2007). These techniques provide insight into the genomic locations of functional TFBSs and are a first step in classifying combinations of binding sites that regulate a specific gene. The identification of binding sites alone does not provide information about how sequences control the level of gene expression, or how this expression might change as a result of removing, adding, or rearranging these binding sites. In order to understand how these variables impact the control of expression, one needs to examine how systematic combinations of TFBS regulate gene expression levels.

Generally speaking, scientists understand how combinations of binding sites control gene expression, but the field lacks a quantitative model to predict how much expression is produced by a sequence. In fact, we have little ability to predict the level of expression that is controlled by even a single DNA sequence. It would be useful to develop a quantitative model to predict expression for two reasons. First, the model would predict how gene expression level is modulated as a result of sequence perturbation. This would be useful in the interpretation of human non-coding variation, which can change affinity or number of regulating binding sites. Second, this model would predict the level of gene expression controlled by a synthetic DNA sequence. This would be useful when engineering a regulatory sequence to express at a desired level. In order to develop quantitative models, we first need to deepen our understanding of how DNA sequence properties impact expression regulation.

Our approach is to gain an understanding of how simple CRE sequences control gene expression. In this study we examine gene expression controlled by CREs composed of only four different TFBS. This design allows us to characterize the relationship between expression controlled by a CRE and the properties of the TFBS in that CRE sequence. We vary TFBS identity, number, affinity, order and orientation but keep other variables constant, such as TFBS spacing and surrounding sequence. We anticipate that patterns found in the study of regulation by simple CREs will be relevant when predicting expression driven by complex regulatory elements that are composed of the same TFBS (Gotea et al. 2010).

The mammalian retina is an ideal system for *cis*-regulatory analysis because it is an accessible tissue with well-characterized transcriptional regulators. The majority of cells in the mammalian retina are photoreceptors that express cascades of TFs to define cell identity during differentiation. Cone/Rod homeobox transcriptional activator CRX is expressed in the early stages of photoreceptor differentiation (Furukawa et al. 1999). In rod photoreceptor cells CRX activates expression of neural retina leucine zipper transcriptional activator NRL (Kautzmann et al. 2011; Montana et al. 2011a). CRX and NRL combine to transcriptionally regulate the level of photoreceptor genes that are critical for rod photoreceptor cell differentiation and homeostasis (Furukawa et al. 1999; Mears et al. 2001; Hsiau et al. 2007). CRX and NRL have been shown to directly interact and regulate synergistically (Chen et al. 1997; Mitton et al. 2000), suggesting that the two TFs may cooperatively regulate expression. We anticipate that this study will not only inform our understanding of rod cell-specific gene regulation, but will also provide use as a general model of mammalian gene regulation.

RESULTS

CRE-seq Library and Measurements

In order to understand how combinations of CRX and NRL transcription factor binding sites combine to regulate gene expression, we created a library of synthetic CREs (Ligr et al. 2006; Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010; Sharon et al. 2012; Smith et al. 2013). These synthetic CREs are composed of: a high affinity CRX_1 site, a low affinity CRX_2 site, a predicted lower affinity CRX_3

site (Lee et al. 2010), and a high affinity NRL site (Kataoka et al. 1994). Previous studies have shown that CRX binds the CRX_1 and CRX_2 binding sites *in vitro* (Lee et al. 2010) and that regulation of these binding sites in the Rhodopsin (Rho) promoter is critical for appropriate expression (Kwasnieski et al. 2012). In addition, we hypothesize that CRX can bind the CRX_3 site based on evidence of TF competition (Kwasnieski et al. 2012) even though CRX_3 has very low predicted affinity. Short constant buffer sequences were added to each binding site to maintain the helical spacing when binding sites were combined. Using these 4 TFBS, we generated all possible synthetic CREs that contain one, two, or three binding sites in both orientations. In addition, we generated 715 of the possible 4096 synthetic elements that are composed of four binding sites.

We used CRE-seq, a high-throughput multiplexed reporter assay, to measure the expression controlled by these synthetic CREs (Kwasnieski et al. 2012). The library of CREs were cloned upstream of the rod cell-specific Rho proximal promoter driving the DsRed reporter gene. The expression of each synthetic CRE was measured by five unique barcodes to average any barcode-specific effects on expression. The library was electroporated into retinal explants and RNA was harvested after 8 days in culture. Expression was calculated as the number of times each barcode was sequenced from the RNA pool normalized to the number of times each barcode was sequenced from the original plasmid pool. We conducted four biological replicates of library expression (Figure 3.1A, R^2 range=0.97 – 0.99).

Synthetic CREs Control Diverse Expression Levels

The library of synthetic CREs was composed of TFBS for transcriptional activators. We anticipated that expression controlled by long CREs would be higher than expression controlled by short CREs because we thought more activator binding sites would result in higher expression. Surprisingly, we observe that for every length CRE, some synthetic elements drive expression that is activated above the Rho proximal promoter, while other synthetic elements drive expression that is repressed below the Rho proximal promoter (Figure 3.1B). The range in expression levels is surprising given that the library was constructed with binding sites for transcriptional activators that are not predicted to repress gene activity. These data show that even simple combinations of TFBS can generate complex gene expression patterns that are not the additive expression of component parts.

Patterns in Regulation by CRX and NRL

In order to understand how these TFBS could generate such a large range in expression values, we sought to identify the combinations of motifs that controlled high expression. Upon examination of synthetic CREs that are two TFBS in length, we find that CREs composed of CRX_1 and NRL binding sites have significantly higher expression than CREs composed of only CRX_1 or only NRL TFBS (Welch's t test, $P<0.01$, Figure 3.2A). We observe the same relationship when we examine synthetic CREs composed of three TFBSs (Welch's t test, $P<0.01$, Figure 3.2B). These data suggest that CRX and NRL synergistically regulate expression because combinations of CRX_1 and NRL binding sites drive significantly higher expression than CRX_1 or NRL binding sites alone. This result is supported by previous studies demonstrating that CRX and NRL synergistically regulate (Chen et al. 1997) and directly interact (Mitton et al. 2000).

We sought to identify combinations of TFBS that control expression levels lower than the expression of the proximal promoter alone. By examining the lowest expressing CREs, we found that synthetic CREs composed of multiple CRX binding sites and few NRL binding sites drive expression that is repressed below the basal promoter alone (Figure 3.3A). This observation is not simply the absence of the CRX and NRL synergism because the addition of NRL binding sites increases expression level but does not relieve the trend that adding more CRX_1 sites lowers expression (Figure 3.3B). Interestingly, we observe similar behavior for NRL binding sites where many NRL TFBS drive low expression (Figure 3.S1), but these synthetic CREs are unable to repress the proximal promoter. There is little evidence to support CRX regulating as both an activator and a repressor (White et al. 2013). Previous studies (Smith et al. 2013) have observed low expression driven by homopolymers of binding sites and that homopolymers are capable of initiating heterochromatin formation (Bulut-Karslioglu et al. 2012).

Repression is Dependent on CRX

To determine whether CRX itself was regulating at the CRX binding site, or if another TF with similar specificity was responsible for this repression, we measured expression of the library in CRX-null retinas. We previously used CD wild-type mice for our experiments but the CRX null mouse was created

in a B6 background, so we measured expression of the library in the B6 wild-type mouse to retina ensure a proper wild-type control for the CRX null mouse (Figure 3.S2). High correlation (R^2=0.96) between the wild-type expression measurements suggests that the genetic variation between the two mouse strain backgrounds does not impact library expression.

Expression measurements from CRX null retinas have two striking patterns when compared to wild-type expression. First, we observe a population of synthetic elements that has relatively consistent expression between the wild-type and null mice and thus lies near the diagonal when wild-type and null expression measurements are compared (Figure 3.4). It has been suggested that the CRX null mouse has a milder phenotype than expected in part because Otx2, or another homeodomain transcription factor, is capable of compensating for CRX in the CRX null retina (Bovolenta et al. 1997; Furukawa et al. 1999; Hennig et al. 2008). This regulatory compensation would explain why many synthetic CREs are still expressed in the null retina. Second, we observe a population of synthetic elements that drive low expression in the wild-type retina but increased expression in the CRX null retina. This population is largely composed of synthetic CREs that contain three or four CRX binding sites (Figure 3.4). These data suggest that the repression we observe in the wild-type mouse is driven by CRX, rather than another homeodomain TF, because we see expression increases when CRX is not present. These data also suggest that CRX repression is dependent on the absence of NRL TFBS because we see that synthetic CREs composed of NRL TFBS are consistently expressed between the wild-type and null conditions.

DISCUSSION

In this study we found that combinations of transcriptional activator binding sites generate diverse expression levels, where synergism between CRX and NRL binding sites drives high expression and CRX-dependent repression drives low expression. We were surprised to see that combinations of CRX binding sites not only drive low expression, but can also repress the Rho proximal promoter itself. We have developed two working models that might explain this observation.

In our first model, we hypothesize that the proximal promoter needs to be occupied by NRL in order to activate transcription. In this model, interactions between CRX bound in the CRE and NRL bound in the proximal promoter would lower the NRL promoter occupancy and decrease the promoter's ability to

recruit polymerase. This model predicts that if only NRL is required at the proximal promoter, rather than both CRX and NRL, CRX binding sites will not be able to repress the Rho proximal promoter with a mutant NRL binding site. This model also predicts that CRX binding sites will not repress a promoter that is not regulated by NRL. Previous studies have found that CRX has very little trans-activation potency but NRL can interact with TATA-binding protein (Chen et al. 1997; Friedman et al. 2004), which adds molecular support to this model.

In our second model, we hypothesize that there is some appropriate ratio of bound CRX to bound NRL that is necessary for expression activation. In this model, the presence of many CRX TFBS increases the occupancy of CRX, disrupting this stoichiometric relationship, and gene activity is repressed. This model predicts that synthetic CREs composed of many CRX TFBS would not repress if the appropriate number of NRL TFBS were included. These two models represent our ideas for how CRX TFBS would be able to repress the proximal promoter activity; we are conducting experiments and developing computational models to distinguish between these two scenarios.

Previously, we have used a statistical thermodynamic framework to develop quantitative models of gene expression (Shea and Ackers 1985; Buchler et al. 2003; Gertz et al. 2009). This framework is powerful because it predicts expression with parameters that represent biophysical interactions, allowing us to gauge whether our understanding of the system is able to mathematically predict the behavior. In this study, we attempted to apply the thermodynamic framework in order to distinguish between the previously described working models. Given the our current data set, this framework is unable to capture the behavior of this library (data not shown) because the CRX binding sites have the ability to at as activating or repressing sites depending on the context. This framework assumes every TFBS behaves in a consistent regulatory manner. In order to model this behavior we would need to add major assumptions to the thermodynamic framework, and before we add these assumptions we need more information about the mechanism of CRX repression.

In order to constrain quantitative models and to understand the TFBS grammar that underlies these complex expression patterns, we are conducting several experiments and analyses. First, we are measuring expression controlled by a library of synthetic CREs upstream of the Hsp68 promoter, rather than the Rho proximal promoter. This experiment will tell us if homopolymers of CRX TFBS can repress

an unrelated promoter element. Second, we are measuring the expression controlled by a library of synthetic CREs upstream of a mutated Rho proximal promoter, rather than the wild-type Rho proximal promoter. This mutant Rho promoter contains two mutations in the annotated NRL binding site. These expression measurements will tell us if CRX repression depends on NRL binding in the proximal promoter. Finally, we are adapting ChIP to our synthetic CREs in order to measure the occupancy of CRX on the synthetic elements. This experiment will tell us if CRX has higher occupancy when there are more CRX binding sites and whether the presence of NRL binding sites increases CRX occupancy.

We conclude that combinations of TFBS for transcriptional activators can generate diverse gene expression patterns and even repress the activity of the proximal promoter alone. We find that combinations of CRX and NRL sites drive high expression, suggesting a synergistic interaction between the two regulators. We also conclude that CRX is capable of acting as a repressor or an activator, depending on the sequence context. CRX sites repress when clustered with other CRX TFBS, but CRX sites activate when NRL sites are present. In general, interactions between TFs and the dual regulatory capacity of TFs can underlie these complex expression patterns. Further genetic and biochemical experiments, paired with the development of quantitative models, will aid in the understanding of the sequence features that make this possible.

MATERIALS AND METHODS

Synthetic CRE Library Design

Synthetic CREs are composed of three CRX binding site variants and one NRL binding site. Short constant buffer sequences were added to each binding site to maintain the helical spacing when binding sites were combined (Table 3.1). Using these 4 TFBS sequences, we computationally generated all possible synthetic CREs that contain one, two, or three binding sites in both orientations. We computationally generated all possible combinations of synthetic elements with four binding sites (4096 total) and randomly chose 715 of these. In addition, the CREs were designed so that after cloning, the basal promoter alone was a part of the library.

CRE-seq Library Construction

A pool of 6,500 unique 150-mer oligos was ordered through a limited licensing agreement with Agilent Technologies. Oligos were structured as follows: 5' priming sequence (GTAGCGTCTGTCCGTC)/XbaI site/CRE/SphI site/randomly generated sequence/EagI site/buffer(ACG)/9 bp barcode/buffer(CGC)/ClaI site/3' priming sequence (GCAACTACTACTACAG). Using this method of synthesis, every oligo must be 150 bp in length so the random sequence length was determined for each CRE length. The random sequence is removed in the second cloning step and only serves to pad the oligo length for synthesis. Barcode sequences were generated as described (Kwasnieski et al. 2012).

The plasmid library was prepared as described (Kwasnieski et al. 2012), except using primers JKP6F and JKP6R (Table 3.S1) and an annealing temperature of 55C. The amplified library product was purified on a polyacrylamide gel as described (White et al. 2013). Purified library amplicons were cloned into pJK01 (Kwasnieski et al. 2012) using XbaI and ClaI. We prepared DNA from 54,000 colonies to generate PL6_1. We then cloned the Rho proximal promoter driving DsRed into PL6_1. A cassette containing the Rho promoter fused to DsRed was amplified from pJK01 with primers JKP7F and JKP7R (Table 3.S1) and cloned into library PL6_1 by using SphI and EagI, creating library PL6_2. To decrease

background in our library, we linearized PL6_2 with HindIII, gel purified the library, and re-circularized to create library PL6_3.

Retinal Electroporation

Electroporations and explant cultures were performed as described (Hsiau et al. 2007) by using 0.5 µg/mL library PL6_3 as well as 0.5 µg/mL CAG promoter driving GFP for visualization of electroporation efficiency (Lee et al. 2010). After 8 d in culture, retinas were washed twice in sterile HBSS (Gibco, Life Technologies), and total RNA was extracted by using TRIzol according to manufacturer's instructions (Ambion, Life Technologies).

Preparing Samples for RNA-Seq

Excess DNA was removed from extracted RNA using the TURBO DNA-free kit (Applied Biosystems), following manufacturer's instructions. First strand cDNA was synthesized from the RNA using SuperScript II Reverse Transcriptase (Life Technologies). Both the cDNA samples and the DNA from the original plasmid library were prepared for sequencing using a custom protocol as described (Kwasnieski et al. 2012). Briefly, we used PCR amplification of the sequence surrounding the barcode in the RNA transcript or plasmid using primers JKP3F and JKP3R (Table 3.S1). We then digested the PCR product using EagI and EcoRI and ligated Illumina adapter sequences (Kwasnieski et al. 2012) to these amplified sequences. Libraries were sequenced using the Illumina HiSeq platform. Reads that perfectly matched the first 13 expected nucleotides were counted, regardless of quality score. This resulted in 77.5 million reads from the cDNA, across 4 biological replicates, and 34.8 million reads from the DNA. Only barcodes with >=50 reads in the DNA pool and >=3 reads in the cDNA pool were used for downstream analysis. The expression of each barcode was calculated as (cDNA reads)/(DNA reads) and then normalized to the expression of the basal promoter alone. The expression of each CRE was calculated as the mean of the expression of each BC associated with it.

Figure 3.1

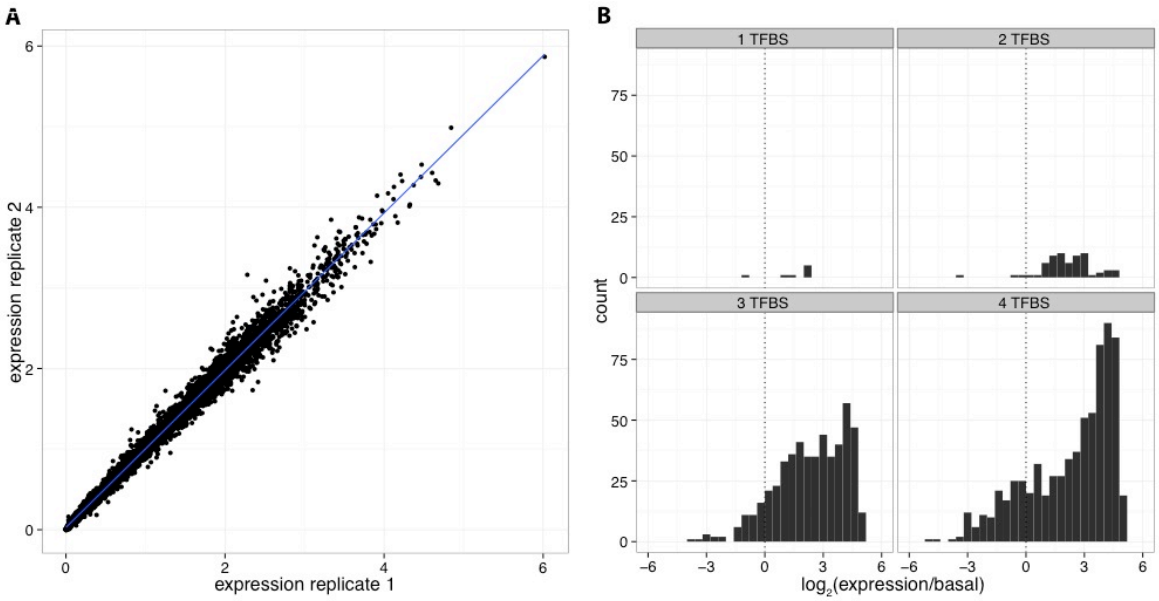


Figure 3.1 Synthetic CREs have strong and reproducible effects on expression. A) Biological replicates of the expression driven by synthetic CREs are highly reproducible. Each dot represents the expression controlled of a single CRE. Blue line is the best fit (R^2=0.99). **B)** Every length of synthetic elements controls a range of expression levels. Histogram shows CRE expression (log 2) normalized to the expression of the basal Rho promoter alone. Dotted line is expression that is identical to the basal promoter. Panels separate CREs by length, or number of component TFBS.

Figure 3.2

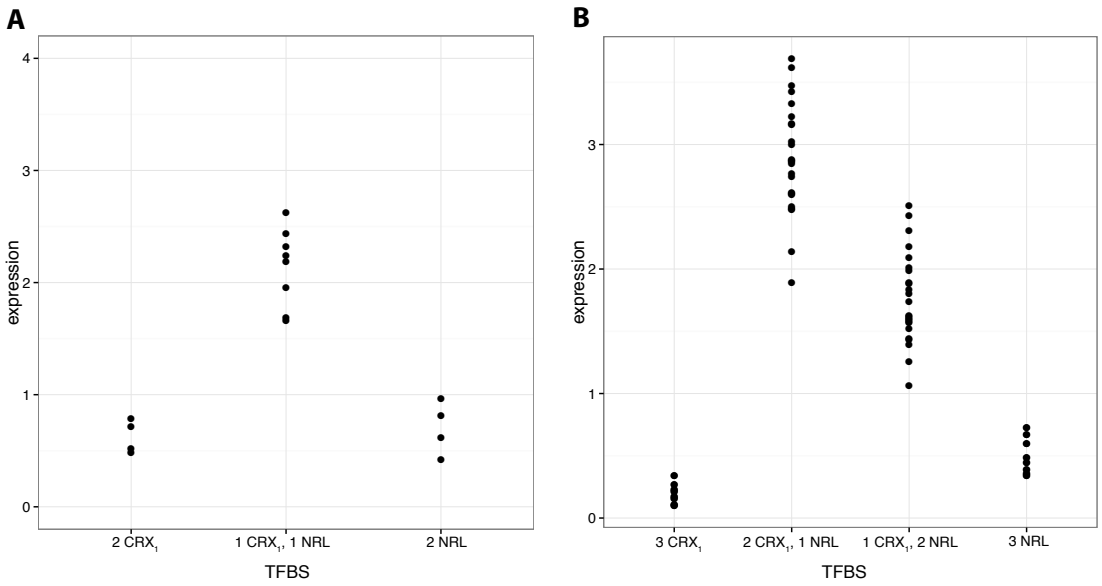


Figure 3.2 Synergistic regulation of CRX1 and NRL binding sites. Synthetic CREs with both CRX1 and NRL binding sites drive significantly higher expression than synthetic CREs with only CRX1 or NRL binding sites. X-axis lists component TFBS, regardless of order and orientation. Y-axis measures CRE expression. A) Synthetic CREs with two TFBS. B) Synthetic CREs with three TFBS.

Figure 3.3

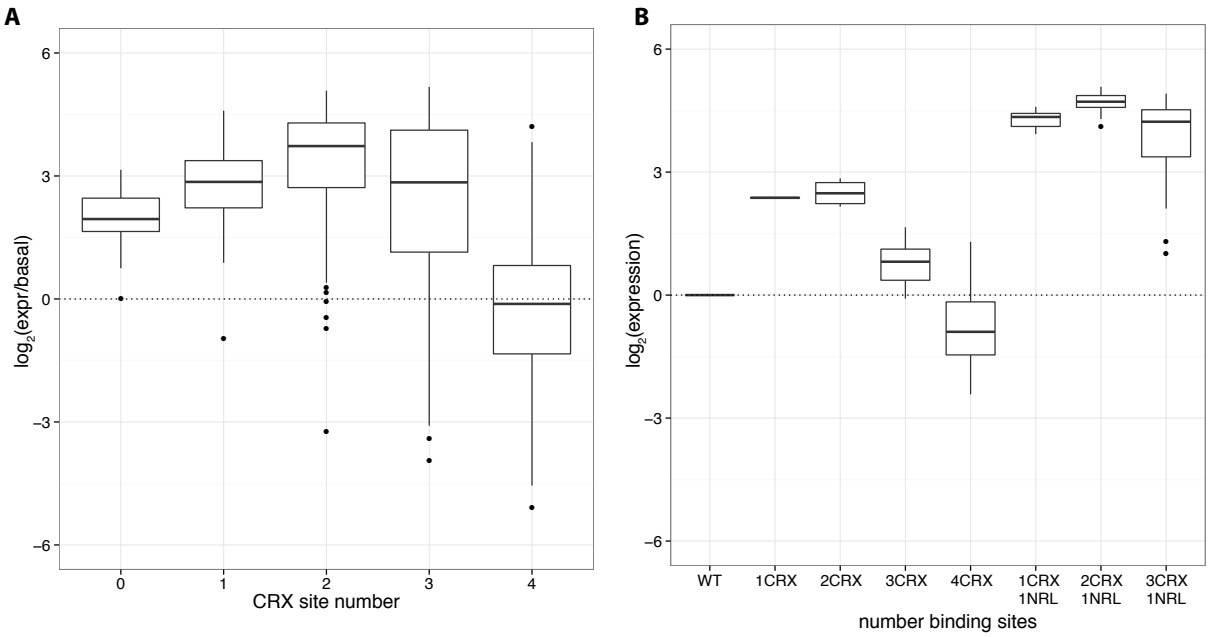


Figure 3.3 Synthetic CREs with many CRX binding sites repress the basal promoter. A) Synthetic CREs composed of many CRX sites drive low expression and can even repress the basal promoter. Boxplot shows CRE-seq expression (log 2 scale) separated by the number of CRX_1, CRX_2, and CRX_3 TFBS in each CRE. B) CRX_1 homopolymers can repress the basal promoter level. Adding NRL binding sites increases total expression but does not counteract the trend that adding CRX_1 sites decreases expression. Boxplot shows CRE-seq expression (log 2 scale) driven by synthetic CREs that are only composed of CRX_1 and NRL TFBS. X-axis denotes the number of CRX_1 and NRL TFBS in each CRE.

Figure 3.4

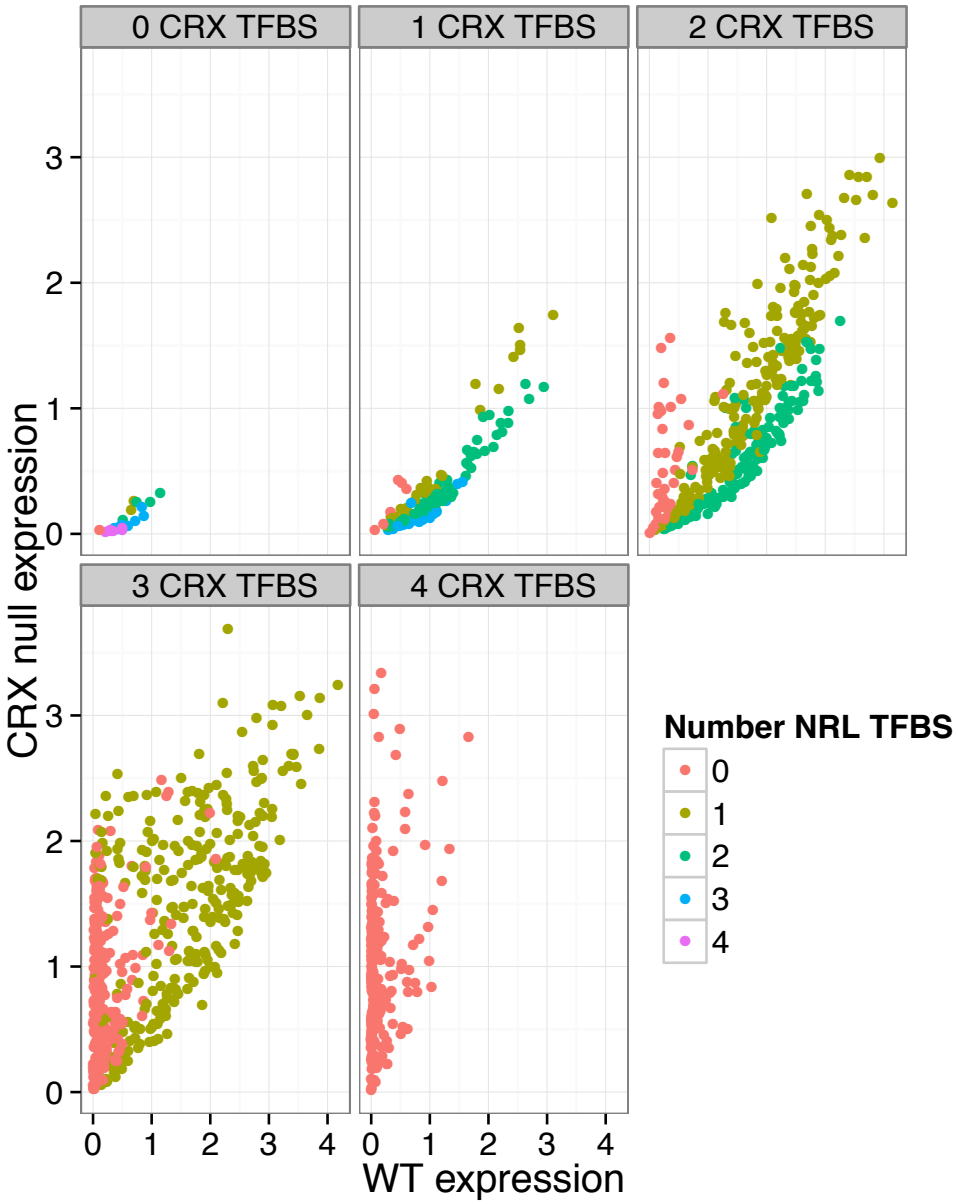


Figure 3.4 Synthetic CREs display CRX-dependent repression but only in the absence of NRL TFBS. Scatterplot compares expression driven by synthetic CREs in the wild-type (x-axis) and the CRX null (y-axis) mouse retinas. Each dot represents the expression controlled by a single CRE. Panels separate expression by the number of CRX_1, CRX_2, and CRX_3 TFBS in each CRE. Dot color denotes number of NRL TFBS in each CRE.

Table 3.1

CRX_1	GCTAATCCCA
CRX_2	GCTAAGCCAA
CRX_3	GCTGATTCAA
NRL	GACTGCTGACTCAGCATCA

Table 3.1 Sequences of binding sites that were used in this study.

SUPPLEMENTARY FIGURES AND TABLES

Table 3.S1

JKP3F	TGTTCCTGTAGCGGCCGACG
JKP3R	GAATTCTAGCCAGAAGTCAGATGCTCAAG
JKP6F	GTAGCGTCTGTCCGTCTCTAG
JKP6R	CTGTAGTAGTAGTTGCATCGATG
JKP7F	TAGCATGCATGTCACCTTGGCCCCTC
JKP7R	GGCCATCGATCTCGAGTAAC
JKP8F	TCTCTAGACTCCTCCGGCTCGCTGATTG
JKP8R	CAGGTACCCCATGGCGCCGCGCTC
JKP9F	TAGCATGCCTCCTCCGGCTCGCTG
JKP9R	AGCCTGCACCTGAGGAGT

Table 3.S1 Primer sequences used in this study.

Figure 3.S1

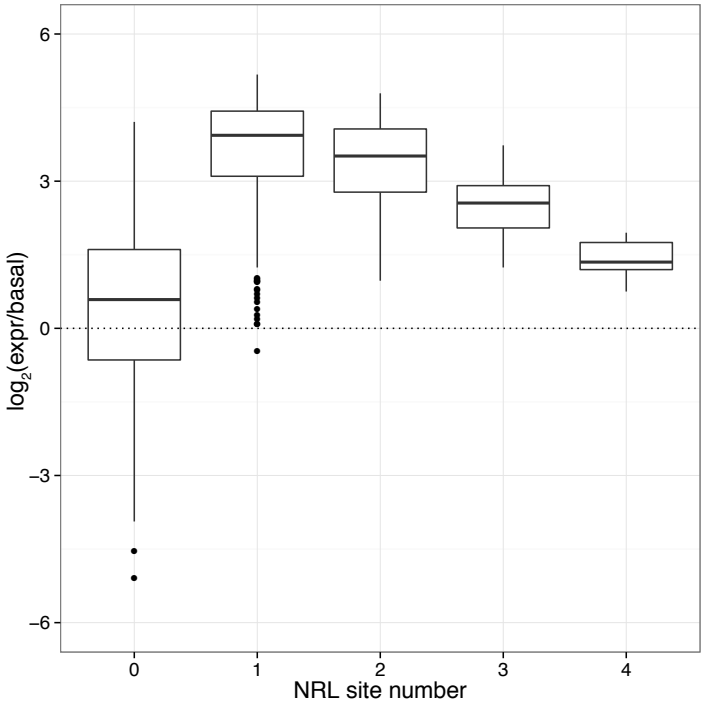


Figure 3.S1 Synthetic CREs with many NRL binding sites drive low expression. Synthetic CREs composed of many NRL TFBS drive low expression. Boxplot shows CRE-seq expression (log 2 scale) separated by the number of NRL TFBS in each CRE.

Figure 3.S2

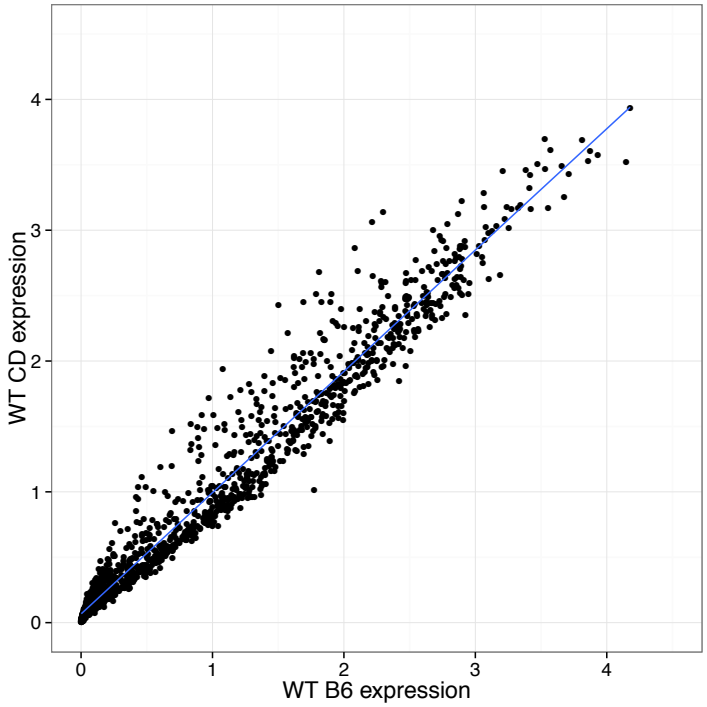


Figure 3.S2 Synthetic CREs control similar expression levels in two wild-type mouse backgrounds. Comparison of expression driven by synthetic CREs in wild-type B6 (x-axis) and wild-type CD (y-axis) mice backgrounds shows high correlation (R^2=0.96).

CHAPTER 4: DISCUSSION

There are three main aims of my thesis: to develop a high-throughput multiplexed reporter assay technique, to use this method to understand how nucleotides contribute to the control of gene expression, and to use this method to understand how combinations of transcription factor binding sites (TFBS) regulate expression. Toward the first aim, I developed a method for *cis*-regulatory element (CRE) sequencing (CRE-seq) that allows high throughput construction and assay of thousands of CREs in a single experiment. This method produces reproducible gene expression measurements that have a high correlation with traditional reporter assay measurements. Toward the second aim, I utilized this method to understand the impact of each nucleotide in the Rhodopsin (Rho) promoter. I found that most single nucleotide variants have a significant impact on gene expression levels and that most combinations of variants drive expression that is not an additive sum of the single variant expression levels. Toward the third aim, I tested combinations of CRX and NRL TFBS to understand how they drive expression. I found that seemingly simple synthetic CREs generated diverse expression levels due to dual regulation and TF synergy. These findings have broad impacts for the study of mammalian *cis*-regulation, as I discuss below.

Comparison of Methods for High Throughput Reporter Analysis

Early methods for high throughput reporter analysis used *in vitro* transcription (Patwardhan et al. 2009) or quantitative PCR (q-PCR) for expression measurement (Nam et al. 2010). Our CRE-seq method was published along side several other studies (Melnikov et al. 2012; Patwardhan et al. 2012; Sharon et al. 2012; Arnold et al. 2013) developing similar methods. All of these methods multiplex reporter gene expression measurements by barcoding the reporter gene 3' UTR, but each has differences, mainly in the process of library construction. Here I compare the methods that multiplex reporter expression measurements in eukaryotic cells with next-generation sequencing.

Two of these methods utilize random construction strategies so they do not have limitations on CRE length or library complexity. First, in the Starr-seq (Arnold et al. 2013) method the enhancer sequence is cloned into the 3' UTR and it serves as the CRE as well as the barcode. Studies have shown

that 3' UTR sequences can have large effects on gene expression outside of transcriptional control (Shalem et al. 2013), so the accuracy of Starr-seq approach is limited.

Second, In the Massively Parallel Functional Dissection (MPFD) (Patwardhan et al. 2012) method, stitching PCR and random library construction is paired with a PCR-based subassembly method to match CRE and barcode sequences. Expression is calculated as the fraction of biological replicates where a barcoded mRNA is detected, which is in contrast to most other methods that use the abundance of barcoded mRNA. Published uses of this method have measured expression from very complex library pools, which require large amounts of sequencing coverage and are therefore expensive.

Two other methods use array-based oligo synthesis for library construction, as does CRE-seq. This approach is advantageous because it allows construction of a high fidelity CRE library with systematically designed sequence derivatives. First, the Massively Parallel Reporter Assay (MPRA) (Melnikov et al. 2012) method is most similar to CRE-seq. The main difference between MPRA and CRE-seq is that MPRA sequences DNA barcodes from the original plasmid pool, rather than from the transfected cells. Our work has shown these approaches result in similar DNA counts and therefore we have switched to sequencing the plasmid pool to reduce sequencing costs.

Second, the Segal lab (Sharon et al. 2012) method measures fluorescence, separates cells based on expression, and sequences barcodes from the binned pools. Expression is calculated as the abundance in each expression pool, which results in a more course-grained measurement of gene expression than the other methods.

Taken together, these methods primarily differ in library construction strategies, which allows the user to choose a method based on CRE design constraints. In particular, measuring the abundance of mRNA barcodes is very quantitative and I recommend methods that take advantage of this. Potential improvements to CRE-seq could include the integration of the reporter gene constructs the genome, or the potential to read out spatial or temporal expression patterns with the method.

Potential Improvements to CRE-seq

The limitations of this technology are twofold: the plasmid expression system is transient and CRE-seq does not currently support the measurement of temporal and spatial expression patterns.

Technique improvements could address these issues. Currently, the number of divisions that the transfected cell population undergoes can limit the duration of CRE-seq experiments because the plasmid library is diluted in the cell population. Integrating the CRE-seq libraries with lentiviral transduction or CRISPR editing technologies would allow expression profiling in a stable cell population, expanding the number of tissues amenable to CRE-seq analysis. Genes are expressed in temporal and spatial patterns that are important for appropriate cell function but current CRE-seq methodology can only measure expression from a homogeneous cell population. Harvesting RNA across different time points or isolating single cell RNA with laser capture micro-dissection would allow CRE-seq to measure expression in relevant temporal and spatial patterns.

Phenotypic Interpretation of Mutagenesis Study

In Chapter 2, we found that most sequence variants have a significant impact on expression. This result is surprising because nearly all mutations, even those located in sequences that are not thought to bind a TF, drove expression that was significantly different than wild-type expression. In addition, we concluded that combinations of these variants drove expression that was not accurately predicted from single variant expression levels. Together, these results imply that for non-coding variation in promoter sequences, most variants will have functional consequences for expression control, and many of these variant combinations will generate non-additive expression levels.

To my knowledge, no other studies have examined the activity of a mammalian promoter in such detail. We know that Rho is a highly expressed gene and its precise expression is critical for homeostasis of the retina. It is possible that this promoter is precisely tuned so that correct activity is easily disrupted. Two studies (Melnikov et al. 2012; Patwardhan et al. 2012) have tested the impact of enhancer variants on gene expression and found that few sequence variants drove significantly different expression than wild-type levels. There are several biological reasons to explain the differences in these results; technical reasons are discussed in Chapter 2. First, sequence variants in enhancers may be less likely to disrupt gene expression control than sequence variants located in promoter sequences. Second, the activity of these specific enhancers may be less affected by sequence variants than the activity of the Rho promoter. Third, it may be that this reporter assay design, where all regulatory sequences are cloned

fewer than 500 base pairs upstream of the transcription start site, is more sensitive to expression changes in promoters than enhancers. To further test these ideas, one could use CRE-seq to examine the impact of sequence variants on enhancer activity.

The conclusion that most variants significantly impact expression is based on the results of statistical tests that determined mutant expression measurements were significantly different than wild-type expression levels. A large fraction of variants drove significantly different mean levels of expression, but it was not clear whether this number of significant changes was more or less than expected. There is no established null model for predicting how regulatory sequence activity changes upon mutation. To develop a null model that describes the frequency that of variants that impact expression, one could conduct a similar study, but compare expression of random genomic sequences against their mutated versions. The results of this experiment would further develop our understanding of how sequence variants impact expression control.

It is clear from this mutagenesis study that sequence variants in the Rho promoter can have large effects on the expression level of the gene. Future work should characterize these sequence variants in the context of other genomic elements and at a phenotypic level. We see large expression changes when the mutated promoter drives expression in isolation, but in the mouse genome there are many regulatory elements that control expression of the Rho gene. Experiments to measure the expression changes of these promoter mutations in the Rho locus would develop our understanding of whether sequence variants are buffered in a genomic context. We observed sequence variants that control expression 60-fold lower than wild-type levels and I hypothesize this Rho dysregulation would be sufficient to induce a phenotype. Experiments to generate transgenic mice with single Rho promoter mutations and assay the transgenic retinas for structure and function would answer this question. These experiments will help us understand how expression studies of isolated genomic elements can translate into interpretation of the impact of human non-coding variation.

This study is one of the first detailed examinations of how sequence variants impact the activity of a mammalian promoter. It demonstrates that mammalian promoters are precisely tuned and therefore subject to significant expression changes upon mutation. These results imply that human non-coding variation in promoter sequences will likely have functional consequences for gene expression control.

Complex Expression Controlled by Simple *cis*-Regulatory Elements

In Chapter 3, I discussed our work to understand how combinations of TFBS control expression. The main conclusion of this study is that simple regulatory elements, only containing TFBS for transcription activators, generate diverse expression levels. These simple elements can drive expression levels that are activated above the basal promoter and expression levels that are repressed below the basal promoter. This result motivates the continued study of *cis*-regulation in controlled model systems where annotated transcriptional regulators have measured binding specificity because we are surprised by the complexity in expression that is controlled by simple regulatory elements.

Synthetic regulatory elements provide a useful paradigm for the study of *cis*-regulation because they allow TFBS interrogation while minimizing the number of confounding variables. The simplicity of these reporter constructs allows for the detection of *cis*-regulatory patterns but it can also limit the study. For example, if TFs are regulating synthetic CREs differently than genomic CREs, due to the design of synthetic CREs, discovered patterns may not be useful in studying genomic CREs. Developing a hybrid approach where one may study the expression controlled by CREs that are rooted in genomic sequences but have synthetic attributes, would allow for statistical power while maintaining relevant CRE design. This might include genomic sequences where TFBS identity and/or orientation are swapped. The results of this experiment would continue to build our understanding of how TFBS control expression, but easily translate into conclusions about genomic elements.

This work contributes to the study of *cis*-regulation in the mammalian rod photoreceptor cell because we have learned new principles that govern how combinations of NRL and CRX binding sites regulate gene expression. Future work, as discussed in Chapter 3, will generate data for the development of a quantitative framework. This model will be used to describe regulation by CRX and NRL, and this model will improve expression predictions for rod-cell specific genes.

APPENDIX 1: MASSIVELY PARALLEL SYNTHETIC PROMOTER ASSAYS REVEAL THE *IN VIVO* EFFECTS OF BINDING SITE VARIANTS

Ilaria Mogno[1], Jamie C. Kwasnieski[1], and Barak A. Cohen

[1]These authors contributed equally to this work
Center for Genome Sciences and Systems Biology, Department of Genetics, Washington University School of Medicine in St. Louis, MO, 63108

This work was done in collaboration with Ilaria Mogno and Barak Cohen. Ilaria Mogno, Barak Cohen and I conceived the project and designed experiments. Ilaria Mogno and I performed the experiments and Ilaria Mogno did the data analysis. Ilaria Mogno, Barak Cohen and I wrote the paper. This chapter is a manuscript published in 2013 in the journal *Genome Research*.

ABSTRACT

Gene promoters typically contain multiple transcription factor binding sites (TFBS), which may vary in affinity for their cognate transcription factors (TF). One major challenge in studying *cis*-regulation is to understand how TFBS variants affect gene expression. We studied the *in vivo* effects of TFBS variants on *cis*-regulation using synthetic promoters coupled with a thermodynamic model of TF binding. We measured expression driven by each promoter with RNA-seq of transcribed sequence barcodes. This allowed reporter genes to be highly multiplexed and increased our statistical power to detect the effects of TFBS variants. We analyzed the effects of TFBS variants using a thermodynamic framework that models both TF-DNA interactions and TF-TF interactions. We found that this system accurately estimates the *in vivo* relative affinities of TFBS and predicts unexpected interactions between several TFBS. Our results reveal that binding site variants can have complex effects on gene expression due to differences in TFBS affinity for cognate TFs and differences in TFBS specificity for non-cognate TFs.

INTRODUCTION

Transcription factors (TFs) orchestrate programs of gene expression by binding promoters and interacting with the core transcriptional machinery. Promoters typically contain multiple transcription factor binding sites (TFBS) with varying affinities for their cognate TFs. Analyses of TFBS variants must account for the effects of low affinity sites, which often have important and surprising roles in gene regulation, especially when TFs bind cooperatively (Driever et al. 1989; Jiang and Levine 1993; Wharton et al. 2004; Gertz et al. 2009; Parker et al. 2011; Peterson et al. 2012). Position weight matrix (PWM) models (Stormo 2000) of binding affinities facilitate the study of TFBS variants; however, these models are often developed *in vitro* and offer a limited picture of the *in vivo* effects of variants on gene expression. The effect of a TFBS variant on gene expression is a function of the sum of its effects on binding by, potentially all, other TFs present in the nucleus. In support of this model, recent genome-wide binding studies show a striking overlap of TF binding profiles (Neph et al. 2012). Therefore, given all the possible interactions between TFs and between TFs and DNA, it is difficult to model and predict the *in vivo* effects of TFBS variants. The analysis of TFBS variants is particularly relevant in light of studies of human genetic variation (Abecasis et al. 2012) and the role of non-coding polymorphisms in complex traits and disease (Degner et al. 2012; Maurano et al. 2012). Progress in this field requires methods to study the effects of combinations of TFBS variants inside cells.

Synthetic promoters are powerful tools for studying *cis*-regulation (Cox et al. 2007; Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010; Raveh-Sadka et al. 2012). Recent advances in DNA synthesis and high-throughput sequencing have driven the development of novel techniques for measuring large numbers of synthetic promoters (Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012; Sharon et al. 2012; Arnold et al. 2013). These techniques add transcribed sequence barcodes to traditional fluorescent reporter genes, allowing reporter genes to be highly multiplexed and assayed by RNA-seq. To date, all of these methods rely on plasmid-based reporter gene libraries. Limitations in the length of synthesized DNA restrict some of these techniques to assaying relatively short synthetic regulatory elements. Here we present a method to assay large numbers of chromosomally-integrated synthetic promoters of arbitrary size. We implemented the method in the yeast *S. cerevisiae* and used it to study the effects of TFBS variants on *cis*-regulation.

The method we developed is a variant of CRE-seq (Kwasnieski et al. 2012), a technique created to transiently assay large numbers of *cis*-regulatory elements in mammalian cells. The modifications we made to this technique allow us to sample large numbers of chromosomally-integrated synthetic promoters consisting of combinations of TFBS with differing affinity. This large sampling was necessary to obtain the statistical power necessary to model the effects of TFBS variants on *cis*-regulation. We fit a thermodynamic model to the resulting data, which provides a formal framework to describe the system in terms of TF binding to DNA, and interactions between TFs. We found that binding site variants have complex effects on gene expression that are due to both differences in affinity for their cognate TFs, and differences in specificity for non-cognate TFs.

RESULTS

Construction Of A Barcoded Synthetic Promoter Library

We sought to understand how sequence variants of TFBSs affect gene expression. We previously used libraries of fluorescent reporter genes to study *cis*-regulatory interactions between four TFBSs, which correspond to binding sites for Mig1, Reb1, Rap1, and Gcr1 (Gertz et al. 2009; Mogno et al.). To build on our previous results with consensus TFBS, we chose to create libraries consisting of variants of these same four sites. For each of the four TFs we chose three variants, with differing predicted affinity, for a total of twelve TFBS (e.g. $Mig1_1$, $Mig1_2$, and $Mig1_3$ denote three variants of the Mig1 TFBS). Table A1.1 shows the specific sequences we chose and the estimated affinities to their cognate TFs as calculated with a position weight matrix (PWM) model (Stormo and Fields 1998; MacIsaac et al. 2006) based on ChIP data (Harbison et al. 2004). We tested TFBS with a wide variety of predicted affinities, from very high ($Mig1_1$), to very low predicted affinity ($Reb1_3$ and $Rap1_2$). Because the increase from four to twelve TFBSs entails an exponential increase of the number of possible synthetic *cis*-regulatory elements (CREs), we implemented CRE-seq technology to multiplex our expression measurements.

We built a CRE-seq reporter library in which each synthetic CRE reporter gene contained a unique sequence barcode in its 3' UTR. We first synthesized double-stranded oligonucleotides (oligos) corresponding to each of the three TFBS variants for each TF: Mig1, Reb1, Rap1 and Gcr1 (Table A1.1).

These oligos were pooled and then randomly ligated to form a library of synthetic CREs (Figure A1.1 panel A, Supplementary Figure A1.S1), which was cloned into a plasmid. We then inserted a library of random fifteen nucleotide barcodes downstream of the CREs, such that each barcode uniquely identified a specific CRE. We performed this cloning step in such a way that each CRE was attached to more than one unique barcode: our final library contained 7289 BCs representing 2534 unique CREs (Supplementary Figure A1.S2). This redundancy increases our statistical power by providing multiple expression readouts for any specific CRE.

We matched the barcodes to their specific CRE using paired-end sequencing of the plasmid library containing the CREs and barcodes (Figure A1.1 panel B, Supplementary Figure A1.S1), coupled with a naïve clustering algorithm (Methods). We were careful not to use PCR to prepare the library for sequencing as we found that PCR amplification creates chimeric products that scramble the CRE-barcode associations. After determining the CRE-barcode associations we cloned a cassette containing a basal promoter (*TSA1*) driving *Yellow Fluorescent Protein* (YFP) into the library, between the CREs and the barcodes. The entire library cassette was then excised and inserted into the *S. cerevisiae* genome at the *TRP1* locus (Figure A1.1 panel C).

To measure, in parallel, the expression driven by each CRE, we grew the integrated yeast library, and then sequenced the barcodes from harvested RNA and genomic DNA (gDNA). We computed the cDNA/gDNA ratio of each barcode and used the median ratio for all barcodes corresponding to a particular CRE as the expression of that CRE.

CRE-Seq Accurately Measures Gene Expression

To test the accuracy of the CRE-seq method in *S. cerevisiae*, we compared expression measurements made by CRE-seq to those made by flow cytometry. We picked 337 CREs containing sites for $Mig1_1$, $Reb1_1$, $Rap1_1$ and $Gcr1_1$ and measured their expression in glucose minimal media by flow cytometry. We then pooled all strains and measured their expression in glucose minimal media by CRE-seq. The high correlation (r=0.92) between pooled CRE-seq measurements and individual flow cytometer measurements confirms that CRE-seq accurately measures gene expression in our system (Figure A1.2 panel A).

To verify that the fifteen base pair barcodes in the 3'UTR of the reporter genes do not affect our measurements, we assayed the effects of barcode sequences on expression. Using CRE-seq we assayed the expression of 602 clones of the same promoter, in which each clone contained a different barcode sequence in its 3' UTR. We performed two replicates of this experiment. If the barcodes have an effect on gene expression we expect to see a positive correlation between the two replicates, as barcodes that increased reporter expression would be correlated between replicates. However, we observed a low correlation between the two replicates (r=0.04). The lack of correlation demonstrates that the random barcodes in the 3'UTR do not have reproducible effects on expression (Figure A1.2 panel B).

Model Selection

After verifying the accuracy of our assay, we analyzed the full library, composed of 7289 BCs for 2534 CREs. To understand the rules of combinatorial regulation we applied a thermodynamic model to our data. This model is a formal framework that describes the data in terms of TF binding to DNA, and interactions between TFs, and provides an automated method to detect the effects of TFBS in large sets of promoters (Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010). Because the differences in expression between members of our library were not correlated with predicted nucleosome occupancy (R^2=0.033), we did not explicitly model interactions with nucleosomes (Kaplan et al. 2009).

We first analyzed CREs containing only $Mig1_1$, $Reb1_1$, $Rap1_1$, and $Gcr1_1$ sites, and recapitulated the results in (Gertz et al. 2009; Mogno et al. 2010) (Supplementary Table A1.S1), showing that $Mig1_1$ sites act cooperatively to repress expression, while the $Reb1_1$, $Rap1_1$, and $Gcr1_1$ sites all have activating effects. We therefore demonstrated that the trends in expression data from CRE-seq recapitulate the trends in expression measured by traditional reporter gene assays.

To explore different potential mechanisms that could account for the effects of TFBS variants we applied several thermodynamic architectures to the full data set with all twelve TFBSs. We started with the simplest set of hypotheses: each TF binds at its three cognate TFBS with different affinities (Figure A1.3 panel A). We also included a parameter to represent the Mig1-Mig1 cooperative interaction that was found in Gertz et al (Gertz et al. 2009) and verified in our data.

We then asked whether our data supported a model with additional interactions. We started by generating a list of additional features (hypotheses) that were not present in the initial simple model, including: 1) an interaction term between Rap1 and Gcr1, as suggested by (Scott and Baker 1993; Tornow et al. 1993), 2) a cooperativity term for Gcr1, as suggested by our expression data (Supplementary Figure A1.S3) and by (Scott and Baker 1993), 3) a term allowing a protein (X) other than Reb1 to bind the $Reb1_3$ site, and 4) a term allowing a protein (Y) other than Rap1 to bind the $Rap1_2$ site. We included features 3) and 4) because $Reb1_3$ and $Rap1_2$ had strong effects on expression even though PWM analysis of these sites indicates that they have very low affinities for their TFs, which suggests that their effects may be mediated through the binding of other TFs.

After identifying a set of features that might improve the simple model, we constructed several model architectures including these additional features in various combinations. Each model was fit to the measured expression values and scored based on the sum of squares of the residuals (RSS) and the number of free parameters needed for the fit, introducing a greater penalty for models with more free parameters. When we rank our models based on this score, a clear pattern appears (Figure A1.4 panel A): the addition of the Rap1-Gcr1 interaction consistently lowers the model score (worse model), while adding Gcr1 cooperativity always results in a higher score (better model). Moreover, allowing unknown proteins to bind the $Reb1_3$ and $Rap1_2$ sites (6 TFs in total), results in a better model even after penalization for increased parameter number. The best performing model includes parameters representing Mig1 self-cooperativity, Gcr1 self-cooperativity, a protein (X) other than Reb1 binding at site $Reb1_3$, and a protein (Y) other that Rap1 binding at site $Rap1_2$ (Figure A1.3 panel B). Scoring $Reb1_3$ and $Rap1_2$ against PWMs for known TFs (Spivak and Stormo 2012) suggests that $Reb1_3$ may be bound by Rtg1 (p= 0.0032) and that $Rap1_2$ may be bound by Yer130C (p=0.0059). This result is not surprising given that the PWM models for these two sites ($Reb1_3$ and $Rap1_2$) predicted extremely low affinity for their cognate TFs (see Table A1.1). It is therefore reasonable to expect other TFs to bind to these particular sites.

This model explains 57% (p<<0.01) of all the variance in expression in our twelve-site synthetic promoter library (Figure A1.4 panel B). We performed repeated random sub-sampling validation (Supplemental Figure A1.S5), showing that we obtain similar results with approximately 1000 unique

promoters. However, to obtain reliable estimates of some parameters, at least 2000 unique promoters are necessary. Thus, the extra statistical power afforded by CRE-seq allowed us to identify features of this system that were undetectable in our previous experiments with smaller libraries (Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010).

The Thermodynamic Model Predicts *In Vivo* Relative Affinities Between Tfs and DNA

We next asked whether the *in vivo* predicted affinities estimated from our thermodynamic model match the PWM predictions from ChIP-seq data. We computed DDG for pairs of binding sites. A negative DDG indicates a stronger TFBS (with higher affinity), while a positive DDG indicates a weaker site. When the PWM and our thermodynamic model are in agreement, the DDG calculated with the PWM and the DDG calculated with the thermodynamic model are proportional. For example, our model is in good agreement with the PWM model for Mig1 (Table A1.2, rows 1 and 2). Our model also agrees with PWM predictions that $Rap1_3$ is stronger than $Rap1_1$ (Table A1.2, row 3). In contrast, our model predicts $Reb1_2$ to be stronger than $Reb1_1$, while the PWM model predicts the opposite (Table A1.2, row 4).

Our relative affinity predictions for Gcr1 do not agree with PWM models (Table A1.2, rows 5 and 6). Our model predicts that $Gcr1_2$ is the strongest site for Gcr1, while the PWM model predicts $Gcr1_3$ to be the strongest site. The PWM model was generated using genome-wide chromatin immunoprecipitation (ChIP) data collected in a rich media (Harbison et al. 2004; MacIsaac et al. 2006); our predictions are estimated from measuring synthetic promoter expression in minimal media. It is possible that the inconsistencies in these predictions can be explained by differences in growing conditions or by differences between measuring binding versus activity through a gene expression based reporter assay. We also tried using different PWM models for Gcr1, which came from different experiments. None of these PWMs for Gcr1 are in good agreement with each other, nor do they agree well with our predictions (Harbison et al. 2004; MacIsaac et al. 2006; Pachkov et al. 2007; Foat et al. 2008; Spivak and Stormo 2012). The relationship between occupancy at the Gcr1 sites and the sites' effect on gene expression may be complicated by condition-specific binding of Gcr1, or the binding of other factors. In the following analysis, we refer to $Gcr1_2$ as the strongest site and $Gcr1_1$ as the weakest site, as predicted by our model.

Gcr1 Participates In Complex TF-TF Interactions

The Gcr1 binding sites used in this study showed the ability to enhance the activity of surrounding TFBS, regardless of whether activators or repressors bind to those TFBSs. When $Gcr1_1$ sites are added to promoters containing only $Mig1_1$ sites, their average expression decreases (Figure A1.5, panel A). In contrast, when $Gcr1_1$ sites are added to promoters containing only $Reb1_1$ or $Rap1_1$ sites, the expression of the reporter gene increases. We also observe a similar behavior when $Gcr1_1$ is added to $Reb1_2$, $Reb1_3$, $Rap1_2$ and $Rap1_3$. However, the ability of $Gcr1_1$ to repress is weaker when it is coupled with weaker sites for Mig1 (e.g. $Mig1_2$ and $Mig1_3$, 5, panel B). The data suggest that the $Gcr1_1$ site acts as an activator when next to any activator site, but it acts as a repressor when next to a strong Mig1 site, and has little effect next to a weak Mig1 site. Increasing the predicted affinity of the Gcr1 site hides this behavior: $Gcr1_3$, a stronger site, has a smaller effect on $Mig1_1$ site (Figure A1.5, panel C). The repressing effect disappears when we use $Gcr1_2$, the highest affinity site. Moreover, this effect is particularly strong in promoters in which $Gcr1_1$ and $Mig1_1$ sites are adjacent to each other (Supplementary Figure A1.S4). These data seem to suggest a role of the Gcr1 sites in facilitating the binding of other TFs, and increasing their regulatory potential.

DISCUSSION

We adapted CRE-seq for use with synthetic promoters of arbitrary size integrated into the genome of *S. cerevisiae*. With the development of CRE-seq we can assay thousands of integrated synthetic promoters, a 10-fold increase over what was previously possible with fluorescent reporter genes. We showed that the method is accurate and reproducible, and that the barcodes in the 3' UTR of the reporter gene do not affect gene expression. As technologies for genome editing (Christian et al. 2010; Bogdanove and Voytas 2011) become more efficient, we anticipate using CRE-seq to study synthetic promoters integrated into the genomes of mammalian cells.

An advantage of CRE-seq is that it allows us to build larger libraries, since all clones are built and assayed in parallel. It also overcomes some of the limitations of traditional assays based on flow cytometry, such as limited dynamic range. CRE-seq measures the abundance of mRNA rather than stable fluorescent proteins, whose long half-lives could mask the true promoter activity.

We used CRE-seq to obtain the statistical power necessary to study *cis*-regulation in promoters containing combinations of TFBS variants. The increased power we obtained from analyzing large libraries revealed TFBS effects that we could not detect in smaller libraries composed of the same binding sites. This demonstrates the utility of CRE-seq when applied to synthetic promotes. In many cases our binding affinity predictions agree well with established PWM models of binding (MacIsaac et al. 2006). In cases where our predictions were discordant with PWM predictions, as was the case for $Reb1_3$ and $Rap1_2$, we found that our data supported a model in which these variant TFBSs are recognized by other TFs. We think the differences in these predictions stem from different experimental conditions and the fact that *in vitro* binding is not equivalent to *in vivo* expression potential.

Our work uncovered an unusual interaction between Gcr1 and Mig1. Although Gcr1 sites behave as weak activators, when put in combination with repressive Mig1 sites, Gcr1 sites increase the repressive effects of Mig1. One possible explanation is that Gcr1 binding opens the locus, thus facilitating the binding of Mig1. This manifests as a greater repressive effect of Mig1, but only when the activating potential of Gcr1 is weak.

With the increasing power and affordability of next generation sequencing technologies, we anticipate that CRE-seq will be a useful tool for unraveling other kinds of interactions between *cis*-regulatory sequences.

ACKNOWLEDGEMENTS

We thank Chris Fiore and Robi Mitra for discussions on the CRE-seq technical details; Aaron Spivak for assistance with PWM analysis; Jessica Hoisington-Lopez, Henning Seedorf, Michelle Smith, Brian Muegge and Jeffrey Gordon for next-generation sequencing support. This work was supported by a grant from the National Institutes of Health (GM092910-02).

MATERIALS AND METHODS

Construction of the CRE-BC Library

E. coli strain DH5a was used for all bacterial cloning steps. Plasmid pIM202 was derived from pIM102 (Mogno et al. 2010) by removing the *TSA1* promoter-*YFP* cassette and replacing it with a multiple cloning site (containing sites for BglII, XmaI, BamHI, KpnI, ClaI, EagI, AvrII and XbaI restriction enzymes). CREs were cloned into pIM202 at the BamHI site as in (Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010), and approximately 7000 colonies were scraped for DNA extraction using a maxiprep kit (Sigma GenElute HP Plasmid Maxiprep Kit).

To create random barcodes two oligos containing 15 random nucleotides flanked by 6 or 7 bases (oligos prIM01 and prIM02, Supplementary Table A1.S2), were denatured at 95C in a water bath and then annealed for 16 hours until the water reached room temperature. The BCs were then cloned into the CRE plasmid library using restriction sites EagI HF and XbaI. The ligations were digested with AvrII before transformation to reduce background. Roughly 20000 colonies were then scraped and maxiprepped (Sigma GenElute HP Plasmid Maxiprep Kit) at this step. The TSA1 promoter-*YFP* cassette was amplified from plasmid pIM102 (98°C for 1min, 5 cycles: 98°C for 15s, 56°C for 30s, 72°C for 60s, 25 cycles: 98°C for 15s, 63°C for 30s, 72°C for 60s, and 72°C for 5 min, NEB HF Phusion MM) using primers prIM03 and prIM04 (Supplementary Table A1.S2) and cloned into the library using restriction enzymes KpnI and EagI HF. The ligation mix was digested ClaI after ligation to reduce background. About 35000 colonies were picked at this step. The CRE-BC plasmid library was integrated into *S. cerevisiae* BY4742 (MATa his3D1 leu2D0 lys2D0 ura3D0) at the *TRP1* locus following the procedure described in (Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010). Between 7000 and 8000 *S. cerevisiae* colonies were replicated into SC media with 2% glucose and 5-FAA (5-Fluoroanthranilic acid) to enrich for the colonies carrying the correct integration. These colonies were scraped and pooled for growth and expression assays.

Matching CREs to BCs

CREs and BCs were matched after cloning the BCs into the plasmid library but before inserting the

TSA1 promoter -*YFP* cassette. The plasmid library was digested with restriction enzymes XmaI and XbaI. Illumina paired end adaptors were ligated, and the DNA molecules between 250 and 500 base pairs in length were selected on agarose gel. No PCR was performed to prevent chimeric products that mask the correct CRE-BC pairs. The purified DNA was then sequenced with an Illumina MiSeq using a paired-end 250x50bp protocol to sequence the CRE and BC regions respectively. We obtained about 1 million reads. BCs represented by fewer than 5 reads were not used in the analysis. Occasionally, more than one CRE was associated with a particular BC. In this case the CRE with the highest number of reads was assigned to the BC if and only if it was represented by at least 90% of the total number of reads associated to the BC. Otherwise the BC was not included in our analysis. Subsequently, all BCs associated to the same CRE were analyzed. We calculated the pairwise sequence distance for all BCs representing the same CRE and we eliminated the ones that had similar sequences to another BC of higher rank, assuming that they arise from sequencing errors.

Flow Cytometer Assay

The strains used for the validation experiment, (Figure A1.2 panel A) were picked from the transformation plate and arrayed into 96 well microplates. The CREs and the BCs were sequenced with a Sanger protocol (Beckman Coulter Genomics). Cultures were grown in 500 ml of synthetic complete media lacking uracil with 2% glucose with shaking at 30°C in 2 ml 96-well plates for four hours. The cells were then fixed with paraformaldehyde as described in (Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010). The fluorescence intensities and electronic volumes of 25000 cells from each well were measured on a Beckman Coulter Cell Lab Quanta SC with a multi-plate loader. Fluorescence was then divided by volume to obtain a normalized fluorescence value for every cell. For each well, mean and variance were calculated from the normalized fluorescence values for 25000 events.

CRE-seq

The *S. cerevisiae* library was grown in synthetic complete media lacking uracil with 2% glucose with shaking at 30°C. After 5 hours gDNA and total RNA were harvested. RNA was then treated with Turbo DNaseI (Ambion) to eliminate genomic DNA contamination and cDNA was synthesized using

Superscript II reverse transcriptase (Invitrogen), with oligo-dT primers (IDT). The 3'UTR of the *YFP* gene, containing the BC, was amplified from gDNA and from cDNA (98°C for 1min, 5 cycles: 98°C for 15s, 54°C for 30s, 72°C for 45s, 10 cycles: 98°C for 15s, 58°C for 30s, 72°C for 45s, and 72°C for 5 min, NEB HF Phusion MM) using primers prIM05 and prIM06 (Supplementary Table A1.S2). We also used primers that amplify across the integration region, prIM05 and prIM07 (Supplementary Table A1.S2), on the gDNA to select for correct integrations. Only the BCs represented in this second control gDNA PCR were included in our analysis. The PCR products were purified with QIAquick PRC purification kit (Qiagen), digested with EagI HF and XhoI, and ligated to Illumina adaptor sequences. The final product was amplified (98°C for 1min, 12 cycles: 98°C for 15s, 63°C for 30s, 72°C for 45s, and 72°C for 5 min) with primers prIM08 and prIM09 (Supplementary Table A1.S2) to enrich for molecules containing both adaptors sequences. This library was run on two lanes of Illumina HiSeq machine generating approximately 102 million reads. Only barcodes with >25 reads in the gDNA pool, and at least one read in the cDNA pool were used for the analysis, for a total of 7289 BCs. Expression associated to each BC was then calculated as the number of reads in the cDNA pool divided by the number of reads in the gDNA pool (for the same set of primers). These 7289 BCs mapped to 2633 unique CREs. Subsequently we determined that 99 of these CREs were likely to contain mutations that altered their expression (see Outlier Detection below). The distribution of BCs identifying each promoter was uneven, 16.1% of the promoters had at least 3 BCs associated with them, while the remaining 83.9% had 2 or fewer (Supplementary Figure A1.S2). Finally, expression driven by each CRE was calculated as the median ratio of all the BCs associated to it.

Thermodynamic Model

To model gene expression, we implemented a thermodynamic model of polymerase occupancy originally proposed by (Shea and Ackers 1985). The model and implementation were described previously in (Gertz and Cohen 2009; Gertz et al. 2009; Mogno et al. 2010), and it includes parameters proportional to DGs of the interactions between proteins and DNA and between proteins. We did not model nucleosome effects. We scanned our promoter sequences with the Nucleosome Positioning prediction software (Kaplan et al. 2008), and found very low correlation between predicted nucleosome occupancy averaged across the TFBS region and the measured expression (r=0.184). Moreover, the

averaged nucleosome occupancy predictions were very similar across our promoter sequences (CV=0.06). Akaike Information Criterion (Akaike 1974), which introduces a penalty term for the number of parameters, was used for model selection. Repeated random sub-sampling validation was performed for cross-validation with training sets containing 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of the total number of data points. All calculations were performed using the Matlab package from The Mathworks, Inc. (Natick, MA).

Outlier Detection

Given the high rate of mutations in *S. cerevisiae* transformants, we expect 5-6% of the colonies to contain mutations that could affect gene expression. The CREs and the BCs are sequenced and matched before inserting the basal promoter and YFP gene, and before the transformation into *S. cerevisiae*; therefore we do not detect mutations in subsequent steps. CREs represented by 3 or more BCs are not affected by this problem, since outlier detection is an easy task in these cases. However, our library contains 1806 CREs associated either with one BC only, or with multiple BCs and high variance in expression (CV>0.5). Replicate experiments showed that 95% of the CREs represented by only one BC produce an accurate measure of gene expression. Instead of eliminating all the CREs represented by a low number of BCs, we used the thermodynamic model in a recursive way to identify the true outliers.

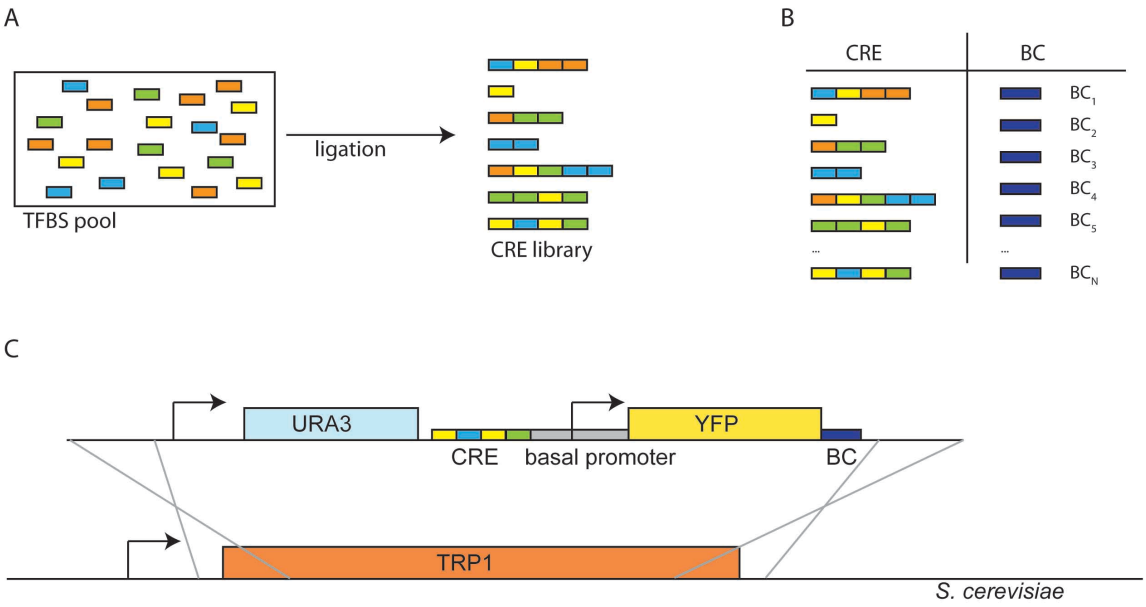
The first step was to apply the thermodynamic model only to the 827 CREs represented by 2 or more BCs, and characterized by low expression variance (CV<0.5). The fit model was used to calculate the error for each of the excluded 1806 CREs. The excluded CREs were ranked based on the error and reintroduced to the model one at the time until the overall R^2 dropped 10% with respect to the original model. This resulted in the exclusion of about 100 CREs. Then the thermodynamic model was applied only to the selected CREs. The CREs excluded from our analysis represent the ones whose expression cannot be explained by the model. There could be two reasons for this: 1) they contain high measurement error, or 2) they contain a specific feature not included in the model. To test whether these CREs contain features that we were not capturing with our model, we looked at the sequence contents of these excluded CRE: they were not enriched in length (number of building blocks), they were not enriched in any specific TFBS, or in any pair of TFBSs. We also tested several models: we added

parameters to include 4 or 6 TFs, and to capture the Gcr1-Gcr1 and the Gcr1-Rap1 sites interactions. Each time we repeated this recursive procedure, excluding between 96 and 115 CREs, and found no common sequence feature in the excluded sets. Moreover, the pairwise intersections of the excluded sets were always between 96% and 100%, indicating a small, reproducible set of outliers. After these analyses we concluded that the unexplained expression for these outlier promoters must be due either to sequencing errors, or to secondary mutations that occurred during their transformation into *S. cerevisiae*. We excluded these outliers from our final analysis, obtaining a final set of 2534 CREs.

PWM Analysis

PWM models for TF binding (MacIsaac et al. 2006; Pachkov et al. 2007; Foat et al. 2008; Spivak and Stormo 2012) were used to estimate the affinity of TFs to their cognate TFBSs. We used *patser* (Stormo et al. 1982) to calculate these scores. The PWM scores are proportional to the –DG of the interaction.

FIGURE A1.1



D

BC	number of reads in cDNA	number of reads in gDNA	E= number of reads in cDNA / number of reads in gDNA
BC_1	1000	500	2
BC_2	680	200	3.4
BC_3	390	50	7.8
BC_4	450	100	4.5
BC_5	700	1000	0.7
...			
BC_N	400	200	2

Figure A1.1 Schematic of the CRE-seq method adapted for this study. A) Double stranded oligonucleotides encoding TFBS are mixed in a pool and ligated randomly to create a CRE library. B) After cloning CRE and BC sequences into a reporter plasmid, the concordance between CREs and BCs is determined with a paired-end next generation sequencing run. Each BC identifies a single CRE. C) The cassette containing the library of CREs upstream of a basal promoter driving YFP and BC is integrated into the *S. cerevisiae* genome at the *TRP1* locus by selecting for URA+ cells. D) Cells are grown in liquid culture, and gDNA and RNA is harvested. The fraction of reads in the cDNA pool divided by the fraction of reads in the gDNA pool for each BC is a quantitative measurement of the expression driven by the corresponding CRE.

FIGURE A1.2

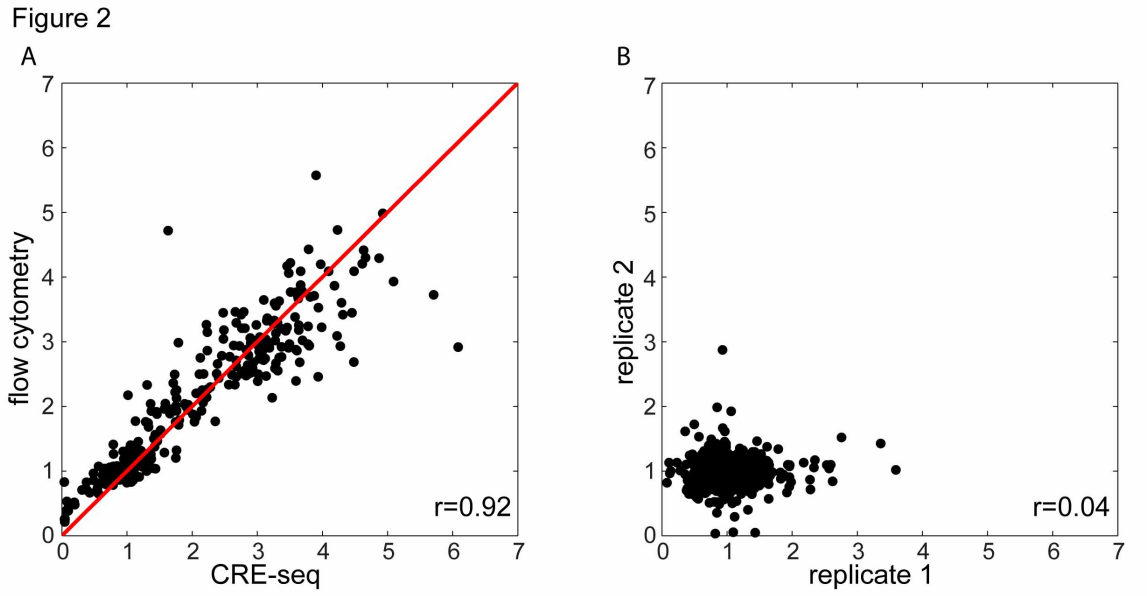


Figure A1.2 CRE-seq accurately measures gene expression. A) Comparison between expression measured by CRE-seq and flow cytometry. Each dot represents a CRE whose activity has been measured with a traditional fluorescent assay (y-axis) and with CRE-seq (x-axis). The high correlation indicates that CRE-seq expression measurements are as accurate as those measured by traditional fluorescent assay. The red line represents the perfect model (r=1). B) Biological replicates of a CRE-seq library where expression is controlled by one CRE matched to 602 different BCs. The library was grown and harvested two times; CRE-seq was performed independently on each replicate. Replicate measurements of BC expression are plotted on the x-axis (replicate 1) and on the y-axis (replicate 2). The absence of correlation reveals that the BCs have no reproducible effects on gene expression.

FIGURE A1.3

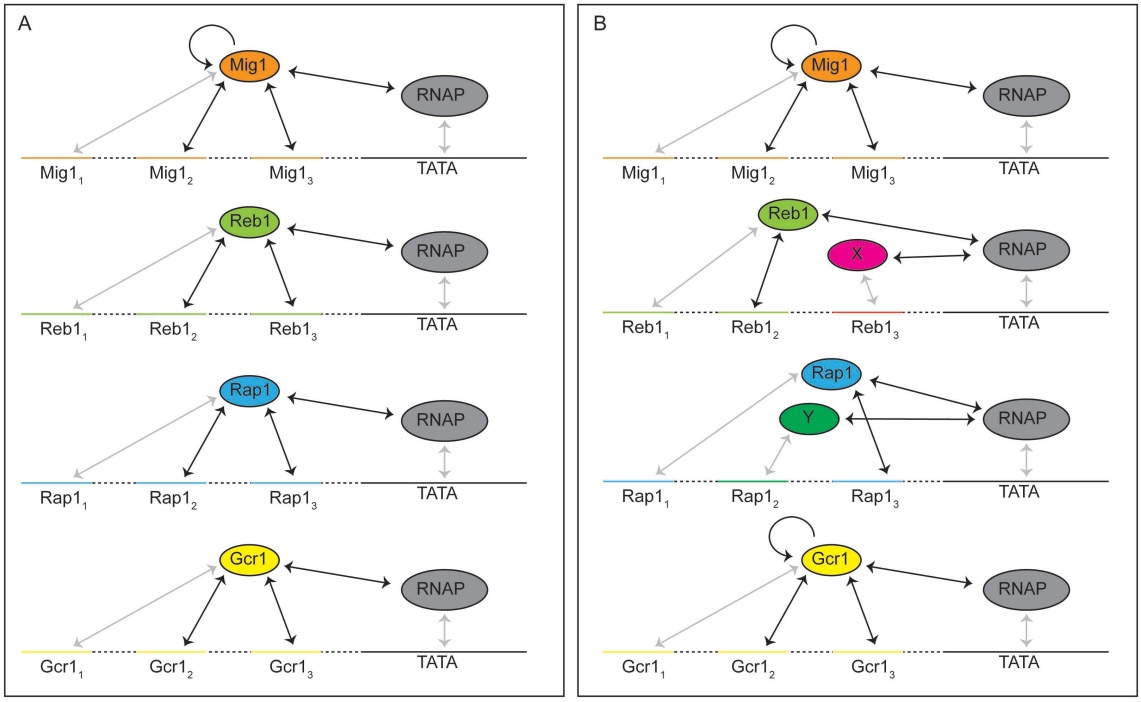


Figure A1.3 The thermodynamic model consists of a set of interactions that govern TF-DNA and TF-TF binding. Each arrow represents an interaction included in the model in the form of a parameter proportional to the DG. Black arrows represent the free parameters. A) The set of interactions allowed in the simplest model: each TF is allowed to bind to its cognate TFBSs, and to interact with polymerase. Mig1 is allowed to interact with itself when two or more Mig1 sites are present in the same promoter. B) The set of interactions applied to the model with the highest score: a protein X other than Reb1 is allowed to bind at site $Reb1_3$, and a protein Y other than Rap1 is allowed to bind at site $Rap1_2$. Both Mig1 and Gcr1 are allowed to interact with themselves when two or more of their sites are present in the same promoter.

FIGURE A1.4

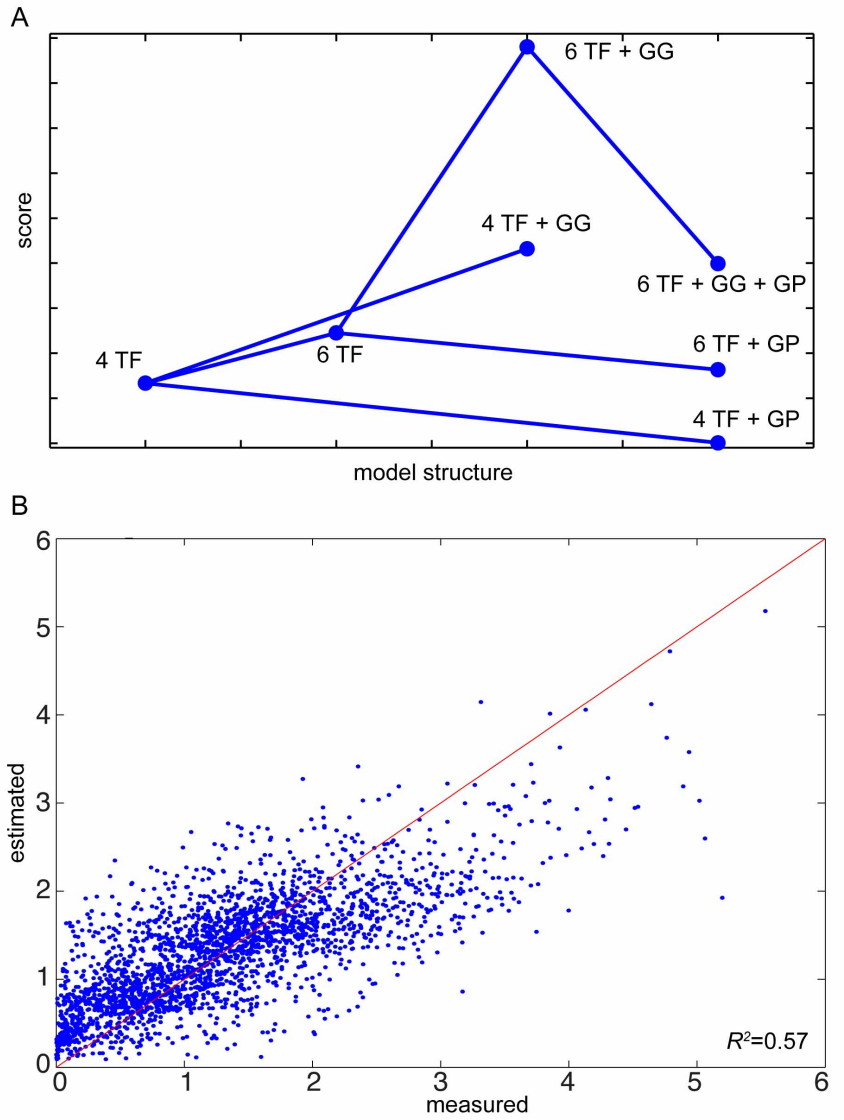


Figure A1.4 A) Several model structures with different sets of rules have been applied to the data. Each dot represents a specific model, whose score is on the y-axis. The score is plotted as the absolute value of the AIC score, calculated taking into account the RSS and a penalty term for the number of free parameters (Methods). An increase in the plotted score (thus a decrease in the AIC score) indicates a better model. Increasing the number of TFs included in the model from 4 to 6 increases the score. The addition of the Gcr1-Gcr1 (GG) interaction always results in a better score. The addition of the Gcr1-Rap1 (GP) interaction always results in a worse score. The model with the best score is the one with 6 TFs and the Gcr1-Gcr1 interaction. All model represented in this plot include the Mig1-Mig1 interaction. B) The thermodynamic model with 6TFs and Gcr1-Gcr1 interaction accurately predicts synthetic promoter gene expression. Each dot represents expression driven by one of the 2534 CREs we assayed in this study. The measured CRE-seq expression is on the x-axis, while the predicted expression from the thermodynamic model is shown on the y-axis. R^2=0.57 shows that our model explains 57% of the variance in the data. The red line represents the perfect model (R^2=1).

FIGURE A1.5

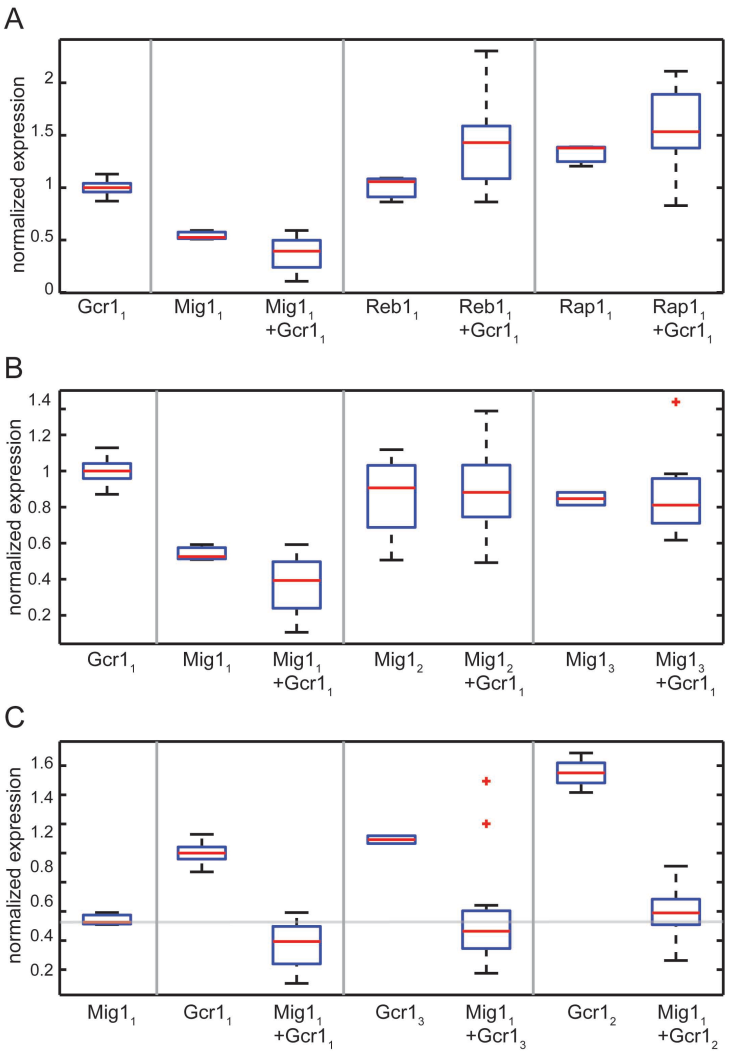


Figure A1.5 Gcr1 binding sites have complex effects on expression. A) When $Gcr1_1$ sites are added to promoters containing only $Reb1_1$ or $Rap1_1$ TFBSs, their effect is to increase the activation of gene expression. However, when $Gcr1_1$ sites are added to promoters containing only $Mig1_1$ sites their effect on gene expression is repressive. B) $Gcr1_1$ TFBS has a weaker repressive effect when added to low affinity Mig1 sites ($Mig1_2$ and $Mig1_3$). C) Gcr1 TFBS with low affinity ($Gcr1_3$ and $Gcr1_2$) have weak repressive interactions when combined with high affinity Mig1 sites ($Mig1_1$).

TABLE A1.1

TFBS	Sequence	Promoter	MacIsaac PWM score
Mig1$_1$	CCCCGGATTT	SUC2	10.4
Mig1$_2$	CCCCACAAAT	MAL61	9.82
Mig1$_3$	CCCCAGGTAT	GAL3	6.69
Reb1$_1$	TTACCCGT	TPI	8.68
Reb1$_2$	TCACCCGT	TRPI	6.15
Reb1$_3$	CAGCCCTT	GALI	-3.11
Rap1$_1$	ACACCTGGACAT	TPI	7.66
Rap1$_2$	ACCCCTTTTTAC	TPI	-3
Rap1$_3$	ACACCCAAGCAT	TPFI	9.95
Gcr1$_1$	CAGCTTCCT	TPI	2.88
Gcr1$_2$	CGGCATCCA	TPI	7.7
Gcr1$_3$	CGACTTCCT	ADH1	8.76

Table A1.1 Summary of binding sites used in this study. This table reports the 12 TFBS sequences in our library, including *S. cerevisiae* promoters where they are present and the PWM score (MacIsaac et al. 2006).

TABLE A1.2

TFBS A	TFBS B	PWM $DG_B - DG_A$	Thermodynamic model $DG_B - DG_A$
$Mig1_1$	$Mig1_2$	0.58	1.10 ± 0.12
$Mig1_1$	$Mig1_3$	3.71	4.40 ± 1.88
$Rap1_1$	$Rap1_3$	-2.29	-1.52 ± 0.51
$Reb1_1$	$Reb1_2$	2.53	-0.14 ± 0.11
$Gcr1_1$	$Gcr1_2$	-4.82	-0.86 ± 0.24
$Gcr1_1$	$Gcr1_3$	-5.88	-0.37 ± 0.30

Table A1.2 Comparison between TFBS affinities predicted by thermodynamic modeling and PWM analysis. For each combination of variant binding sites (columns 1 and 2), we show PWM predicted relative affinities (column 3) and thermodynamic modeled relative affinities (column 4). Each numeric value represents the change in DG for the variant in the second column with respect to the variant in the first column (DDG). A positive number predicts that site B has a weaker affinity than site A, while a negative number predicts site B has a stronger affinity than site A.

SUPPLEMENTARY FIGURES AND TABLES

TABLE A1.S1

Parameter	Values ±95% CI
DGMig1-RNAP	3.76±1.95
DGMig1$^{(R)}$-RNAP	3.95±2.68
DGReb1-RNAP	-0.14±0.08
DGReb1$^{(R)}$-RNAP	-0.05±0.11
DGRap1-RNAP	-0.35±0.11
DGRap1$^{(R)}$-RNAP	-0.21±0.11
DGGcr1-RNAP	-0.27±0.10
DGGcr1$^{(R)}$-RNAP	-0.37±0.08
DGX-RNAP	-0.24±0.08
DGX$^{(R)}$-RNAP	-0.13±0.07
DGY-RNAP	-0.18±0.06
DGY$^{(R)}$-RNAP	-0.20±0.06
DGMig1-Mig1	-0.68±0.33
DGGcr1-Gcr1	-0.15±0.12
DGMig1$_2$-DNA - DGMig1$_1$-DNA	1.10 ± 0.12
DGMig1$_3$-DNA - DGMig1$_1$-DNA	4.40 ± 1.88
DGRap1$_3$-DNA - DGRap1$_1$-DNA	-1.52 ± 0.51
DGReb1$_2$-DNA - DGReb1$_1$-DNA	-0.14 ± 0.11
DGGcr1$_2$-DNA - DGGcr1$_1$-DNA	-0.86 ± 0.24
DGGcr1$_3$-DNA - DGGcr1$_1$-DNA	-0.37 ± 0.30
R^2	0.57

Table A1.S1 This table reports the estimated parameters of our thermodynamic model as proportional to DG or the difference of DGs (DDG). X and Y denote the unknown TFs predicted to bind at site Reb1$_3$ and Rap1$_2$ respectively. (R) denotes the TFBS in the reverse orientation. A negative value for DG between TF and RNAP or TF-TF indicates a favorable interaction, while a positive value indicates an unfavorable interaction. A negative value for the difference of DGs between TF and DNA indicates a stronger site (higher affinity), while a positive value indicates a weaker site (lower affinity).

TABLE A1.S2

Primer name	Sequence
prIM01	GGCCGAANNNNWNNNNWNNNNNAGCTCGT
prIM02	CTAGACGAGCTNNNNNWNNNNWNNNNTTC
prIM03	TATAGGTACCGGCTCGGGTTGGCAAA
prIM04	GTTCCGGCCGTTATTTGTACAATTCATCCATACCATGGGT
prIM05	TGAATTGTACAAATAAGCGGCCGAA
prIM06	TTAACTCGAGAAGTAGATAAAGTCAGTGCTTAAACAC
prIM07	ATATCTCGAGCCTGCGATGTATATTTTCCTGTAC
prIM08	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
prIM09	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Table A1.S2 Primer sequences used in this study.

FIGURE A1.S1

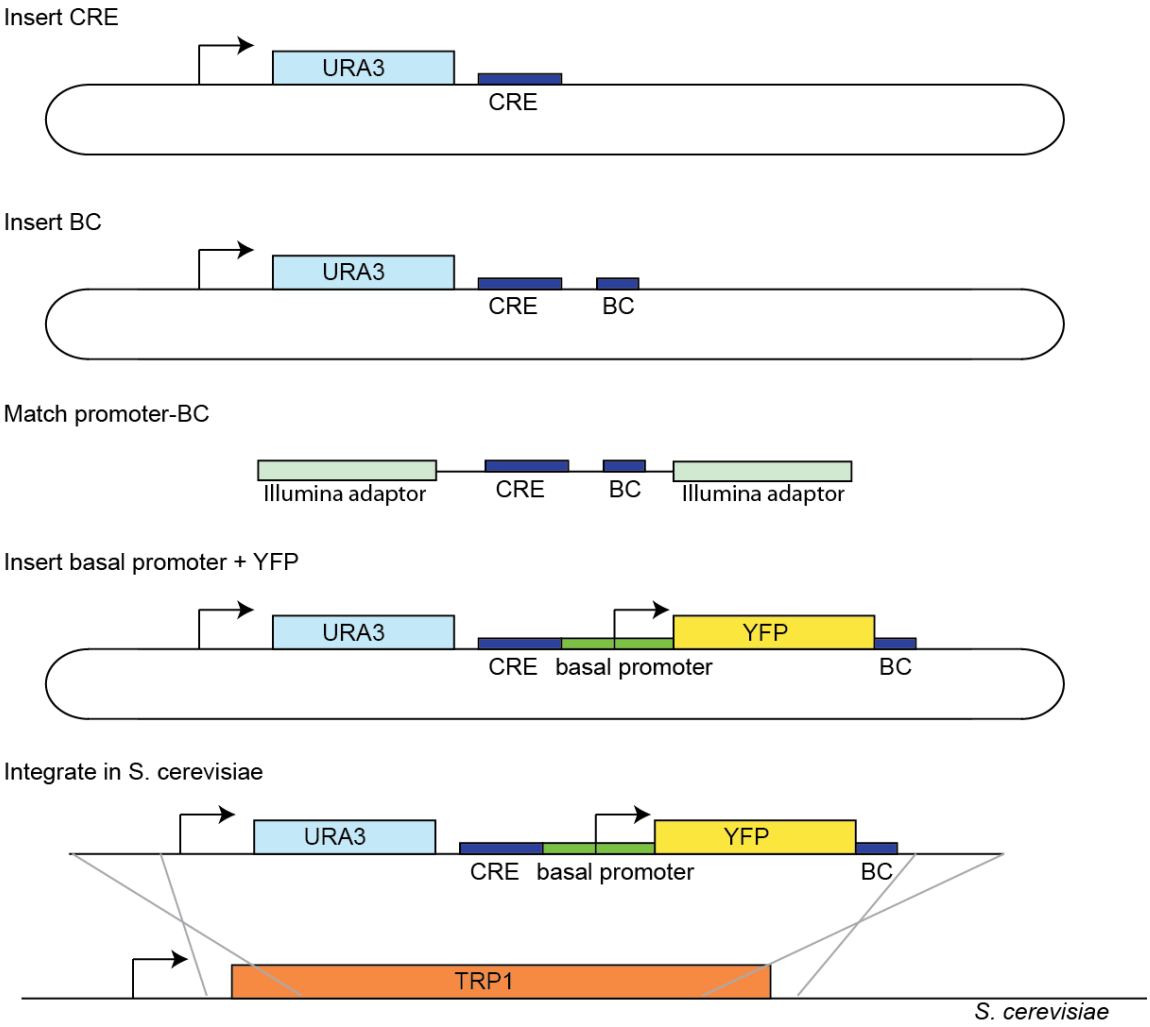


Figure A1.S1 The complete CRE-seq cloning strategy with chromosomal integration. First, we cloned the CREs into a plasmid creating a library of CRE sequences. Into this library we then cloned a library of BCs downstream of the CREs. To match CREs to BCs we digested away the plasmid backbone, ligated Illumina adapters, and sequenced the DNA with a paired end run. We then inserted the basal promoter and YFP gene between the CREs and the BCs. Finally, we integrated the library cassette into *S. cerevisiae* at the *TRP1* locus, and pooled colonies for use in the CRE-seq assay.

FIGURE A1.S2

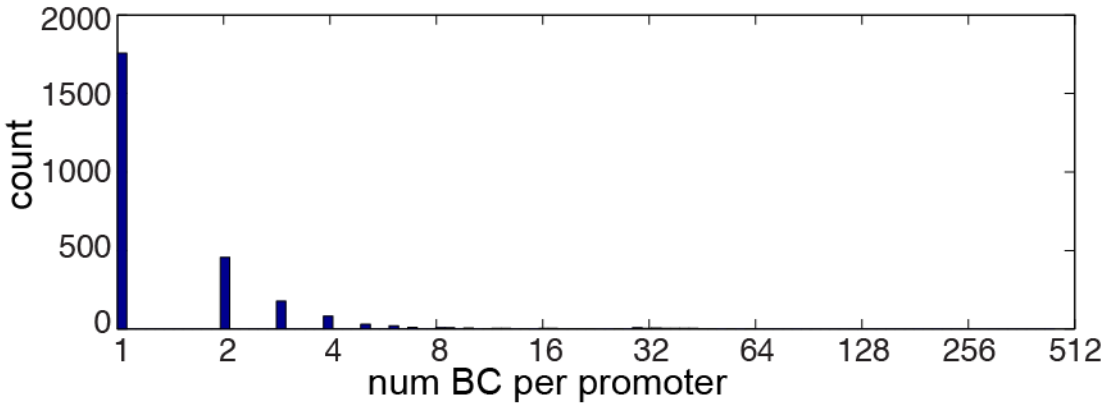


Figure A1.S2 Distribution of BC number per CRE. 66.7% of the CREs are represented by 1 BC only, and 16.1% of the CREs are represented by 3 or more BCs.

FIGURE A1.S3

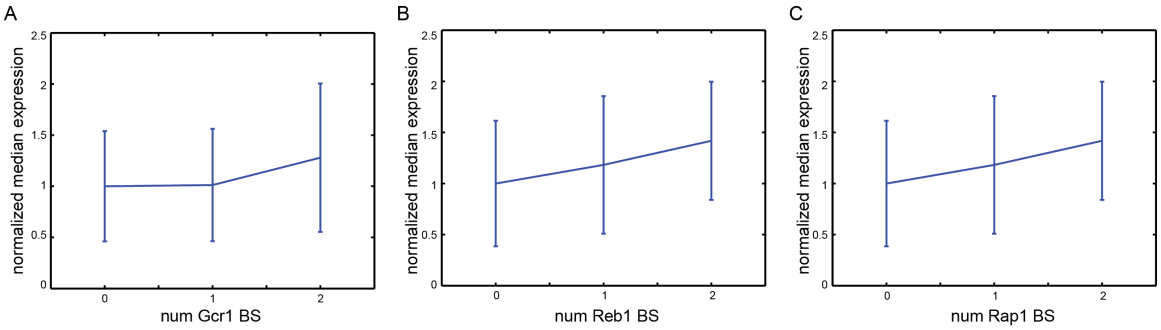


Figure A1.S3 The effect of 0, 1 and 2 Gcr1, Reb1 and Rap1 sites on gene expression. While expression (y-axis) grows linearly with the number of Reb1 and Rap1 sites (x-axis, panels B and C), Gcr1 shows a different behavior (panel A). We observe no increase over basal expression until two Gcr1 sites are present. This plot includes all promoters containing 0, 1, or 2 sites of interest, among other sites.

FIGURE A1.S4

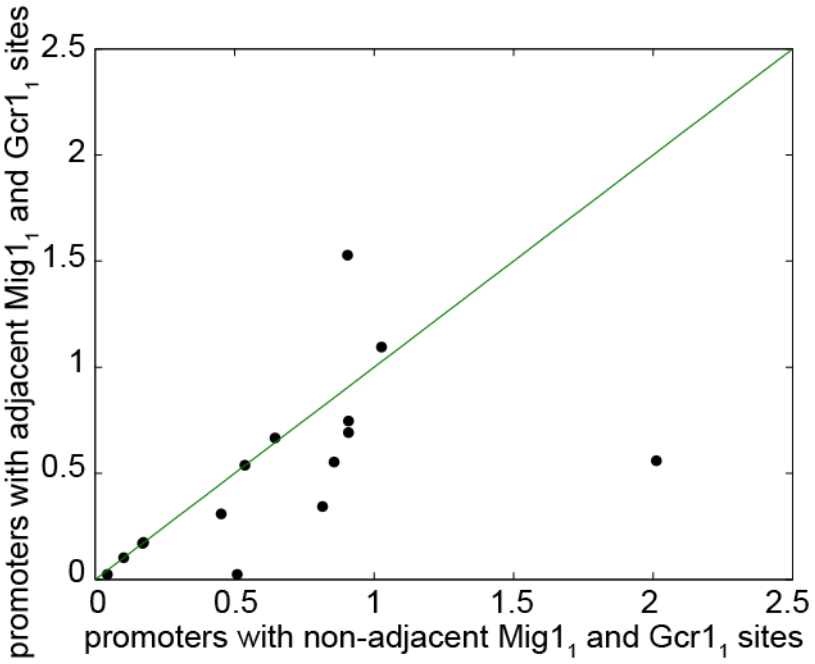


Figure A1.S4 Expression of promoters composed of the same TFBSs. The expression driven by CREs containing adjacent $Gcr1_1$ and $Mig1_1$ sites (y-axis), is, in general, lower than the expression driven by CREs with the same TFBS content but non-adjacent $Gcr1_1$ and $Mig1_1$ TFBSs (x-axis).

FIGURE A1.S5

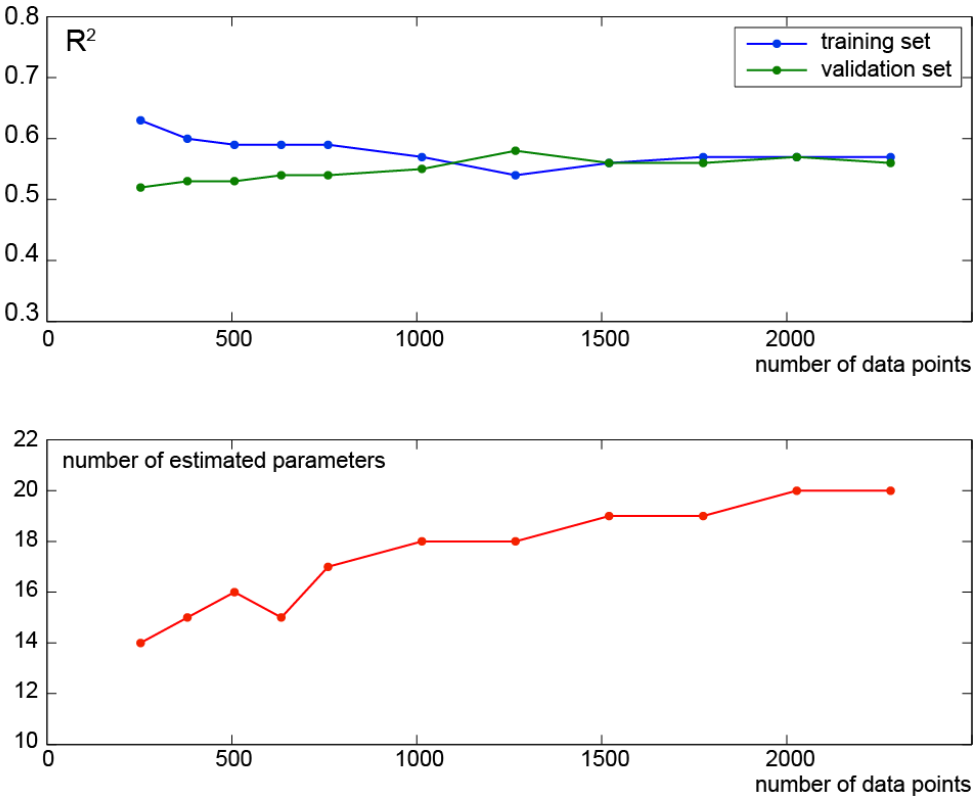


Figure A1.S5 Repeated random sub-sampling validation results. The 2534 data points were split in two sets, a training set and a validation set. The size of the training set varied; sets were defined by randomly picking 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80% or 90% of the total number of data points. The model was fit on the training set and then applied to the validation sets. R^2 values were estimated for the two sets. The R^2 values reported in the first panel of the figure are averaged across the 20 trials of this procedure. The figure shows that with 1000 unique data points (CREs), we have similar performances in the training set as in the validation set. With fewer data points the model is over-fit and predicts the validation set less well. The second panel shows that the number of significant parameters increases with the number of data points. When we include fewer than 2000 promoters we lose significant parameters, while the values of the others do not change significantly. This suggests that in order to recover weak behaviors we need to have a large number of unique promoters.

APPENDIX 2: HIGH-THROUGHPUT FUNCTIONAL TESTING OF ENCODE SEGMENTATION PREDICTIONS

Jamie C. Kwasnieski[1], Christopher M. Fiore[1], Hemangi G. Chaudhari, and Barak A. Cohen

[1]These authors contributed equally to this work
Center for Genome Sciences and Systems Biology, Department of Genetics, Washington University School of Medicine in St. Louis, MO, 63108

This work was done in collaboration with Chris Fiore, Hemangi Chaudhari, and Barak Cohen. Chris Fiore, Hemangi Chaudhari, Barak Cohen, and I conceived the project and designed experiments. Chris Fiore, Hemangi Chaudhari, and I performed the experiments. Chris Fiore and I did the data analysis. Chris Fiore, Barak Cohen, and I wrote the paper. This chapter is a manuscript currently under review.

ABSTRACT

The histone modification state of genomic regions is hypothesized to reflect the regulatory activity of the underlying genomic DNA. Based on this hypothesis, the ENCODE consortium measured the status of multiple histone modifications across the genome in several cell types and used this data to segment the genome into regions with different predicted regulatory activities (Dunham et al. 2012; Hoffman et al. 2013). We measured the *cis*-regulatory activity of more than 2000 of these predictions in the K562 leukemia cell line. We tested genomic segments predicted to be Enhancers, Weak Enhancers, or Repressed elements in K562 cells, along with other sequences predicted to be Enhancers specific to the H1 human embryonic stem cell (H1-hESC) line. Regions annotated as Repressed in K562 cells and Enhancer elements in H1-hESC did not show *cis*-regulatory activity in K562 cells greater than that produced by negative controls. In contrast, both Enhancer and Weak Enhancer sequences in K562 cells were more active than negative controls, although surprisingly, Weak Enhancer segmentations drove higher expression than Enhancer segmentations. Lower levels of the covalent histone modifications H3K36me3 and H3K27ac, thought to mark active enhancers and transcribed gene bodies, associate with higher expression and partly explain the higher activity of Weak Enhancers over Enhancer predictions, suggesting that our understanding of these particular modifications is incomplete. While DNase hypersensitivity (HS) is a good predictor of active sequences in our assay, transcription factor (TF) binding models need to be included in order to accurately identify highly expressed sequences. Overall our results support the notion that histone modification states reflect the *cis*-regulatory activity of sequences across the genome, but also suggest that specific sequence preferences such as transcription factor binding sites are the causal determinants of *cis*-regulatory activity.

INTRODUCTION

It is widely reported that specific combinations of covalent histone modifications reflect the regulatory function of underlying genomic DNA sequence (Strahl and Allis 2000). As part of the ENCODE project the genomic locations of a variety of covalent histone modifications were determined by chromatin immunoprecipitation sequencing (ChIP-seq) in a number of cell types and cell lines. Two studies used these data to train computational models that predict different functional regions of the human genome. These unsupervised learning algorithms, Segway (Hoffman et al. 2012) and ChromHMM (Ernst and Kellis 2010; Ernst and Kellis 2012), take functional genomics data as input (DNase-seq; FAIRE-seq; and ChIP-seq of histone modifications, RNA Polymerase II, and CTCF) and return segmentation classes, which are then assigned a hypothesized function using current knowledge of histone modification function. As part of the ENCODE project, these two sets of predictions were consolidated to create a unified annotation of the entire human genome with seven functional classes in multiple cell types. These segmentations include Transcription Start Site, Promoter Flanking, Transcribed, CTCF-bound, Enhancer, Weak Enhancer, and Repressed or Inactive segments (Dunham et al. 2012; Hoffman et al. 2013). If histone modifications accurately reflect the regulatory activity of their associated DNA, then these segmentation classes should have measurably different *cis*-regulatory activities.

In this study we tested whether the segmentation classes determined by ENCODE have different effects on gene regulation in their predicted cell type. We used CRE-seq, a massively parallel reporter assay, to determine whether 1) sequences in the Enhancer, Weak Enhancer and Repressed classes drive expression that is different from that produced by negative controls, 2) sequences in different segmentation classes drive different levels of gene expression, and 3) sequences control gene expression level consistent with their predicted segmentation labels. We find that segmentation predictions drive distinct levels of expression. In particular, enhancer predictions drive expression that is different than the expression levels driven by negative control sequences. We find that chromatin features can distinguish highly expressed sequences with some accuracy, but transcription factor binding preferences better identify the most highly expressed sequences.

RESULTS

CRE-seq Library and Measurements

We used a high-throughput multiplexed reporter assay (Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012) to characterize the regulatory activity of 2100 randomly chosen sequences annotated as Enhancer, Weak Enhancer, or Repressed. Specifically, we tested sequences with the following annotations in the K562 cell line: 600 Enhancer regions, 600 Weak Enhancer regions, and 300 Repressed regions. In order to test the cell-type specificity of the segmentation predictions, we also tested 600 Enhancer predictions from the H1-hESC cell line that are not annotated as Weak Enhancers or Enhancers in K562 cells.

We sought to establish an empirical null distribution as a negative control for activity in this assay, against which to compare the activities of sequences from the different segmentation classes. We randomly selected 284 sequences from each class of predictions and scrambled the nucleotide sequence of each while maintaining dinucleotide content, in order to preserve basic sequence features of the segment such as CpG frequency and nucleosome favoring signals. We designed our experiment to compare the expression distribution for each segmentation class to the expression distributions from their corresponding scrambled negative controls. Including predicted *cis*-regulatory elements (CREs) and scrambled negative controls, our final experimental design included 3237 distinct reporter gene constructs.

We used CRE-seq, a massively parallel reporter gene assay (Kwasnieski et al. 2012) to simultaneously measure the expression of all constructs. We first synthesized 13,000 unique 200-mer DNA sequences using array-based oligonucleotide (oligo) synthesis (LeProust et al. 2010). Each predicted CRE was replicated four times on the array, and each replicate was tagged with a unique nine basepair (bp) barcode, providing redundancy in the expression measurements. The 200 bp limit of oligonucleotide synthesis, along with the requirement to include priming sites and restriction enzyme sites, limited our tested CREs to 130 bp of each segmentation prediction. For the Enhancer and Weak Enhancer classes, we selected the entire region of 300 short (121-130 bp) genomic segments, and the central 130 bp of 300 longer genomic segments (>130 bp). As only a small fraction of Repressed segments are less than 130 bp in length, we tested only central sequences from this class. We chose the

center because it is an unbiased portion that does not incorporate additional histone or sequence features beyond the algorithms' output. This allows us to appropriately test the predictive power of the segmentations. Finally, we used the array-synthesized oligos to create a library of these CREs cloned upstream of the Hsp68 minimal promoter in which each reporter construct contains a unique sequence barcode in its 3' UTR (Kwasnieski et al. 2012). The resulting plasmid library was then transfected into K562 cells, and RNA was isolated after 22 hours.

To measure CRE activity, we quantified the level of each barcode in the transfected cells using RNA-seq, and normalized the RNA barcode counts by the abundance of each barcode in the plasmid DNA pool. The RNA/DNA ratio of barcode counts is a quantitative measure of the expression driven by each CRE in the library (Kwasnieski et al. 2012). We performed four independent transfections in K562 cells and found that our expression measurements are precise, displaying high reproducibility between biological replicates (Figure A2.1A, R^2 range: 0.95-0.97).

Expression of Segmentation Classes

We compared the activity of each class of segmentation prediction to the activity controlled by its matched scrambled distribution. We used two independent metrics to classify individual segmentations as "active" or "inactive". First, we computed the fraction of CREs within a segmentation class that drive expression higher than that of the 95th percentile of the matched scrambled expression distribution. Second, for each CRE we compared its sixteen replicate measurements (four barcodes per CRE in four independent experiments) with the distribution of the scrambled controls (Wilcoxon Rank Sum Test, $p<0.05$, Bonferroni correction with N=2100). Using these metrics, a significant number of Enhancer and Weak Enhancer CRE predictions are active (Figure A2.1B, A2.1C, Table A2.1). In contrast, neither the K562 Repressed regions nor the H1-hESC Enhancer regions show activity that is significantly different from their scrambled negative controls (Figure A2.1D, A2.1E, Table A2.1). Enhancer and Weak Enhancer regions show distinct levels of activity from both the K562 Repressed and H1-hESC Enhancer regions (Wilcoxon Rank Sum, $P<0.01$). Moreover, segmentations from the Repressed category did not repress expression below the 5th percentile of their matched scrambled controls, suggesting that these sequences are transcriptionally inactive and not repressive. We get the same results regardless of whether the

sequences are short segmentations included in their entirety, or longer predictions from which we included only the central 130 bp (Supplemental Figure A2.S1), indicating that our results are not biased by the method of choosing 130 bp sequences for testing. Taken together, we conclude that sequences annotated as Enhancer and Weak Enhancer segments have increased levels of activity over their corresponding null distributions, and that different segmentation classes produce distinct median levels of activity in our assay.

Unexpectedly, we found that sequences classified as Weak Enhancers drive a higher median level of activity than sequences classified as Enhancers (Supplemental Figure A2.S2, P=3.7e-4 by Wilcoxon Rank Sum). The difference between the two classes is even greater when comparing the fraction of CREs we designated as “active” relative to their matched scrambled sequences. (Table A2.1). Compared to Weak Enhancers, segmentations in the Enhancer class have higher GC content (Supplementary Figure A2.S3B), a sequence feature associated with higher *cis*-regulatory activity (Landolin et al. 2010; Lidor Nili et al. 2010; White et al. 2013). Indeed scrambled sequences derived from the Enhancer class drive higher expression than scrambled sequences from the Weak Enhancer class (Supplementary Figure A2.S3A). Therefore, despite having higher GC content, a feature associated with higher expression, the Enhancer predictions drive lower expression than the Weak Enhancer predictions. This suggests that some additional determinant is responsible for the higher activation of segments labeled as Weak Enhancers.

We asked whether differences in covalent histone modifications might explain the difference in expression between Weak Enhancers and Enhancers. We compared the levels of all histone modifications (Hoffman et al. 2013) that were measured in K562 cells between the two classes. Weak Enhancers were segmented from Enhancers by their lower levels of the histone modification H3K27ac (Creyghton et al. 2010) (Figure A2.2B), thought to signify active enhancers, and H3K36me3 (Barski et al. 2007) (Figure A2.2D), often thought to signify a transcribed gene body but recently also found in silenced genes (Chantalat et al. 2011). Surprisingly, lower levels of both of these covalent histone modifications are associated with higher expression of enhancers in our assay (Wilcoxon Rank Sum Test, $p<10^{-5}$, Figure A2.2A, A2.2C), even within the Enhancer or Weak Enhancer classes (Supplemental Figure A2.S4). We did not find an association of H3K27ac signal in the larger context (up to 500bp surrounding the

selected regions) or H3K27ac dip score with active enhancers (Kheradpour et al. 2013) (data not shown). Thus, Weak Enhancers may have more activity than Enhancers in part because they have lower enrichment of H3K27ac and H3K36me3, which associate with higher activity in our assay. These histone modifications do not fully explain the expression differences between these two classes indicating that other sequence features must explain the higher activity of Weak Enhancers.

Sequence and Chromatin Features

We searched for sequence and chromatin features that could predict activity across all segmentation classes in our assay. Two primary sequence features (GC content and minor groove width as estimated by ORChID2 (Rohs et al. 2009; Bishop et al. 2011) score) and six chromatin features (Dunham et al. 2012; Hoffman et al. 2013) (DNase HS from Duke; DNase HS from University of Washington [UW]; Faire-seq; and ChIP-seq of H3K4me1, H3K36me3, and RNA Pol2) are significantly enriched in sequences that drive high expression in our assay (Wilcoxon Rank Sum test, $P<0.05$ corrected with Bonferroni method, N=16). We used these data to develop a quantitative model that distinguishes active CREs from inactive CREs. Of these eight features, DNase HS (UW) signal best separated the active from inactive sequences (Figure A2.3A, A2.3B, AUC=0.685), suggesting that DNA accessibility is a good indicator of the *cis*-regulatory potential of a sequence (Thurman et al. 2012). No other single feature performed as well as DNase HS signal. A logistic regression model with the above-mentioned six chromatin features and two primary sequence features (PSF), improves the classification of active sequences (Figure A2.3A, AUC=0.733), but only marginally above that of DNase HS alone. We therefore hypothesized that additional sequence-specific binding features, such transcription factor binding motifs, may better explain expression.

We investigated whether the inclusion of transcription factor (TF) binding specificities improved our ability to explain the expression differences we observed in our assay. Using several libraries of TF binding models (Newburger and Bulyk 2009; Jolma et al. 2013; Mathelier et al. 2013), we searched for motifs enriched or depleted in activated CREs and found 50 significant, non-redundant motifs. A logistic regression model that incorporated these binding models performs better at distinguishing active sequences than the chromatin and PSF model (Figure A2.3A, AIC (Akaike 1974): 1881 vs. 1729 for

model with motifs; AUC=0.802). We performed 5-fold cross-validation on all of the models and observed little decrease in predictive power, suggesting that our model is not over-fit (Supplemental Table A2.S2). The predicted motif for Activator Protein 1 (AP1), a heterodimer of TFs in the FOS and JUN families(Hess et al. 2004), is the most significantly enriched motif in highly expressed CREs. The expression driven by CREs with predicted AP1 motifs is significantly higher than the expression driven by sequences without the motif (Figure A2.3C, $\log_2$ ratio of 0.96, $p<2.2x10^{-16}$). Furthermore, highly expressing CREs are significantly enriched for sequences that are bound by FOS and JUN family TFs in K562 cells (Consortium 2011). (Figure A2.3D; $p=8.8x10^{-10}$ by Fisher's exact test, odds ratio=4.2). These data suggest that AP1 is responsible for the activity of many enhancers in K562 cells, as previously reported (Muthukrishnan and Skalnik 2009; Kheradpour et al. 2013), and, as a consequence, the enhancers' histone modification state.

DISCUSSION

In this study we directly tested the *cis*-regulatory activity of segmentation predictions based on histone modification data from the ENCODE project. We found that these predictions were cell type-specific in K562 cells and could accurately distinguish enhancer sequences from non-enhancer sequences. Our results suggest that combinations of TF binding preferences, not histone modifications alone, are most predictive of actively expressing genomic sequences, a result supported by other attempts to define the sequence features of enhancers (Heinz et al. 2010; Lee et al. 2011; Arvey et al. 2012; Gorkin et al. 2012; Smith et al. 2013). These results support a model where TF binding and subsequent transcriptional regulation configure the immediate chromatin environment (Struhl and Segal 2013), leading to the constellation of histone modifications observed in segments with high *cis*-regulatory activity. However, even our model incorporating all of the available features is only moderately predictive (AUC=0.84) and cannot quantitatively predict expression level. This suggests that more complex features determine the quantitative expression levels controlled by enhancers.

We conclude that the Repressed segmentation class consists mostly of sequences with no transcriptional activity rather than *cis*-regulatory sequences that actively repress transcription. We have previously shown transcriptional repression by short enhancers (White et al. 2013), indicating that the length of CREs we tested cannot explain this result. There are two possible explanations for why we did

not see repression in this assay. First, the Repressed segmentation class contains mostly sequences with predicted low activity by either the ChromHMM or Segway algorithms, with only a small fraction of the sequences predicted to have repressive activity by these algorithms. Second, it is possible that we are unable to predict combinations of histone modifications that signal repression such that no segmentation successfully defines repressive activity. Because a large fraction of regulated gene expression works through the activity of transcriptional repressors, identifying combinations of histone modifications that reflect repression is still an important challenge.

Finally, we conclude that combinations of histone modifications identify functional enhancers, but our interpretation of these combinations needs to be refined. In particular, high levels of the covalent histone modifications H3K27ac and H3K36me3 are thought to mark active enhancers and transcribed gene bodies or even heterochromatic regions (Barski et al. 2007; Creyghton et al. 2010; Chantalat et al. 2011). Among segments marked as Enhancers or Weak Enhancers, lower enrichment of these modifications is found at segments with high activity in this assay. This finding suggests that the precise function of these modifications needs to be explored, as it is clear that there is no simple linear relationship between the level of these modifications and expression.

MATERIALS AND METHODS

CRE-seq Library Construction

A pool of 13,000 unique 200-mer oligos was ordered through a limited licensing agreement with Agilent Technologies. Oligos were structured as follows: 5' priming sequence (GTAGCATCTGTCC)/NheI site/CRE/HindIII site/XhoI site/SphI site/ barcode/SacI site/3' priming sequence (CGACTACTACTACG). A more detailed diagram of array sequence is provided in Supplemental Figure A2.5.

The plasmid library was prepared as described (Kwasnieski et al. 2012), except using primers CF166 and CF167 (Supplemental Table 1) and an annealing temperature of 57C. The amplified library product was purified on a polyacrylamide gel as described (White et al. 2013). The library plasmid backbone, CF10, was created from the plasmid pGL4.23, by cloning dsRed-Express2 between the Acc65I and FseI sites. Purified library amplicons were cloned into CF10 using NheI and SacI. We prepared DNA from 100,000 colonies to generate PL7_1. We then cloned the Hsp68 promoter driving DsRed into PL7_1. A cassette containing the Hsp68 promoter was amplified from pGL-hsp68 with primers CF121 and CF168 (Supplemental Table 1). pGL-hsp68 was created by amplifying the hsp68 promoter from hsp68LacZ (kind gift of M. de Bruijn, Oxford Stem Cell Institute, Oxford, UK) using primers JKO25F and JKO25R (Supplemental Table 1). The hsp68 DsRed amplicon was cloned into library PL7_1 by using HindIII and SphI, creating library PL7_2.

Cell culture and Transfection

K562 cells were maintained in Iscove's Modified Dulbecco's Medium (IMDM) medium with 10% Fetal Bovine Serum and 1% Amino Acids (Life technologies). The plasmid library was purified by phenol-chloroform extraction and ethanol precipitation before transfection. The Neon transfection system (Life technologies) was used to transfect the plasmid library. For each replicate, 1.2 million cells were pelleted by centrifugation, washed with PBS and resuspended in 100µl of Buffer R. 27µg plasmid library DNA along with 3 µg of pMax-GFP as a positive control was transfected into the cells by using three 10ms pulses at 1450V. The transfected cells were seeded into T-25 flasks with 5ml of the growth medium and incubated at standard conditions. Transfection efficiency was greater than 90% (data not shown).

Selection of Segmentation Predictions

Segmentation predictions (Dunham et al. 2012; Hoffman et al. 2013) were downloaded from the ENSEMBL genome browser and converted to UCSC notation. We filtered predictions that overlapped with the ENCODE DAC Blacklisted Regions (http://moma.ki.au.dk/genome-mirror/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability) or Repeat Masker regions (http://www.repeatmasker.org/species/homSap.html).

We also removed predictions that contained restrictions site sequences that we intended to use for cloning sequences into a plasmid library. To select H1-hESC Enhancer predictions, we removed H1-hESC Enhancer predictions that overlapped with K562 Enhancer or Weak Enhancer predictions. Next we sorted predictions by chromosome, and separated them by length into long (>130 bp) and short (121-130 bp). To choose the predictions to test, we selected lines of this file at regular intervals, so the tested CREs span all chromosomes of the human genome.

Preparing Samples for RNA-Seq

RNA was extracted from K562 cells 22 hours after transfection using the PureLink RNA mini kit (Life Technologies) and then excess DNA was removed using the TURBO DNA-free kit (Applied Biosystems), following manufacturer's instructions. First strand cDNA was synthesized from the RNA using SuperScript II Reverse Transcriptase (Life Technologies). Both the cDNA samples and the DNA from the original plasmid library were prepared for sequencing using a custom protocol as described(Kwasnieski et al. 2012). Briefly, we used PCR amplification of the sequence surrounding the barcode in the RNA transcript or plasmid using primers CF150 and CF151b (Supplemental Table 1). We then digested the PCR product using SphI and XhoI and ligated Illumina adapter sequences (MO576/582, MO577/583, MO578/584, MO579/585, Supplemental Table 1) to these amplified sequences. Two lanes of the Illumina HiSeq machine were used to sequence this barcode region from the cDNA and DNA, and reads that perfectly matched the first 13 expected nucleotides were counted, regardless of quality score. This resulted in 77.5 million reads from the cDNA, across 4 biological replicates, and 34.8 million reads from the DNA. Only barcodes with >=50 reads in the DNA pool and >=3 reads in the cDNA pool were

used for downstream analysis. The expression of each barcode was calculated as (cDNA reads)/(DNA reads) and then normalized to the expression of the basal promoter alone. The expression of each CRE was calculated as the mean of the expression of each BC associated with it.

Data sources

We used the normalized chromatin ChIP-seq, Faire-seq, and DNase-seq data used in the integrated segmentation of the genome by Hoffman et al. (Hoffman et al. 2013), which can be accessed at https://sites.google.com/site/anshulkundaje/projects/wiggler. These included (all from K562 cell line): CTCF, Duke DNase, UW DNase, Faire, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H4K20me1, Pol2, and Control. This data was produced by the ENCODE consortium (Dunham et al. 2012). The signal associated with each CRE we analyzed was the average signal over that segment.

The TF binding matrices were taken from three databases: JASPAR vertebrate (146 matrices) (Mathelier et al. 2013), uniPROBE (757 matrices) (Newburger and Bulyk 2009), and high-throughput SELEX (820 matrices) (Jolma et al. 2013). FIMO (Grant et al. 2011) was used to find binding sites in the CREs used in the assay (both genomic and scrambled), using the default options with a p-value threshold of 10^{-4}. The AP-1 binding matrix that was enriched in highly expressed sequences in our assay was from JASPAR (MA0099.2).

TF ChIP-seq data was obtained from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/.

GC-content and ORChID2 (Rohs et al. 2009; Bishop et al. 2011) scores were calculated from the nucleotide sequences of the CREs.

Logistic regression models

A logistic regression model was developed to predict sequences activated over the scrambled 90^{th} percentile. The parameters for the model were chosen from a filtered list of available genomic data and sequence features. Each of the three sets of parameters was filtered separately: histone data including primary sequence features (GC-content and ORChID2 scores), binding matrices, and a set of

peaks from TF ChIP-seq. Those scores that had a significantly different distribution of values in the active CREs (expression greater than the 90th percentile of the matched scrambled distribution) vs. the inactive CREs passed the filter. For the parameter set with histone data and primary sequence features (PSF) and the parameter set with binding matrices we used Wilcoxon rank sum test ($p<0.05$, corrected using Bonferroni with N=16 for histone and N=1687 for binding matrices). For the TF ChIP-seq peak data (which is in binary form) we used Fisher's exact test ($p<0.05$, corrected using Bonferroni with N=16). 73 binding matrices, 8 histone with PSF parameters (including GC-content and ORChID2 scores), and 8 TF ChIP-seq parameters passed the filter. The binding matrices were further filtered to remove ones that showed nearly identical binding patterns across the CREs (>=99% similar), resulting in 50 binding matrices.

A logistic regression model for predicting actively expressed CREs was created for each of the three sets of parameters separately and with all sets of parameters together (66 total parameters). Only additive terms were used. We then created receiver operating characteristic (ROC) curves attempting to correctly predict the activated CREs (over 90th percentile of the matched scrambled distribution). The area under the curve (AUC) was calculated for each model as well as the best performing histone parameter (UW DNase), GC-content, and ORChID2 scores. Additionally, five-fold cross validation was used to ensure our models were not over-fit. The CREs were split into five training groups, and the model was trained on the data holding out each group in turn (beginning with the filtering of the parameters) and tested on the group held out. AUC was calculated for each of these sets, and the mean AUC from the five sets was calculated (Supplementary Table 2).

Figure A2.1

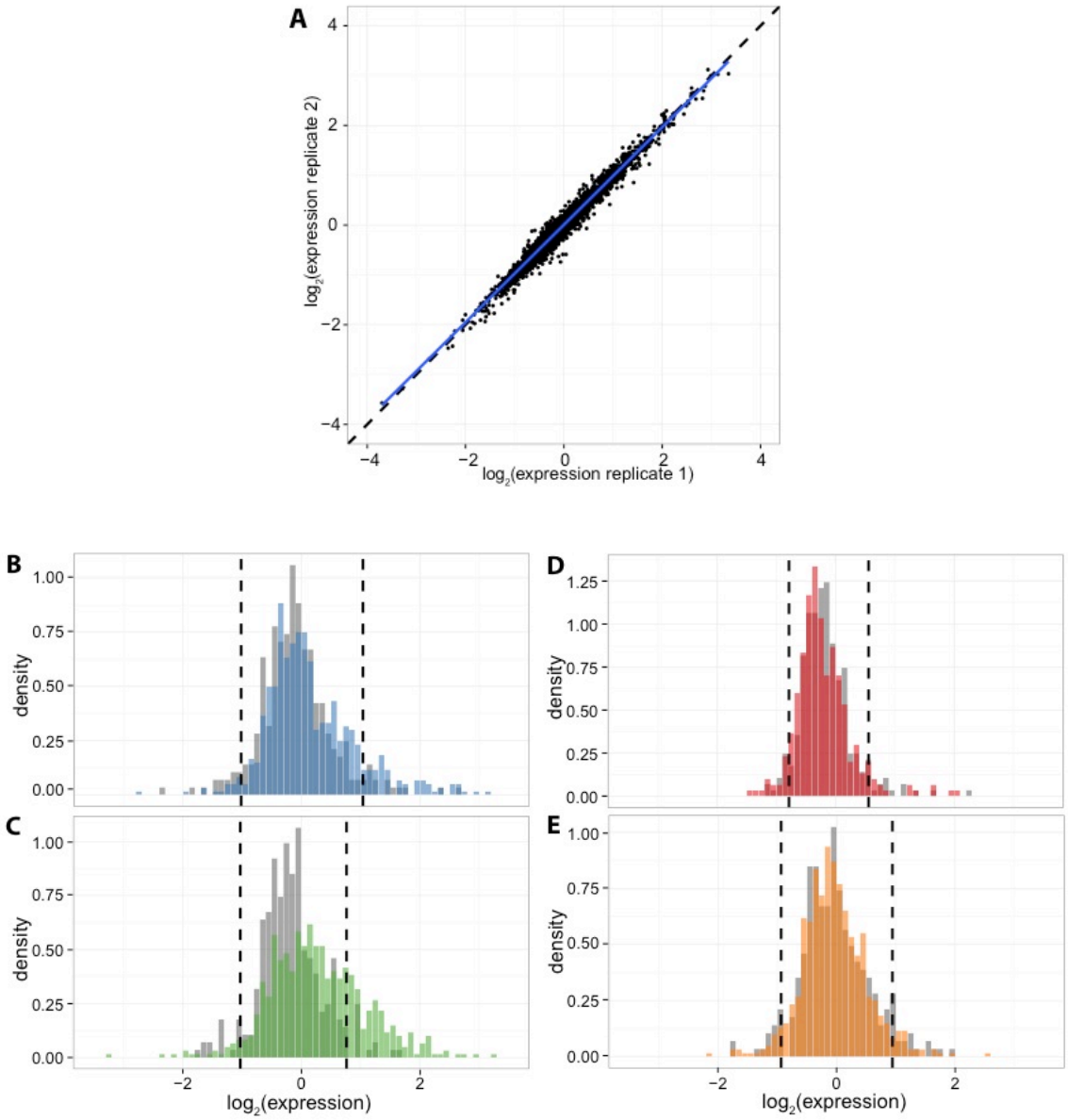


Figure A2.1. Reproducible expression measurements show differences in expression by segmentation class. A) Representative scatterplot showing expression of each CRE in two biological replicates (R^2=0.95, range of R^2 between all replicates: 0.95-0.97). Dashed black line is line of equality and blue line is best fit. **B-E)** Histograms of genomic CRE expression measurements in K562 cells. Each class is compared to scrambled controls with equivalent GC and dinucleotide content (grey). Dashed lines are the 5th and 95th percentiles of the scrambled distributions. **B)** K562 Enhancer class (blue), **C)** K562 Weak Enhancer class (green), **D)** K562 Repressed class (red), **E)** H1-hESC Enhancer class (orange).

Figure A2.2

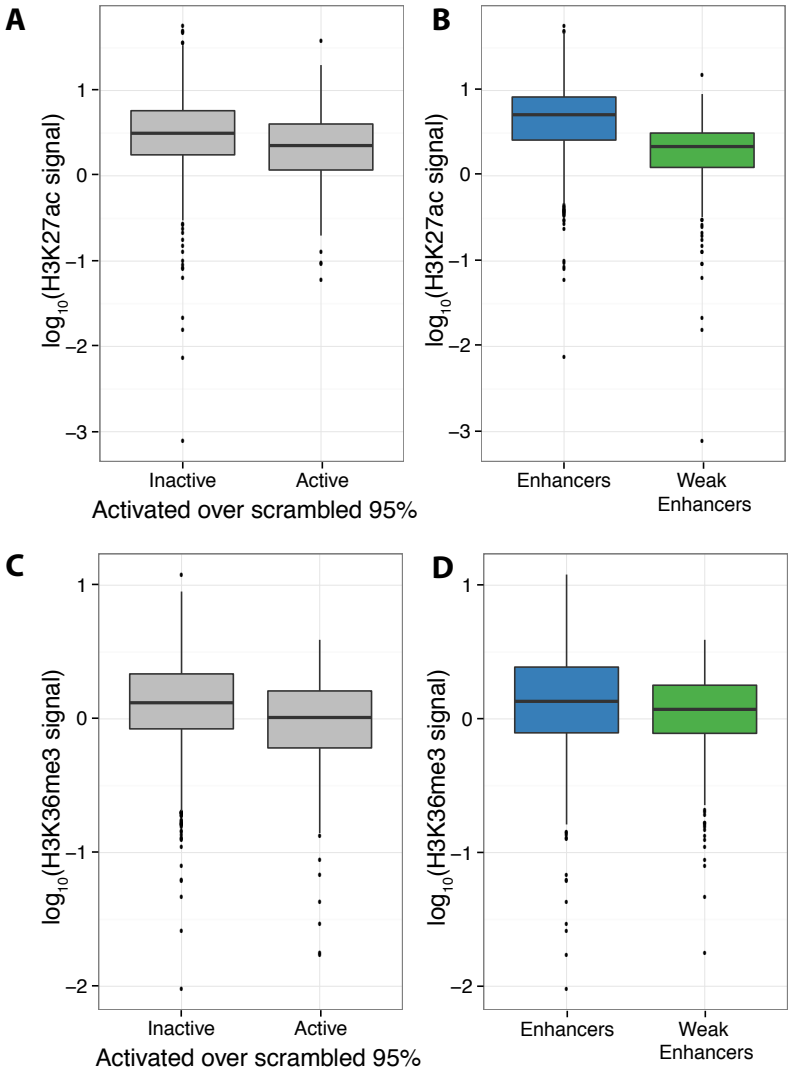


Figure A2.2. Lower H3K27ac and H3K36me3 signal are associated with higher Weak Enhancer expression. Boxplots showing that **A)** H3K27ac signal and **C)** H3K36me3 signal are depleted in active CREs compared to inactive CREs. **B)** H3K27ac signal and **D)** H3K36me3 signal are also depleted in Weak Enhancers (green) compared to Enhancers (blue). Active CREs are those above 95th percentile of scrambled distribution (Table 1).

Figure A2.3

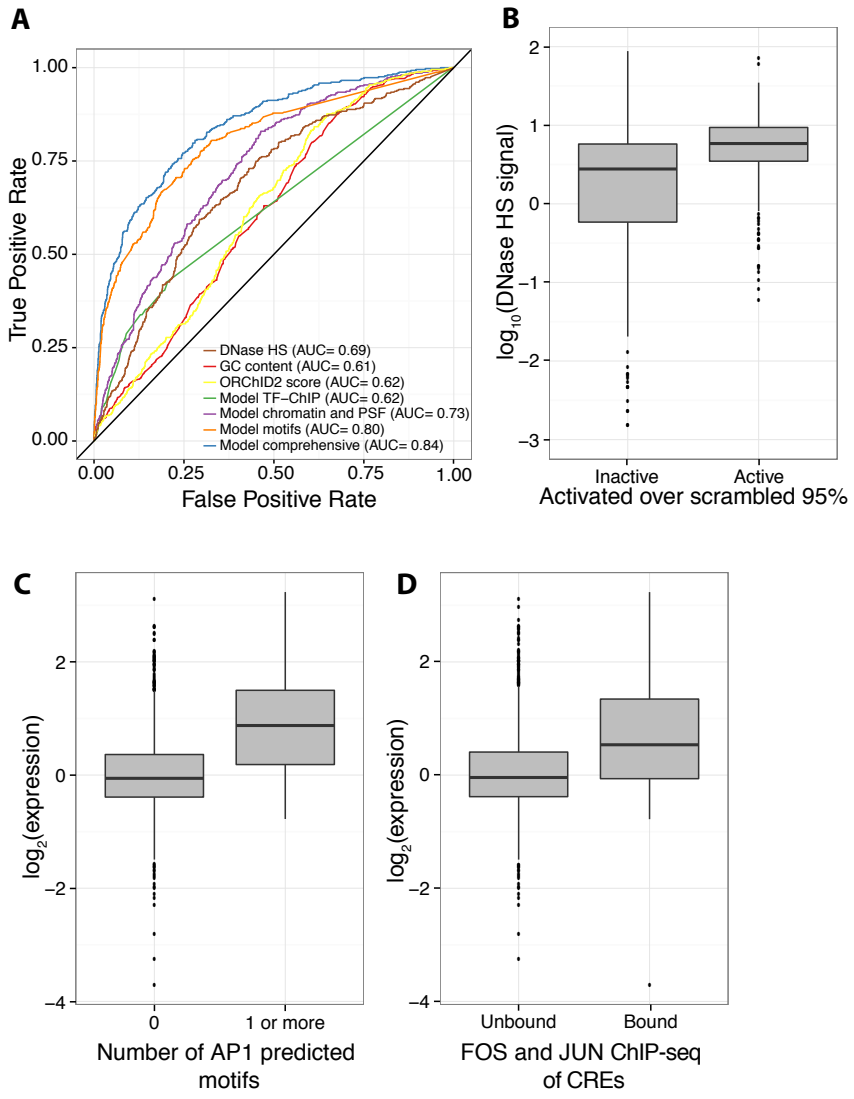


Figure A2.3. Chromatin features and sequence-specific binding identify active sequences. **A)** Receiver operating characteristic (ROC) curve shows that a logistic regression model ("Model comprehensive") incorporating sequence-specific binding motifs, chromatin features, primary sequence features (PSF), and TF ChIP data is best able to identify active sequences. Of logistic regression models with fewer features, one with sequence-specific binding motifs ("Model motifs") does best, followed by a model incorporating chromatin and primary sequence features ("Model chromatin and PSF"), a model with only significant TF-ChIP features ("Model TF-ChIP"). Minor groove width as predicted by ORChID2 score, GC content and DNase HS are also shown. Area under the curve (AUC) is indicated in legend. **B)** Boxplot showing that active CREs are enriched in high DNase HS signal over inactive CREs. **C)** Boxplot showing that CREs with at least 1 predicted AP-1 motif drive higher expression than CREs with no AP-1 predicted motifs. **D)** CREs overlapping with ChIP-seq peaks for a FOS (FOS or FOSL1) family member and a JUN (JUNB or JUND) family member, the constituent proteins of AP-1, drive higher expression than unbound CREs.

Table A2.1

Segmentation Prediction	Active over 95% scrambled	Active by Wilcoxon
K562 Enhancer	11.3% [5%]	26.8% [13.4%]
K562 Weak Enhancer	25.7% [5%]	39.7% [15.5%]
K562 Repressed	5.35% [5%]	7.00% [8.45%]
H1-hESC Enhancer	4.34% [5%]	12.0% [16.6%]

Table A2.1. Percentage of active CREs by segmentation class. For each ENCODE segmentation class, the table shows the percentage of all genomic CREs that are active with the percentage of matched scrambled controls that are active in square brackets. Activation was determined by comparing CRE expression to the 95^{th} percentile of matched scrambled controls (Active over 95% scrambled) or by statistically comparing replicate measurements of expression to matched scrambled control distribution (Active by Wilcoxon, Wilcoxon Rank Sum test, $P<0.05$, corrected using Bonferroni method with N=2100 for genomic CREs and N=1136 for scrambled control CREs).

SUPPLEMENTARY FIGURES AND TABLES

Table A2.S1

Name	Sequence
Primer CF121	TAGCGTCGAGGACATCAAGA
Primer CF150	TACACCGTGGTGGAGCAGTA
Primer CF151b	AGCGTACTCGAGTTGTTAACTTGTTTATTGCAGCTT
Primer CF168	ATGCATGCCTAGAATTACTACTGGAACA
Primer JKO25F	CATCAAGCTTCTCCTCCGGCTCGCT
Primter JKO25R	CGTTGTAAAACGACGGGATC
Adapter MO576	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC TCTTCCGATCTGCTCGATCATG
Adapter MO577	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC TCTTCCGATCTTAGACTATCATG
Adapter MO578	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC TCTTCCGATCTCGCTACCCTCATG
Adapter MO579	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC TCTTCCGATCTATAGTGGACACATG
Adapter MO582	ATCGAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCT CGGTGGTCGCCGTATCATT
Adapter MO583	ATAGTCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATC TCGGTGGTCGCCGTATCATT
Adapter MO584	AGGGTAGCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGAT CTCGGTGGTCGCCGTATCATT
Adapter MO585	TGTCCACTATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGA TCTCGGTGGTCGCCGTATCATT

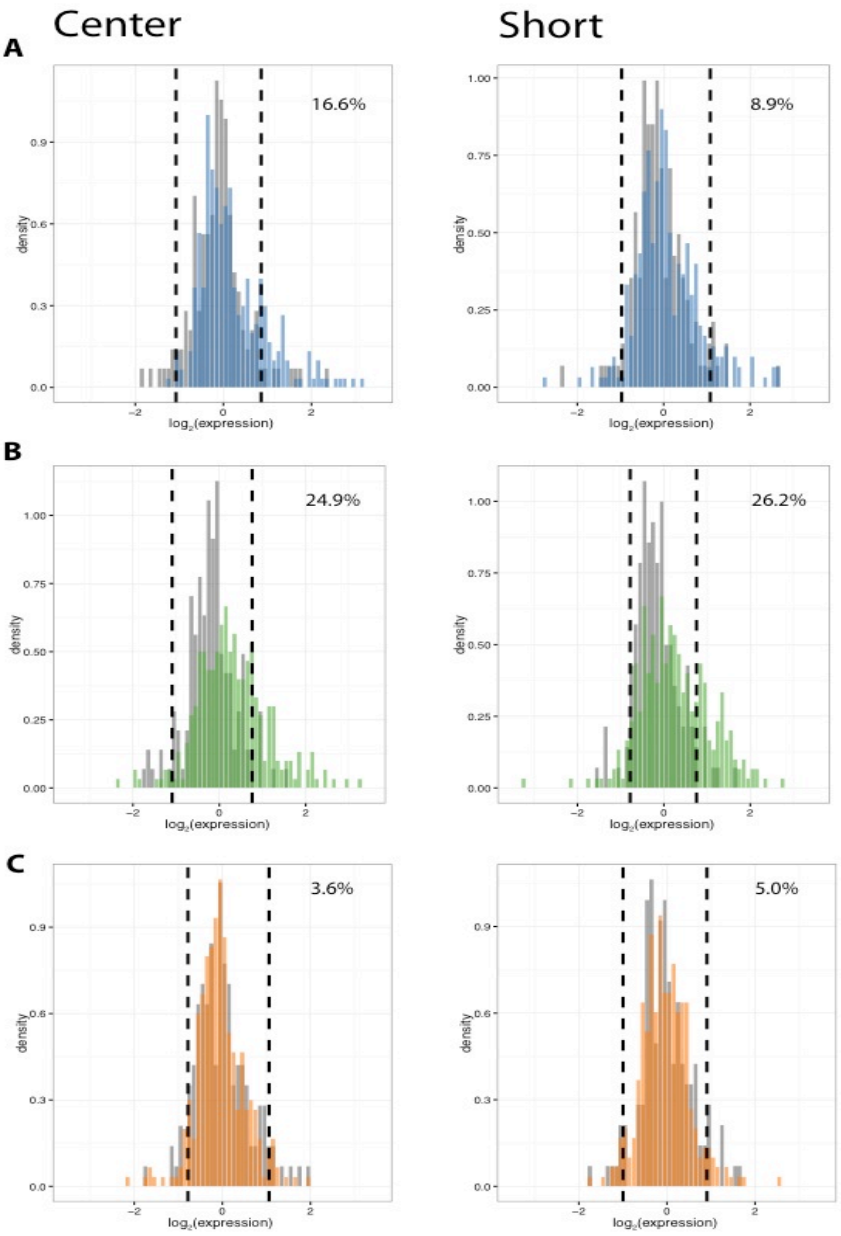
Supplemental Table A2.S1: Oligonucleotide sequences used in this study.

Table A2.S2

Model	Num of Parameters	AIC	AUC	CV mean AUC
Histone and PSF	8	1881.5	0.733	0.722
TF ChIP	8	2007	0.622	0.5974
Binding	50	1728.9	0.802	0.758
Comprehensive	66	1643.3	0.841	0.798

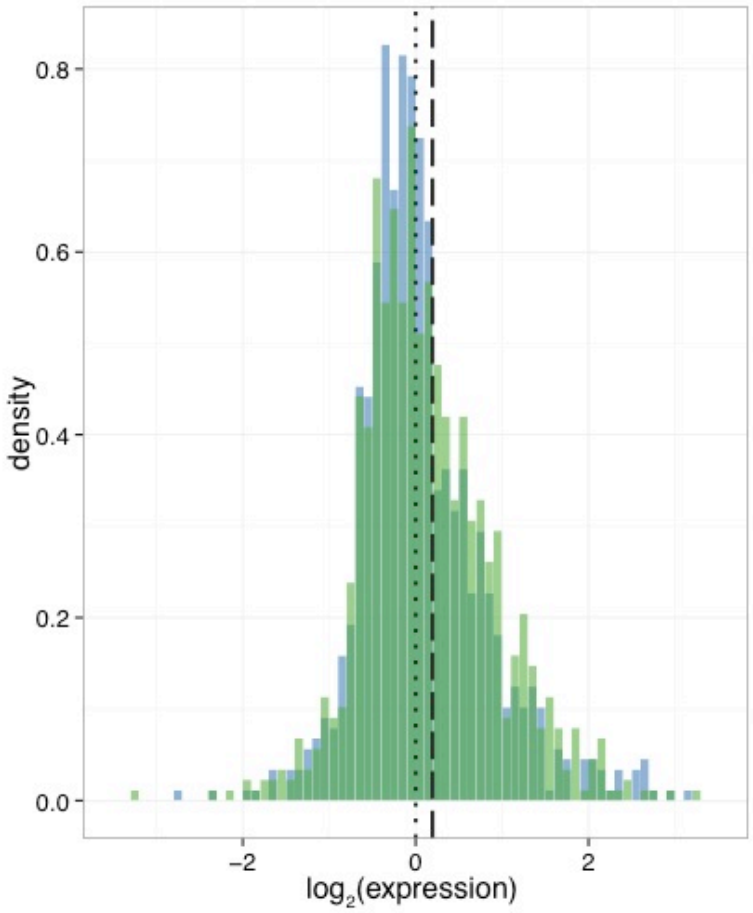
Supplemental Table A2.S2: Logistic Regression Models. The four logistic regression models used to predict active CREs from our assay: Histone and primary sequence features (PSF) included histone features, GC-content, and ORChID2 score; TF ChIP included peaks from TF ChIP-seq; binding included predicted binding models for TFs; and comprehensive included parameters from all of the models. Akaike Information Criteria (AIC), area under the curve (AUC) using the full data for training and testing, and the mean AUC from 5-fold cross-validation (CV) is listed for each model.

Figure A2.S1



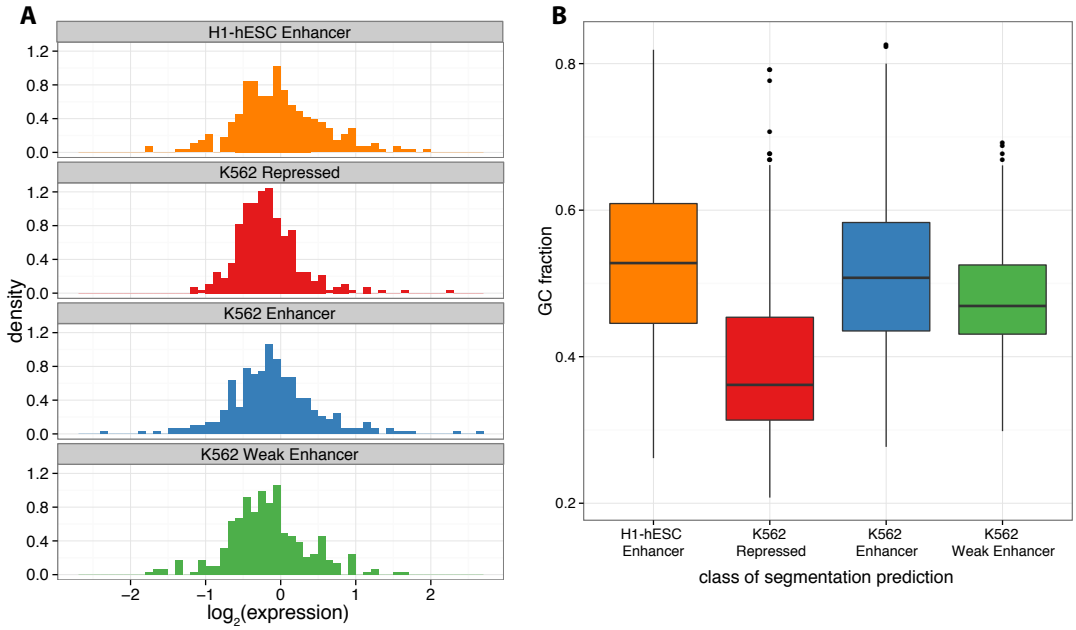
Supplementary Figure A2.S1. Computing the fraction of active sequences does not depend on the method of choosing short segments. Histograms showing the distribution of expression for each class; **A)** Enhancers (blue), **B)** Weak Enhancers (green), **C)** H1-hESC Enhancers (orange); either for the sequences from the center of longer segments ("Center"), or from whole short segments ("Short"), compared to their matched scrambled controls (grey). The dashed lines indicate the 5th and 95th percentiles of the scrambled distribution. The percentage of elements with expression greater than the scrambled 95th percentile is indicated.

Figure A2.S2



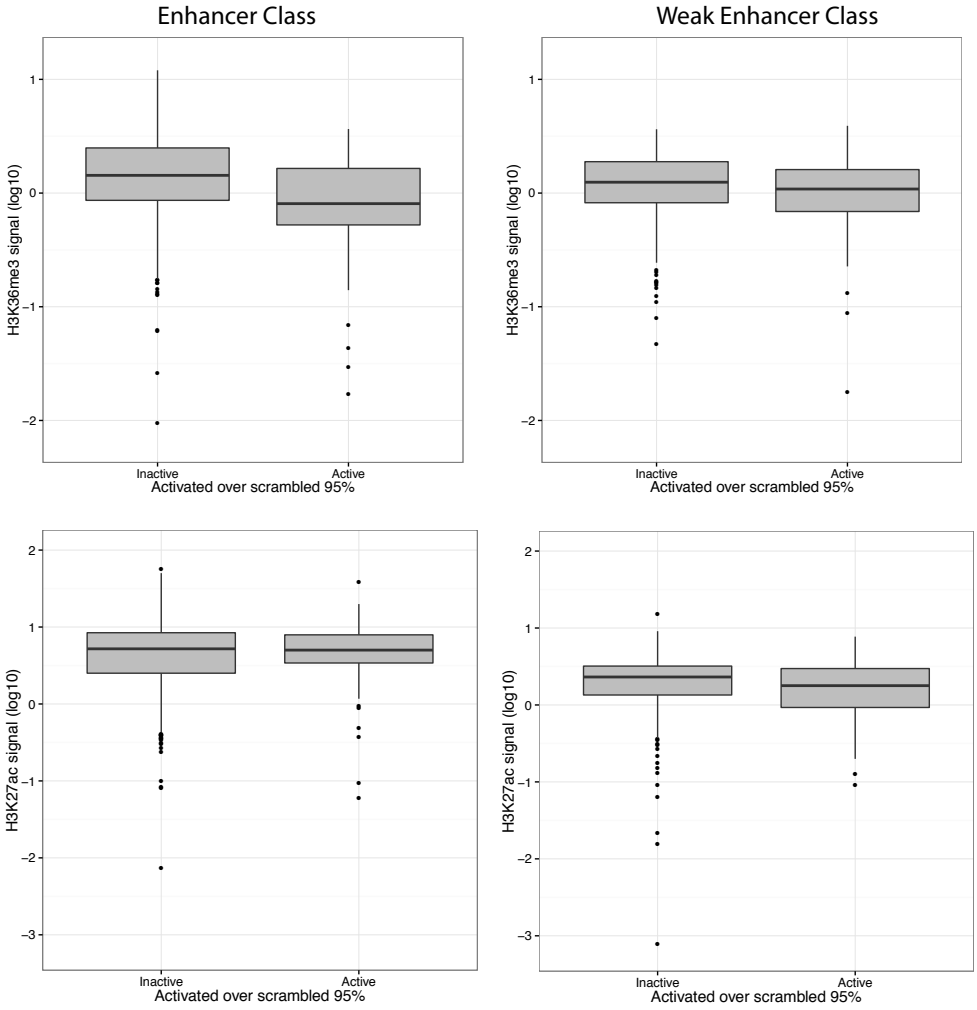
Supplementary Figure A2.S2. Weak Enhancers control higher median expression than Enhancers. Histogram of expression measurements, showing the distribution of Weak Enhancers (green) shifted to the right of that of Enhancers (blue). Lines show median expression for Enhancers (dotted) and Weak Enhancers (dashed).

Figure A2.S3



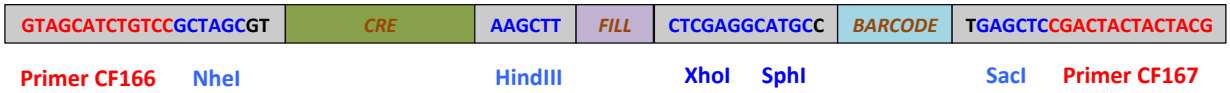
Supplementary Figure A2.S3. Expression and GC Fraction of Scrambled CREs. A) Histograms showing the expression controlled by each set of scrambled sequences. **B)** Boxplots show the distribution of GC fraction for each category of scrambled sequences.

Figure A2.S4



Supplementary Figure A2.S4. Active Weak Enhancer and Enhancer CREs have lower levels of H3K36me3 and H3K27ac. Boxplots showing that H3K27ac signal and H3K36me3 signal are depleted in active CREs compared to inactive CREs. Plots are similar to 2A and 2C except data is separated by segmentation class. Active CREs are those above the 95th percentile of the scrambled distribution (Table 1).

Figure A2.S5



Supplementary Figure A2.S5. Diagram of 200-mer Oligos Used to Construct the CRE-seq Library. Red sequences are used for PCR priming and blue sequences are restriction enzyme sites. The *CRE* is the 121-130bp sequence from the genomic predictions or scrambled controls. The *FILL* is 0-9bp of random sequence to bring the length of the whole oligo sequence up to 200bp. The length of the FILL sequence is calculated as 130-length of CRE. The *BARCODE* is a 9bp sequence that will label the 3' UTR of the mRNA transcript.

REFERENCES

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716-723.

Akimoto M, Cheng H, Zhu D, Brzezinski JA, Khanna R, Filippova E, Oh EC, Jing Y, Linares JL, Brooks M et al. 2006. Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc Natl Acad Sci U S A* **103**(10): 3890-3895.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*.

Arnosti DN. 2003. Analysis and function of transcriptional regulatory elements: insights from Drosophila. *Annual review of entomology* **48**: 579-602.

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **22**(9): 1723-1734.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4): 823-837.

Beer MA, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* **117**(2): 185-198.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**(10): 1045-1048.

Bessant DA, Payne AM, Mitton KP, Wang QL, Swain PK, Plant C, Bird AC, Zack DJ, Swaroop A, Bhattacharya SS. 1999. A mutation in NRL is associated with autosomal dominant retinitis pigmentosa. *Nat Genet* **21**(4): 355-356.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET Thurman RE et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Bishop EP, Rohs R, Parker SC, West SM, Liu P, Mann RS, Honig B, Tullius TD. 2011. A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. *ACS Chem Biol* **6**(12): 1314-1320.

Blackshaw S, Fraioli RE, Furukawa T, Cepko CL. 2001. Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell* **107**(5): 579-589.

Bogdanove AJ, Voytas DF. 2011. TAL effectors: customizable proteins for DNA targeting. *Science* **333**(6051): 1843-1846.

Bovolenta P, Mallamaci A, Briata P, Corte G, Boncinelli E. 1997. Implication of OTX2 in pigment epithelium determination and neural retina differentiation. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **17**(11): 4243-4252.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2): 311-322.

Buchler NE, Gerland U, Hwa T. 2003. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* **100**(9): 5136-5141.

Bulut-Karslioglu A, Perrera V, Scaranaro M, de la Rosa-Velazquez IA, van de Nobelen S, Shukeir N, Popow J, Gerle B, Opravil S, Pagani M et al. 2012. A transcription factor-based mechanism for mouse heterochromatin formation. *Nat Struct Mol Biol* **19**(10): 1023-1030.

Bussemaker HJ, Li H, Siggia ED. 2001. Regulatory element detection using correlation with expression. *Nat Genet* **27**(2): 167-171.

Carey M. 1998. The enhanceosome and transcriptional synergy. *Cell* **92**(1): 5-8.

Carter-Dawson LD, LaVail MM. 1979. Rods and cones in the mouse retina. II. Autoradiographic analysis of cell generation using tritiated thymidine. *The Journal of comparative neurology* **188**(2): 263-272.

Chantalat S, Depaux A, Hery P, Barral S, Thuret JY, Dimitrov S, Gerard M. 2011. Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res* **21**(9): 1426-1437.
Chen J, Rattner A, Nathans J. 2005. The rod photoreceptor-specific nuclear receptor Nr2e3 represses transcription of multiple cone-specific genes. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **25**(1): 118-129.
Chen S, Wang QL, Nie Z, Sun H, Lennon G, Copeland NG, Gilbert DJ, Jenkins NA, Zack DJ. 1997. Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron* **19**(5): 1017-1030.
Chen S, Zack DJ. 1996. Ret 4, a positive acting rhodopsin regulatory element identified using a bovine retina in vitro transcription system. *J Biol Chem* **271**(45): 28549-28557.
Cheng H, Khanna H, Oh EC, Hicks D, Mitton KP, Swaroop A. 2004. Photoreceptor-specific nuclear receptor NR2E3 functions as a transcriptional activator in rod photoreceptors. *Hum Mol Genet* **13**(15): 1563-1575.
Chiang DY, Nix DA, Shultzaberger RK, Gasch AP, Eisen MB. 2006. Flexible promoter architecture requirements for coactivator recruitment. *BMC molecular biology* **7**: 16.
Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**(2): 757-761.
Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. *Genome Res* **19**(9): 1553-1561.
Consortium TEP. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**(4): e1001046.
Corbo JC, Lawrence KA, Karlstetter M, Myers CA, Abdelaziz M, Dirkes W, Weigelt K, Seifert M, Benes V, Fritsche LG et al. 2010. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res* **20**: 1512-1525.
Cox RS, 3rd, Surette MG, Elowitz MB. 2007. Programming gene expression with combinatorial promoters. *Mol Syst Biol* **3**: 145.
Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**(50): 21931-21936.
Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**(7385): 390-394.
Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**(5558): 1306-1311.
Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**(10): 1299-1309.
Driever W, Thoma G, Nusslein-Volhard C. 1989. Determination of spatial domains of zygotic gene expression in the Drosophila embryo by the affinity of binding sites for the bicoid morphogen. *Nature* **340**(6232): 363-367.
Dunham I Kundaje A Aldred SF Collins PJ Davis CA Doyle F Epstein CB Frietze S Harrow J Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.
Elemento O, Slonim N, Tavazoie S. 2007. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28**(2): 337-350.
Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**(8): 817-825.
-. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**(3): 215-216.
Ferraris L, Stewart AP, Kang J, DeSimone AM, Gemberling M, Tantin D, Fairbrother WG. 2011. Combinatorial binding of transcription factors in the pluripotency control regions of the genome. *Genome Res* **21**(7): 1055-1064.

Foat BC, Tepper RG, Bussemaker HJ. 2008. TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors. *Nucleic acids research* **36**(Database issue): D125-131.
Freund CL, Gregory-Evans CY, Furukawa T, Papaioannou M, Looser J, Ploder L, Bellingham J, Ng D, Herbrick JA, Duncan A et al. 1997. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell* **91**(4): 543-553.
Freund CL, Wang QL, Chen S, Muskat BL, Wiles CD, Sheffield VC, Jacobson SG, McInnes RR, Zack DJ, Stone EM. 1998. De novo mutations in the CRX homeobox gene associated with Leber congenital amaurosis. *Nat Genet* **18**(4): 311-312.
Friedman JS, Khanna H, Swain PK, Denicola R, Cheng H, Mitton KP, Weber CH, Hicks D, Swaroop A. 2004. The minimal transactivation domain of the basic motif-leucine zipper transcription factor NRL interacts with TATA-binding protein. *J Biol Chem* **279**(45): 47233-47241.
Furukawa T, Morrow EM, Cepko CL. 1997. Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell* **91**(4): 531-541.
Furukawa T, Morrow EM, Li T, Davis FC, Cepko CL. 1999. Retinopathy and attenuated circadian entrainment in Crx-deficient mice. *Nat Genet* **23**(4): 466-470.
Gertz J, Cohen BA. 2009. Environment-specific combinatorial *cis*-regulation in synthetic promoters. *Mol Syst Biol* **5**: 244.
Gertz J, Siggia ED, Cohen BA. 2009. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**(7226): 215-218.
Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**(6): 877-885.
Goodbourn S, Maniatis T. 1988. Overlapping positive and negative regulatory domains of the human beta-interferon gene. *Proc Natl Acad Sci U S A* **85**(5): 1447-1451.
Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, Loftus SK, Beer MA, Pavan WJ, McCallion AS. 2012. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Res* **22**(11): 2290-2301.
Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**(5): 565-577.
Goutsias J, Kim S. 2004. A nonlinear discrete dynamical model for transcriptional regulation: construction and properties. *Biophysical journal* **86**(4): 1922-1945.
Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**(7): 1017-1018.
Gurnett CA, Bowcock AM, Dietz FR, Morcuende JA, Murray JC, Dobbs MB. 2007. Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. *Am J Med Genet A* **143**(1): 27-32.
Haider NB, Demarco P, Nystuen AM, Huang X, Smith RS, McCall MA, Naggert JK, Nishina PM. 2006. The transcription factor Nr2e3 functions in retinal progenitors to suppress cone cell generation. *Visual neuroscience* **23**(6): 917-929.
Hao H, Kim DS, Klocke B, Johnson KR, Cui K, Gotoh N, Zang C, Gregorski J, Gieser L, Peng W et al. 2012. Transcriptional regulation of rod photoreceptor homeostasis revealed by in vivo NRL targetome analysis. *PLoS Genet* **8**(4): e1002649.
Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004): 99-104.
Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**(4): 576-589.
Hennig AK, Peng GH, Chen S. 2008. Regulation of photoreceptor gene expression by Crx-associated transcription factor network. *Brain Res* **1192**: 114-133.
Hess J, Angel P, Schorpp-Kistner M. 2004. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci* **117**(Pt 25): 5965-5973.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**(5): 473-476.
Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**(2): 827-841.
Hsiau TH, Diaconu C, Myers CA, Lee J, Cepko CL, Corbo JC. 2007. The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS ONE* **2**(7): e643.
Hu Z, Killion PJ, Iyer VR. 2007. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* **39**(5): 683-687.
Humphries MM, Rancourt D, Farrar GJ, Kenna P, Hazel M, Bush RA, Sieving PA, Sheils DM, McNally N, Creighton P et al. 1997. Retinopathy induced in mice by targeted disruption of the rhodopsin gene. *Nat Genet* **15**(2): 216-219.
Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318-356.
Jiang J, Levine M. 1993. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* **72**(5): 741-752.
Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830): 1497-1502.
Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**(1-2): 327-339.
Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**(7236): 362-366.
Kaplan S, Bren A, Dekel E, Alon U. 2008. The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Mol Syst Biol* **4**: 203.
Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE et al. 2010. Variation in transcription factor binding among humans. *Science* **328**(5975): 232-235.
Kataoka K, Noda M, Nishizawa M. 1994. Maf nuclear oncoprotein recognizes sequences related to an AP-1 site and forms heterodimers with both Fos and Jun. *Mol Cell Biol* **14**(1): 700-712.
Kautzmann MA, Kim DS, Felder-Schmittbuhl MP, Swaroop A. 2011. Combinatorial regulation of photoreceptor differentiation factor, neural retina leucine zipper gene NRL, revealed by in vivo promoter analysis. *J Biol Chem* **286**(32): 28247-28255.
Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**(5): 800-811.
Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**(7295): 182-187.
Kinkhabwala A, Guet CC. 2008. Uncovering cis regulatory codes using synthetic promoter shuffling. *PLoS ONE* **3**(4): e2030.
Kinney JB, Tkacik G, Callan CG. 2007. Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences of the United States of America* **104**(2): 501-506.
Kobayashi M, Takezawa S, Hara K, Yu RT, Umesono Y, Agata K, Taniwaki M, Yasuda K, Umesono K. 1999. Identification of a photoreceptor cell-specific nuclear receptor. *Proc Natl Acad Sci U S A* **96**(9): 4814-4819.
Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**(47): 19498-19503.
Lam FH, Steger DJ, O'Shea EK. 2008. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**(7192): 246-250.
Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, Weng Z, Myers RM. 2010. Sequence features that drive human promoter function and tissue specificity. *Genome Res* **20**(7): 890-898.
Lee D, Karchin R, Beer MA. 2011. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21**(12): 2167-2180.

Lee J, Myers CA, Williams N, Abdelaziz M, Corbo JC. 2010. Quantitative fine-tuning of photoreceptor cis - regulatory elements through affinity modulation of transcription factor binding sites *Gene Therapy* **17**: 1390-1399.
Lee TI, Young RA. 2000. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**: 77-137.
Lem J, Applebury ML, Falk JD, Flannery JG, Simon MI. 1991. Tissue-specific and developmental regulation of rod opsin chimeric genes in transgenic mice. *Neuron* **6**(2): 201-210.
LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH. 2010. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res* **38**(8): 2522-2540.
Lettice LA, Williamson I, Wiltshire JH, Peluso S, Devenney PS, Hill AE, Essafi A, Hagman J, Mort R, Grimes G et al. 2012. Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. *Dev Cell* **22**(2): 459-467.
Lidor Nili E, Field Y, Lubling Y, Widom J, Oren M, Segal E. 2010. p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome Res* **20**(10): 1361-1368.
Ligr M, Siddharthan R, Cross FR, Siggia ED. 2006. Gene expression from random libraries of yeast promoters. *Genetics* **172**(4): 2113-2122.
Livesey FJ, Furukawa T, Steffen MA, Church GM, C L, Cepko. 2000. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx. *Curr Biol* **10**(6): 301-310.
MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. 2006. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**: 113.
Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**: 387-402.
Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H et al. 2013. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*.
Matsuda T, Cepko CL. 2004. Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc Natl Acad Sci U S A* **101**(1): 16-22.
Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**(6099): 1190-1195.
Mears AJ, Kondo M, Swain PK, Takada Y, Bush RA, Saunders TL, Sieving PA, Swaroop A. 2001. Nrl is required for rod photoreceptor development. *Nat Genet* **29**(4): 447-452.
Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**(3): 271-277.
Merika M, Thanos D. 2001. Enhanceosomes. *Curr Opin Genet Dev* **11**(2): 205-208.
Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17**(12): 1797-1808.
Mitton KP, Swain PK, Chen S, Xu S, Zack DJ, Swaroop A. 2000. The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. *J Biol Chem* **275**(38): 29794-29799.
Mogno I, Vallania F, Mitra RD, Cohen BA. 2010. TATA is a modular component of synthetic promoters. *Genome Res* **20**(10): 1391-1397.
Montana CL, Lawrence KA, Williams NL, Tran NM, Peng GH, Chen S, Corbo JC. 2011a. Transcriptional regulation of neural retina leucine zipper (Nrl), a photoreceptor cell fate determinant. *J Biol Chem* **286**(42): 36921-36931.
Montana CL, Myers CA, Corbo JC. 2011b. Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation. *J Vis Exp*(52).
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.
Muthukrishnan R, Skalnik DG. 2009. Identification of a minimal cis-element and cognate trans-factor(s) required for induction of Rac2 gene expression during K562 cell differentiation. *Gene* **440**(1-2): 63-72.

Myers RM, Tilly K, Maniatis T. 1986. Fine structure genetic analysis of a beta-globin promoter. *Science* **232**(4750): 613-618.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.
Nam J, Dong P, Tarpine R, Istrail S, Davidson EH. 2010. Functional cis-regulatory genomics for systems biology. *Proc Natl Acad Sci U S A* **107**(8): 3930-3935.
Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**(7414): 83-90.
Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **37**(Database issue): D77-82.
Oh EC, Khan N, Novelli E, Khanna H, Strettoi E, Swaroop A. 2007. Transformation of cone precursors to functional rod photoreceptors by bZIP transcription factor NRL. *Proc Natl Acad Sci U S A* **104**(5): 1679-1684.
Olsson JE, Gordon JW, Pawlyk BS, Roof D, Hayes A, Molday RS, Mukai S, Cowley GS, Berson EL, Dryja TP. 1992. Transgenic mice with a rhodopsin mutation (Pro23His): a mouse model of autosomal dominant retinitis pigmentosa. *Neuron* **9**(5): 815-830.
Pachkov M, Erb I, Molina N, van Nimwegen E. 2007. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic acids research* **35**(Database issue): D127-131.
Parker DS, White MA, Ramos AI, Cohen BA, Barolo S. 2011. The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Sci Signal* **4**(176): ra38.
Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324**(5925): 389-392.
Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**(3): 265-270.
Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**(12): 1173-1175.
Paulsson J. 2005. Models of stochastic gene expression. *Physics of Life Reviews* **2**(2): 157-175.
Peng GH, Ahmad O, Ahmad F, Liu J, Chen S. 2005. The photoreceptor-specific nuclear receptor Nr2e3 interacts with Crx and exerts opposing effects on the transcription of rod versus cone genes. *Hum Mol Genet* **14**(6): 747-764.
Peng GH, Chen S. 2007. Crx activates opsin transcription by recruiting HAT-containing co-activators and promoting histone acetylation. *Hum Mol Genet* **16**(20): 2433-2452.
Peterson MD, Liu D, Iglayreger HB, Saltarelli WA, Visich PS, Gordon PM. 2012. Principal component analysis reveals gender-specific predictors of cardiometabolic risk in 6th graders. *Cardiovasc Diabetol* **11**: 146.
Phillips K, Luisi B. 2000. The virtuoso of versatility: POU proteins that flex to fit. *J Mol Biol* **302**(5): 1023-1039.
Polach KJ, Widom J. 1996. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *Journal of molecular biology* **258**(5): 800-812.
Ptashne M. 2005. Regulation of transcription: from lambda to eukaryotes. *Trends in biochemical sciences* **30**(6): 275-279.
Ptashne M, Gann A. 1997. Transcriptional activation by recruitment. *Nature* **386**(6625): 569-577.
Ptashne M, Gann A. 2002. *Genes & signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Rastegar S, Hess I, Dickmeis T, Nicod JC, Ertzer R, Hadzhiev Y, Thies WG, Scherer G, Strahle U. 2008. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol* **318**(2): 366-377.
Raveh-Sadka T, Levo M, Segal E. 2009. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* **19**(8): 1480-1496.
Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E. 2012. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature genetics* **44**(7): 743-750.

Rehemtulla A, Warwar R, Kumar R, Ji X, Zack DJ, Swaroop A. 1996. The basic motif-leucine zipper transcription factor Nrl can positively regulate rhodopsin gene expression. *Proc Natl Acad Sci U S A* **93**(1): 191-195.
Rembold M, Ciglar L, Yanez-Cuna JO, Zinzen RP, Girardot C, Jain A, Welte MA, Stark A, Leptin M, Furlong EE. 2014. A conserved role for Snail as a potentiator of active transcription. *Genes Dev* **28**(2): 167-181.
Remenyi A, Scholer HR, Wilmanns M. 2004. Combinatorial control of gene expression. *Nat Struct Mol Biol* **11**(9): 812-815.
Rivolta C, Peck NE, Fulton AB, Fishman GA, Berson EL, Dryja TP. 2001. Novel frameshift mutations in CRX associated with Leber congenital amaurosis. *Hum Mutat* **18**(6): 550-551.
Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. 2009. The role of DNA shape in protein-DNA recognition. *Nature* **461**(7268): 1248-1253.
Roorda A, Williams DR. 1999. The arrangement of the three cone classes in the living human eye. *Nature* **397**(6719): 520-522.
Schlabach MR, Hu JK, Li M, Elledge SJ. 2010. Synthetic design of strong promoters. *Proc Natl Acad Sci U S A* **107**(6): 2538-2543.
Scott EW, Baker HV. 1993. Concerted action of the transcriptional activators REB1, RAP1, and GCR1 in the high-level expression of the glycolytic gene TPI. *Mol Cell Biol* **13**(1): 543-550.
Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**(7178): 535-540.
Setty Y, Mayo AE, Surette MG, Alon U. 2003. Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences of the United States of America* **100**(13): 7702-7707.
Shalem O, Carey L, Zeevi D, Sharon E, Keren L, Weinberger A, Dahan O, Pilpel Y, Segal E. 2013. Measurements of the impact of 3' end sequences on gene expression reveal wide range and sequence dependent effects. *PLoS computational biology* **9**(3): e1002934.
Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**(6): 521-530.
Shea MA, Ackers GK. 1985. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* **181**(2): 211-230.
Shimell MJ, Simon J, Bender W, O'Connor MB. 1994. Enhancer point mutation results in a homeotic transformation in Drosophila. *Science* **264**(5161): 968-971.
Shin I, Kim J, Cantor CR, Kang C. 2000. Effects of saturation mutagenesis of the phage SP6 promoter on transcription activity, presented by activity logos. *Proc Natl Acad Sci U S A* **97**(8): 3890-3895.
Singh K, Tokuhisa JG, Dennis ES, Peacock WJ. 1989. Saturation mutagenesis of the octopine synthase enhancer: correlation of mutant phenotypes with binding of a nuclear protein factor. *Proc Natl Acad Sci U S A* **86**(10): 3733-3737.
Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ et al. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**(6): 1270-1282.
Smith JR, Osborne TF, Goldstein JL, Brown MS. 1990. Identification of nucleotides responsible for enhancer activity of sterol regulatory element in low density lipoprotein receptor gene. *J Biol Chem* **265**(4): 2306-2310.
Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**(9): 1021-1028.
Sohocki MM, Sullivan LS, Mintz-Hittner HA, Birch D, Heckenlively JR, Freund CL, McInnes RR, Daiger SP. 1998. A range of clinical phenotypes associated with mutations in CRX, a photoreceptor transcription-factor gene. *Am J Hum Genet* **63**(5): 1307-1315.
Solomon MJ, Larsen PL, Varshavsky A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**(6): 937-947.
Spivak AT, Stormo GD. 2012. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic acids research* **40**(Database issue): D162-168.

Stemmer WP. 1994. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci U S A* **91**(22): 10747-10751.
Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16**(1): 16-23.
Stormo GD, Fields DS. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* **23**(3): 109-113.
Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **10**(9): 2997-3011.
Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**(6765): 41-45.
Struhl K, Segal E. 2013. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**(3): 267-273.
Swaroop A, Kim D, Forrest D. 2010. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nature reviews Neuroscience* **11**(8): 563-576.
Tanay A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res* **16**(8): 962-972.
Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**(7414): 75-82.
Tornow J, Zeng X, Gao W, Santangelo GM. 1993. GCR1, a transcriptional activator in Saccharomyces cerevisiae, complexes with RAP1 and can function without its DNA binding domain. *Embo J* **12**(6): 2431-2437.
Vashee S, Melcher K, Ding WV, Johnston SA, Kodadek T. 1998. Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein-protein interactions. *Curr Biol* **8**(8): 452-458.
Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**(7231): 854-858.
Wang T, Stormo GD. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**(18): 2369-2380.
Wasson T, Hartemink AJ. 2009. An ensemble model of competitive multi-factor binding of the genome. *Genome Res* **19**(11): 2101-2112.
Webber AL, Hodor P, Thut CJ, Vogt TF, Zhang T, Holder DJ, Petrukhin K. 2008. Dual role of Nr2e3 in photoreceptor development and maintenance. *Exp Eye Res* **87**(1): 35-48.
Weiher H, Konig M, Gruss P. 1983. Multiple point mutations affecting the simian virus 40 enhancer. *Science* **219**(4585): 626-631.
Wharton SJ, Basu SP, Ashe HL. 2004. Smad affinity can direct distinct readouts of the embryonic extracellular Dpp gradient in Drosophila. *Current biology : CB* **14**(17): 1550-1558.
White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci U S A* **110**(29): 11952-11957.
Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**(1): 238-241.
Yuh CH, Bolouri H, Davidson EH. 2001. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128**(5): 617-629.
Yun Y, Adesanya TM, Mitra RD. 2012. A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. *Genome Res*.
Zack DJ, Bennett J, Wang Y, Davenport C, Klaunberg B, Gearhart J, Nathans J. 1991. Unusual topography of bovine rhodopsin promoter-lacZ fusion gene expression in transgenic mouse retinas. *Neuron* **6**(2): 187-199.
Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**(7269): 65-70.

CURRICULUM VITAE

Jamie C Kwasnieski, Ph.D. candidate
Center for Genome Sciences and Systems Biology, Washington University in St. Louis
4444 Forest Park Blvd., Room 5510 · jkwasnieski@wustl.edu

Education

2008 – present	Ph.D. in Molecular Genetics and Genomics Washington University in St. Louis, School of Medicine, St. Louis, MO Mentor: Barak A. Cohen Degree anticipated, May 2014
2004 – 2008	B.A. in Chemistry Northwestern University, Evanston, IL

Research Experience

2009 – present	Graduate Student, Washington University in St. Louis, St. Louis, MO Mentor: Barak A. Cohen, Ph.D. Developed a massively parallel reporter assay to identify *cis*-regulatory sequences that control gene expression in mammalian rod photoreceptor cells.
2008 – 2009	Rotation Student, Washington University in St. Louis, St. Louis, MO 1) Mentor: S. Kerry Kornfeld, M.D., Ph.D. Conducted epistatic analysis of Zinc transport mutations in *C. elegans*. 2) Mentor: Stephen L. Johnson, Ph.D. Conducted screen looking for FDA-approved drugs that inhibit melanocyte regeneration in *D. rerio*. 3) Mentor: Anne Bowcock, Ph.D. Characterized the dysregulation of multiple long RNA species in psoriatic tissue samples.
2007 – 2008	Undergraduate Research Internship, Northwestern University, Evanston, IL Mentor: Andreas Matouschek, Ph.D. Utilized a synthetic biology approach to understand the peptide features necessary for proteasomal degradation.
2007	Summer Research Assistant, Boystown National Research Hospital, Omaha, NE Mentor: Dominic E. Cosgrove, Ph.D. Characterized and determined the localization of a protein complex in mammalian retinal pigment epithelial cells.

Teaching Experience

2010	Teaching assistant, “Introduction to Cell Biology”. Washington University in St Louis, St. Louis, MO.
2009	Teaching assistant, “Methods of DNA Manipulation”. Washington University in St Louis, St. Louis, MO.

Honors and Awards

2012	Best Student Talk, Department of Genetics Retreat, Washington University in St. Louis
2011	Best Student Poster, Department of Genetics Retreat, Washington University in St. Louis

Publications

1) **Kwasnieski JC***, Mogno I*, Myers CA, Corbo JC, Cohen BA. (2012) Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc. Nat. Acad. Sci.* 109(47): 19498-503. doi: 10.1073/pnas.1210678109.

2) Mogno I, **Kwasnieski JC***, Cohen BA. (2013) Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* 23(11): 1908-1915. doi: 10.1101:gr.157891.113

3) **Kwasnieski JC***, Fiore CM*, Chaudhari HG, Cohen BA. High-throughput testing of ENCODE putative enhancers suggests that combinations of histone marks predict *cis*-regulatory activity. *In review.*

4) **Kwasnieski JC**, Myers CA, Corbo JC, Cohen BA. Combinations of CRX and NRL binding sites generate diverse expression levels. *In preparation.*

***** Denotes equal contributions

Conference Presentations

1) **Kwasnieski JC**, Mogno I, Myers CA, Corbo JC, Cohen BA. (2012) Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. Poster presentation given at the *Cell and Molecular Biology Symposium*, St. Louis, MO

2) **Kwasnieski JC**, Mogno I, Myers CA, Corbo JC, Cohen BA. (2011) A method to dissect the mouse Rhodopsin promoter at single nucleotide resolution. Poster presented at the *Systems Biology: Global Regulation of Gene Expression Meeting*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

3) **Kwasnieski JC**, Mogno I, Myers CA, Corbo JC, Cohen BA. (2011) A method to dissect the mouse Rhodopsin promoter at single nucleotide resolution. Poster presentation given at the *Center for Genome Sciences Symposium*, St. Louis, MO

Computational and Mathematical Skills

- Fluency in Perl, Python, R
- Experience using Java, C, MATLAB
- Unix/Linux and use of cluster computing systems
- Next-Generation sequencing analysis, including generation of custom scripts and use of existing computational packages (Bowtie, MACS, Novoalign, etc)
- Experience building statistical thermodynamic models of gene expression

Laboratory Skills

- Standard molecular biology protocols including PCR, qPCR, PCR mutagenesis, stitching PCR, bacterial cloning, etc
- Custom Next-Generation sequencing library design and preparation
- RNA isolation and RT-PCR
- Large scale plasmid library design and creation

- Protein biochemistry techniques including ChIP, Western Blotting, Immunohistochemistry and Co-immunoprecipitation
- Standard tissue culture protocols (Stem cell and Cancer cell lines)
