Summer 8-15-2015

Functional Identification and Characterization of cis-Regulatory Elements

Christopher Michael Fiore
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Barak A. Cohen, Chair
Joe Dougherty
Justin Fay
Gary Stormo
Andrew Yoo

Functional Identification and Characterization of *cis*-Regulatory Elements

by
Christopher M Fiore

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

August 2015
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

My graduate work has been supported by countless others, both through direct support of my research and through support of my professional development. The Cohen lab has been an amazing place to do research and learn about science; I have been consistently been impressed with the intelligent people in the lab. Gatherings of lab members, whether in lab meetings or random conversations, were the source of incredibly insightful conversations that helped me to learn and grow as a scientist. The leader of this bunch, Barak Cohen, has consistently led us to be better scientists through his sharp feedback and wise questions. He is essential to the formative spirit that the Cohen lab exhibits and has helped make me the scientist I am today. I would also like to thank Robert Zeigler and Mike White for helping me with concepts related to the thermodynamic model. Ilaria Mogno assisted with the thermodynamic model as well as CRE-seq. Jamie Kwasnieski also helped with CRE-seq and was a great collaborator with whom I had many great scientific conversations. She is also a great friend to hang out with, both inside and outside of lab. Additionally, I owe every member of the Cohen lab a debt of gratitude for making it a wonderful place to spend my graduate career.

The Center for Genome Sciences and Systems Biology (CGS) has been an amazing place to work and is full of intelligent and wonderful people willing to assist you and listen to your ideas. I am especially indebted to Xuhua Chen in the Mitra lab for teaching me to use cell culture and Jess Hoisington-Lopez for being an Illumina sequencing guru for all my unusual sequencing samples.

My thesis committee of Gary Stormo, Justin Fay, Joe Dougherty, Ting Wang, and Andrew Yoo provided great feedback and advice on all aspects of my research. It was always incredibly useful to hear their opinions. I am also grateful to other professors I have worked with before coming to Washington University for introducing me to scientific research. These include Kevin J. Peterson at Dartmouth College and Massimo Loda at Dana-Farber Cancer Institute.

Washington University has provided countless other friends who made life enjoyable and offered conversations of all things science and graduate school. Kevin Forsberg and Keith Jacobs were particularly lively partners in conversation about science and otherwise, but many others contributed to my professional development. I owe much to all my friends who made my graduate school experience a great one. I also spent more than two years working with great people in The BALSA Group. This experience helped me to develop professionally and learn great collaborative skills.

Through graduate school I was also able to meet the wonderful Beth Selleck, who soon became my wife. Having a partner in science and graduate school has countless benefits, and Beth provided loving support and entertainment throughout the process. She is an intelligent and wonderful person, and I wouldn't have managed without her.

I am also incredibly thankful for my family for setting me up for a great education and a successful career. My parents have always been incredibly supportive of me and pushed me to find my passion and a career I can enjoy. It's safe to say I would not be where I am today without them and their efforts.

My research was funded by National Science Foundation Graduate Research Fellowship DGE-1143954 and NIH Grants R01 HG006790-01 and RO1 GM092910-02.

Christopher M Fiore

Washington University in St. Louis

August 2015

ABSTRACT OF THE DISSERTATION

Functional Identification and Characterization of *cis*-Regulatory Elements

By Christopher M Fiore

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2015

Professor Barak A. Cohen, Chair

Transcription is regulated through interactions between regulatory proteins, such as transcription factors (TFs), and DNA sequence. It is known that TFs act combinatorially in some cases to regulate transcription, but in which situations and to what degree is unclear.

I first studied the contribution of TF binding sites to expression in mouse embryonic stem (ES) cells by using synthetic *cis*-regulatory elements (CREs). The synthetic CREs were comprised of combinations of binding sites for the pluripotency TFs Oct4, Sox2, Klf4, and Esrrb. A statistical thermodynamic model explained 72% of the variation in expression driven by these CREs. The high predictive power of this model depended on five TF interaction parameters, including favorable heterotypic interactions between Oct4 and Sox2, Klf4 and Sox2, and Klf4 and Esrrb. The model also included two unfavorable homotypic interaction parameters. These homotypic parameters help to explain the fact that synthetic CREs with mixtures of binding sites for various TFs drive much higher expression than multiple binding sites for the same TF. I then found that the expression of these synthetic CREs largely changes as ES cells differentiate down the neural lineage. However, CREs with no repeat binding sites drove similar levels of expression, suggesting that heterotypic interactions may be similar in the two conditions.

In a separate set of experiments I interrogated the determinants of expression driven by genomic sequences previously segmented into classes based on chromatin features. A set of these sequences was assayed in K562 cells. As expected, we found that Enhancers and Weak Enhancers drove expression over background, while Repressed elements and Enhancers from another cell type did not. Unexpectedly, we found that Weak Enhancers drove higher expression than Enhancers, possibly based on their lower H3K36me3 and H3K27ac, which we found to be weakly associated with lower expression. Using a logistic regression model, we showed that matches to TF binding motifs were best able to predict active sequences, but chromatin features contributed significantly as well.

These results demonstrate that interactions between certain combinations of pluripotency TFs, but not all combinations, are important to transcriptional regulation. Furthermore, chromatin modifications can still contribute to predictions of expression even after accounting for binding site motifs. Better understanding of the process of *cis*-regulation will allow us to predict which sequences can drive expression and how perturbations affect this expression.

Chapter 1: Introduction

Transcriptional regulation helps control the processes that run the cell and is essential for the proper development of multicellular organisms. We know that some complex rules guide this regulation, but it is unclear to what extent these rules operate and what effect they have. It is therefor necessary to study transcriptional regulation from a mechanistic viewpoint in order to learn in which situations these rules apply. In this dissertation, I use this approach to move us closer to understanding the processes behind regulation of transcription.

Importance of transcriptional regulation

The control of the gene expression is essential to the proper functioning of the cell and the development of multi-cellular organisms. Every cell in an organism has the same genes encoded in its genome, but the activity level of these genes varies wildly between cells and across time. A gene's expression level is often tightly controlled, and transcription is the first step in gene regulation. By one estimate, transcriptional regulation is responsible for 73% of a gene's protein level (Li and Biggin 2015). Various processes can influence transcriptional regulation, including external stimuli, developmental stage, and general cellular processes such as cell cycle. The regulation of transcription is thus a key process controlling a cell's identity and its proper function.

It is clear from a number of different approaches that a large amount of non-coding sequence in mammalian genomes is functional. Tests for selection over evolution can give us an idea of the how much non-coding sequence is functional. A number of groups have estimated that 3-10% of the human genome is under selection, the majority of which is non-coding (Bernstein et al. 2012; Meader et al. 2010; Rands et al. 2014; Cooper et al. 2005). Additionally, a large proportion of so-called ultra conserved elements, with 100% sequence identity between mammalian species, are non-coding and thought to be regulatory regions (Bejerano et al. 2004). Even when the sequence identity of two orthologous non-coding sequences differ, their function may still be conserved due to turnover of transcription factor binding sites (Meader et al. 2010; Moses et al. 2006). Furthermore, about 93% of peaks from genome-wide association studies (GWAS) are in non-coding regions, implying that a good deal of disease-associated variation is operating through non-coding regions (Maurano et al. 2012). Thus, there is a significant amount of non-coding sequence with important function, which often operates to regulate transcription.

Non-coding sequence works to regulate transcription in many developmental processes. Transcription factors (TFs) drive the expression of developmental genes in species from sea urchins to mammals, helping to specify cell fate and pattern whole organisms (Davidson et al. 2002; Arnosti 2003; Kranz et al. 2011; Novershtern et al. 2011; Oliveri et al. 2008; Zinzen et al. 2009; Segal et al. 2008; Odom et al. 2004). The regulation of TFs varies across developmental stages, leading to changes in the regulation of downstream genes, which is critical for proper development. This underlies the need to study the function of the non-coding regions of mammalian genomes that regulate transcription, especially in developmental systems.

Mechanisms of transcriptional regulation

Many years of research have shown that regulation of transcription takes place through regulatory proteins recognizing specific DNA sequences and acting through interactions with other proteins. Initial work into transcriptional regulation focused on bacteria and in particular the lambda phage genes. Jacob and Monod's early work into the lac and lambda phage genes showed that the protein working to regulate transcription, the lambda repressor, acted to repress transcription (Jacob and Monod 1961). They thus posited that active transcription was the default state and regulation worked through repressors to turn off transcription. Later work showed that this situation is not always the norm, as the lambda repressor could also activate transcription and the basal level of transcription was relatively low. Further work suggested that interactions between proteins might mediate their ability to activate transcription (Hochschild et al. 1983; Keegan et al. 1986; Ptashne 2005). This established the basic role of these regulatory proteins, referred to as transcription factors (TFs), in regulating transcription.

The mechanisms by which TFs regulate transcription follow general principles of binding and activation. The DNA-binding function of TFs is largely decoupled from the regulatory function of the TF, allowing different TFs to bind the same DNA sequence but have different regulatory functions, or to bind different DNA sequences but have similar regulatory functions (Brent and Ptashne 1985; Keegan et al. 1986). This modularity opens the door for complex regulatory networks. The DNA binding of TFs helps guide these regulatory proteins to proper genomic locations and works through degenerate binding sites, meaning that each TF can recognize a number of different but similar DNA sequences. The mechanism of binding usually operates through hydrogen bonding of the TF with the major or minor groove of the DNA along with general non-specific interactions (von Hippel and Berg 1986). The sequence specificity of a TF can then be used to predict other binding sites and the activity of promoters (von Hippel and Berg 1986; Stormo and Fields 1998). While this is easier in simpler systems, such as bacteria, it becomes much harder in larger genomes.

After binding to DNA, TFs work with other TFs and proteins to regulate transcription. TFs can recruit transcriptional complexes that begin the process of transcription, as well as chromatin remodeling complexes to improve accessibility of other TFs and the general transcriptional machinery. TFs can also interact with each other to help recruit these complexes, demonstrating the great breadth of possible complexity with which TFs can regulate transcription (Cosma et al. 1999; Malik and Roeder 2000; Dröge and Müller-Hill 2001; Arnosti et al. 1996). Furthermore, these interactions can occur over long distances through looping of the DNA (Fullwood et al. 2009; Hochschild and Ptashne 1986). This means both that the potential regulatory region of a gene could lie far from its location and that the gene a specific regulatory element regulates may not be the nearest one. These potentially complex interactions indicate that there may be some specific logic or order with which TFs operate. Two possible scenarios exist to describe TF activity: either they act regardless of the surrounding context, or their action depends on interactions with other proteins and specific spatial requirements. There are examples by which specific arrangements of TFs are important in activating transcription through synergy between the proteins bound at the regulatory element (Thanos and Maniatis 1995; Kim and Maniatis 1997; Arnosti et al. 1996). This cooperativity between TFs could be mediated either at the level of binding or at the level of recruitment of transcriptional machinery (Harbison et al. 2004). Other examples however, show more flexibility in the arrangement and nature of the proteins bound at the regulatory DNA (Kulkarni and Arnosti 2003). An important problem in genetics has been to identify scenarios in which cooperativity, and thus a more detailed *cis*-regulatory logic, plays an important role in transcriptional regulation.

Discovery and modeling of *cis*-regulatory elements

Many studies have attempted to predict the *cis*-regulation of genomic sequences using combinations of sequence features, protein binding, and chromatin features. The hope is that these predictive models can teach us something about the mechanisms of *cis*-regulation. Predictions can be done either at the level of single genes or across the whole genome. Single-gene studies tend to focus on more minute details of regulation, whereas whole genome studies abstract away some detail in order to obtain a comprehensive picture. Since large genomes, such as those of mammals, consist primarily of intergenic space, it can be difficult to determine which regions have *cis*-regulatory function at all. We know that TFs have sequence specificity (von Hippel and Berg 1986), so we can predict their binding sites across a large genome. However, only a small fraction of these are bound by a TF and fewer still have *cis*-regulatory function (Li et al. 2008; Zhang et al. 2005; Whiteld et al. 2012; White et al. 2013). Thus, computational tools are needed to discover functional *cis*-regulatory elements and discern their effect on transcription.

Many studies focus on individual promoter elements in organisms with smaller genomes to gain a mechanistic view of the regulation of transcription. Examining the lac operon in *E. coli*, Uri Alon's group found that the concentrations of small molecule inducers can perform a computation-like logic to regulate the expression of the gene, suggesting that transcription is regulated by a set of rules (Setty et al. 2003). Another study in *E. coli* examined the expression driven by mutated versions of the lac promoter to infer interactions between TFs and RNA Polymerase and their relationship to expression (Kinney et al. 2010). These studies were done in *E. coli*, which is a simpler system to study transcription but lacks some of the elements of eukaryotes, such as nucleosomes. A study in yeast incorporated nucleosomes into a kinetic framework to model transcription at the PHO5 promoter. This allowed them to predict expression driven by this promoter and interactions between TFs and nucleosomes (Kim and O'Shea 2008). These studies are good at working out the mechanisms at a single locus, but it's unclear how often they can be generalized to the whole genome.

A number of statistical and computational strategies have been implemented to investigate the general mechanisms of *cis*-regulation that occur across the genome. Most work has used sequence features, such as matches to TF binding motifs (Stormo and Fields 1998), to predict expression. Simple linear models using sequence motifs in the promoters of yeast genes can be effective in predicting relative expression across conditions and learning sequence motifs relevant to certain cellular conditions (Bussemaker et al. 2001). More complex models, such as Bayesian networks, allow known TF binding motifs to be combined with combinatorial logic. They have been used to predict expression of yeast genes across conditions and find possible TF interactions (Beer and Tavazoie 2004). When examining genomes larger than yeast, it becomes important to select amongst many potential *cis*-regulatory regions. In *Drosophila*, a common strategy has been to use *cis*-regluatory modules (CRMs), which are distinct regulatory regions and can act independently to regulate transcription. A logistic regression framework using binding motifs in CRMs and the concentration of TFs in each cell has been used to predict expression patterns along the anterior-posterior (A-P) axis in the developing Drosophila embryo (Kazemian et al. 2010). A similar model in mammalian cells used TF binding motifs to predict promoter sequences active in a reporter assay using a support vector machine (SVM) (Landolin et al. 2010). Rather than using TF binding motifs, another group used all possible 6bp sequences to predict sequences bound by P300, a transcriptional co-activator and marker of enhancers, using an SVM (Lee et al. 2011a). While these modeling frameworks can be useful in predicting expression and coming up with basic features that contribute to *cis*-regulatory activity, a more mechanistic model may uncover determinants of regulation that explain the functional activity of transcriptional regulators.

Statistical thermodynamic models of transcription allow for mechanistic explanations of *cis*-regulatory features using information about cooperativity between TFs. These models incorporate parameters that are related to the change in free energy of interactions on a *cis*-

regulatory element to explain the thermodynamic stability of protein complexes on a potential regulatory DNA sequence. This model makes the assumption that the proteins are in thermodynamic equilibrium and that any kinetic terms are wrapped up in the thermodynamic parameters. A key assumption of these models is that the probability that RNA Polymerase II (RNAP) is bound at the regulatory element is proportional to the expression driven by the regulatory element. Importantly, these models can incorporate parameters of interaction between TFs, both providing a model of *cis*-regulatory logic and a potential mechanism. Shea and Ackers first used this framework to describe the mechanisms of regulation of the lambda OR control system. They modeled the binding of the cI dimer, the cro dimer, and RNAP at three Or binding sites. Using a thermodynamic modeling framework and the expression of two promoters during induction of lysis, they were able to show that cooperativity between cI repressors is an important element in regulation of these genes (Shea and Ackers 1985). Buchler and colleagues extended the model to a more general framework to show that a thermodynamic model can be used to flexibly encode a number of different types of *cis*-regulatory logic (Buchler et al. 2003). These works laid the groundwork for other applications of the model in new systems.

A common use of the thermodynamic framework has been used to demonstrate the degree to which proteins cooperate on a *cis*-regulatory element. In the Drosophila embryo, thermodynamic modeling has been used to show that homotypic interactions between TFs are important features of the regulation of gene expression patterns across the anterior-posterior axis (Segal et al. 2008). Another group extended this study to show that this cooperativity is best explained by simultaneous interaction of TFs with RNAP (He et al. 2010). Synthetic or mutated *cis*-regulatory elements (CREs) have also been used in Drosophila to show that cooperativity between TFs can explain expression patterns of individual CREs (Parker et al. 2011; Erceg et al.

2014). One of the advantages of the thermodynamic model framework is its flexibility. It can be modified to incorporate proteins such as nucleosomes, with one group showing that nucleosomes dynamics could create the appearance of cooperativity between TFs in yeast (Raveh-Sadka et al. 2009). These examples demonstrate the flexibility of statistical thermodynamic models in learning about *cis*-regulatory mechanisms such as cooperativity between TFs.

Previous work in the Cohen lab has used thermodynamic modeling to explain the expression driven by synthetic promoters and learn about *cis*-regulatory logic in yeast. These synthetic promoters are comprised of chains of binding sites for a few TFs, providing a great system for discovering interactions between TFs in a controlled sequence environment. The expression driven by hundreds of synthetic promoters was measured in various conditions in yeast. Expression was best explained by incorporating interactions between TFs into a thermodynamic modeling framework and allowing TF concentration to vary across conditions (Gertz et al. 2009; Gertz and Cohen 2009). Additional modifications to the thermodynamic model further demonstrate the flexibility that it provides. The same features in the model predicted expression even in the face of changes in the strength of the TATA box, showing that the TATA box does not affect the combinatorial TF interactions (Mogno et al. 2010). The model can also incorporate measured TF occupancy, which allows for the discovery of additional TF interactions (Zeigler and Cohen 2014). While these important studies laid the groundwork for using thermodynamic models with synthetic sequences, they were restricted to yeast. Applying this flexible model framework to a mammalian system will allow us to determine the extent to which TF interactions play a role in important developmental processes.

Regulation of Pluripotency

Embryonic stem (ES) cells are one of the more important model systems used to study development, cellular biology, and genomics. ES cells are formed from the inner cell mass of the developing mammalian embryo, and have the ability to self-renew almost indefinitely when plated in culture (Chambers and Tomlinson 2009; Martin 1981; Evans and Kaufman 1981). They are able to differentiate into many different cell types in culture, form teratomas *in vivo*, and even contribute to whole animals (Martin 1981; Evans and Kaufman 1981; Nagy et al. 1993). Pluripotency is the property that gives ES cells these abilities of differentiation. Formally, it is defined as the ability to differentiate into any of the three primary germ layers (Solter and Solter 2006). ES cells are thus a great model system for studying mammalian development.

Pluripotency is also an excellent model for studying combinatorial regulation of transcription. Pluripotency is tightly regulated through a transcriptional network that is essential to maintaining the ES cell fate. A set of pluripotency TFs including Oct4, Sox2, Nanog, Esrrb, Klf2, Klf4, Klf5, and c-Myc are important for the maintenance of pluripotency and self-renewal of ES cells (Masui et al. 2007; Chambers et al. 2007; Niwa et al. 2000; Li et al. 2005; Ema et al. 2008; Yeo et al. 2014; Ivanova et al. 2006; Nishiyama et al. 2013). Oct4 is a Pou-domain containing TF, Sox2 is from the HMG family of TFs, and Esrrb is an orphan nuclear receptor (Ambrosetti et al. 2000; Feng et al. 2009). Klf2, Klf4, and Klf5 are all zinc finger TFs from the same family (Jiang et al. 2008). These TFs work to transcriptionally regulate a network that both activates genes important to the self-renewal of ES cells as well as represses genes that would lead to differentiation. This network includes a significant amount of cross-regulation, in which the pluripotency TFs themselves are frequent targets of regulation (Chambers and Tomlinson 2009; Chen et al. 2008c; Kim et al. 2008; Boyer et al. 2005; Ivanova et al. 2006; Loh et al. 2006;

Liu et al. 2008; Nishiyama et al. 2009, 2013). Underscoring their importance, the knockdown of many of these TFs, including Nanog, Oct4, Sox2, and Esrrb, leads to loss of self-renewal in ES cells (Ivanova et al. 2006). In addition to their role in maintaining the ES cell state, some of these TFs also help guide differentiation down certain developmental pathways, and overexpression of some of these TFs can lead to differentiation (Thomson et al. 2011; Teo et al. 2011; Niwa et al. 2000). This demonstrates the importance of the action of the whole pluripotency network to maintaining pluripotency and shows that precise transcriptional regulation is key to maintaining the ES cell state. These features make pluripotency, and by extension ES cells, a great model system for studying transcriptional regulation.

Not only are most of these TFs necessary for pluripotency and self-renewal of ES cells, some of them are even sufficient to induce pluripotency in other cell types. This was first shown by converting fibroblasts to induced pluripotent stem (iPS) cells through ectopic expression of Oct4, Sox2, Klf4, and c-Myc in both mouse and human (Takahashi and Yamanaka 2006; Takahashi et al. 2007). These iPS cells posses all the features that make ES cells pluripotent, including the ability to contribute cells to an entire animal. iPS cells can be generated from a number of cell types from all three germ layers, including hematopoietic cells at various stages of differentiation (Eminli et al. 2009), pancreatic β cells (Stadtfeld et al. 2008), and adult neural stem cells (Kim et al. 2008). Later studies refined the picture of what sets of TFs could induce pluripotency. c-Myc was found to be unnecessary for iPS cell generation. Esrrb could be paired with Oct4 and Sox2 to generate iPS cells, and either Klf2 or Klf5 could increase the efficiency of iPS generation when paired with Oct4, Sox2, and Klf4 (Nakagawa et al. 2008; Feng et al. 2009). Oct4 has even been shown to induce pluripotency by itself when paired with a small molecule

(Li et al. 2010; Yuan et al. 2011). This evidence all points to the central role that TFs play in regulating the property of pluripotency.

The TFs that maintain the property of pluripotency often work together to regulate gene expression in ES cells. Groups of these TFs tend to bind common genomic loci to regulate the expression of the same genes important to maintaining pluripotency (Chen et al. 2008c; Kim et al. 2008; Boyer et al. 2005; Loh et al. 2006). For instance, Oct4, Sox2, and Nanog bind a distinct set of loci, with Klf4 and Esrrb joining them in many locations (Boyer et al. 2005; Chen et al. 2008c; Kim et al. 2008). In addition to binding common targets, clusters of bound pluripotency TFs are also associated with transcriptional regulatory signals. P300, a transcriptional co-activator, is associated with clusters of bound pluripotency TFs, and knocking down these TFs can lead to lower P300 binding (Chen et al. 2008c). Furthermore, ES cell gene expression is associated with groups of pluripotency TFs bound at a gene's promoter (Kim et al. 2008). For certain sets of these TFs, there is even evidence of cooperation or physical interaction in regulating genes. Oct4 and Sox2 are known to physically interact and bind a joint sequence motif to regulate transcription in ES cells (Ambrosetti et al. 2000; Chew et al. 2005; Kuroda et al. 2005; Ng et al. 2012; Rodda et al. 2005). Additionally, there is some evidence of cooperativity of Nanog with Sox2 (Gagliardi et al. 2013), Klf4 with Sox2/Oct4 (Nakatake et al. 2006; Wei et al. 2009), Oct4 with Esrrb (van den Berg et al. 2008, 2010), and Klf4 with Oct4 (Wei et al. 2013). These previous observations suggest that *cis*-regulatory logic is important to specify the transcription driven by the pluripotency TFs. A new method is necessary in order to gain the statistical power to fully assay the degree to which these interactions play a role. Synthetic *cis*-regulatory elements coupled with high-throughput assays give us this opportunity.

High-throughput reporter assays and synthetic CREs

In order to study the *cis*-regulatory properties of regulatory elements at a general level, it is necessary to study many hundreds to thousands of sequences in order to gain enough statistical power. In the case of combinatorial regulation, the sequences need to sample many different binding sites combinations. This necessitates an assay that can be used to measure expression driven by many sequences. Recent work by a number of groups, including the Cohen lab, has led to the development of massively parallel reporter assays to more easily assay the *cis*-regulatory potential of up to thousands of sequences in one experiment (Arnold et al. 2013; Kwasnieski et al. 2012; Melnikov et al. 2012; Sharon et al. 2012; Smith et al. 2013a; Patwardhan et al. 2012a). The ease of using these assays has allowed for more high-throughput work into transcriptional regulation of mammalian systems. Most of these assays use a reporter gene with a barcoded transcript to read out the expression level driven by a set of *cis*-regulatory elements using next-generation sequencing. This has been used in a number of systems including yeast, mammalian cell culture, and mouse tissues. The flexibility of these assays makes them ideally suited for a study into the mechanisms of *cis*-regulation.

Focus of Dissertation

My dissertation has focused on the determinants of *cis*-regulation. Gene regulation is essential to proper development of multicellular organisms; as such, mammalian systems are a great for studying how important transcriptional processes. In general, predictions of expression driven by a given sequence can use two types of data: 1) the DNA sequence of the region in question, and 2) measured biochemical features at the given region. Predictions based on sequence often use TF binding motifs but can also use k-mers or simpler features such as GC content. Predictions based on measured biochemical features usually use the presence of histone modifications or open chromatin at the region of interest. In Chapters 2 and 3 I use sequence features, specifically TF binding sites and combinations of these sites, to learn about the regulation of expression by pluripotency TFs in ES cells. In Chapter 4, I make predictions of expression based on both histone modifications and sequence features. These complementary approaches allow for a picture to emerge of mechanisms of *cis*-regulation.

The technical basis of much of my thesis is the ability to measure expression driven by a large library of sequence elements. This is made possible by a massively parallel reporter assay developed in the Cohen lab, CRE-seq (Kwasnieski et al. 2012). It allowed me to easily assay the expression in two cell types using mostly the same protocol. I used both ES cells and K562 cells, a leukemia cell line. I was also able to measure expression in differentiated ES cells. Large numbers of measurements such as those made possible by CRE-seq are essential to learning about mechanisms of transcriptional regulation, especially given the complex picture that is emerging.

In Chapter 2 I used the ES cell system to investigate the degree to which interactions between TFs in the pluripotency network determine the expression driven by *cis*-regulatory elements (CREs) in ES cells. In this work, I assayed the expression of a library of synthetic CREs comprised of many combinations of binding sites for TFs in the pluripotency network. Using a statistical thermodynamic model I found that TF interactions are important for specifying *cis*-regulation in ES cells. This was followed up by bioinformatic analysis of genomic binding sites to support the interactions between TFs. I further investigated homotypic interactions through a small library of CREs comprised of chains of Klf4 sites in various overexpression conditions. I found that competition between Klf factors helps determine expression driven by these chains. I conclude that interactions between certain pluripotency TFs, but not all, help specify expression level, and that exact grammar is important in regulation in some cases.

In Chapter 3, I continued examining the expression of combinations of binding sites for pluripotency TFs in a new condition. I treated ES cells with retinoic acid (RA), which causes them to differentiate down the neural lineage. I then measured the expression of the library of synthetic CREs from Chapter 2 in these cells and compared to the expression in ES cells. I found that CREs with complex mixtures of binding sites had similar expression in both cell types whereas other CREs did not. This indicates that heterotypic interactions are largely consistent between cell types but that homotypic interactions tend to change.

The work in Chapter 4 looked at genomic sequence elements and biochemical features that were associated with high *cis*-regulatory potential. A set of genomic segmentation predictions for enhancers or repressed elements was assayed for their ability to drive expression in K562 cells. Each class drove a distinct expression level, and Enhancers and Weak Enhancers drove expression over background. Interestingly, Weak Enhancers drove higher expression than Enhancers, potentially due to lower H3K27ac. Additionally, I built a logistic regression model using this expression data to predict expression based on chromatin and sequence features, finding that matches to TF binding motifs are best at explaining expression. From these results, we can conclude that both features relating to sequence and biochemical modifications are needed to fully explain gene expression.

The work in this dissertation lays out important principles of *cis*-regulation relating to the role of TFs and chromatin modifications. It lays the groundwork for deciphering the *cis*-regulatory code from the features that we believe have some effect but are unsure of their exact function. Specifically, it investigates the degree to which interactions between TFs regulate expression and the relative contribution of biochemical signals and sequence to expression. As others build upon this work, we will be able to get closer to the goal of quantitative understanding of gene regulation.

Chapter 2: Interactions between transcription factors help specify cis-regulation in pluripotency

A core set of transcription factors (TFs) maintain the pluripotent state in embryonic stem (ES) cells, and a number of these TFs bind to similar regions of the genome. We investigated the degree to which interactions between these TFs affect *cis*-regulation in embryonic stem (ES) cells. To this end, we measured the expression of a library of hundreds of synthetic *cis*-regulatory elements (CREs) comprised of binding sites for Oct4, Sox2, Klf4, and Esrrb in ES cells. CREs with mixtures of unique types of TF binding sites drive the highest expression. A statistical thermodynamic model that incorporates interactions between TFs explains a large portion (72%) of the variance in expression of these CREs. These interactions include favorable interactions between Oct4 and Sox2, Klf4 and Sox2, and Klf4 and Esrrb, and are supported by genomic binding data. Interestingly, an unfavorable homotypic interaction was also a strong component of the model, helping to explain the finding that CREs with many unique binding sites drive the highest expression. We went on to investigate the expression driven by CREs comprised of homotypic chains of Klf4 sites, which can be bound by Klf2, Klf4, and Klf5 in ES cells. Our results show that each of the Klf TFs has a unique contribution to regulation by these CREs. This suggests that competition between Klf2, Klf4, and Klf5 for binding at the Klf4 binding site is an important element of regulation by chains of these sites.

Introduction

Transcription is regulated primarily through the action of sequence-specific transcription factors (TFs), which help direct *cis*-regulatory programs and guide mammalian development. TFs bind DNA at specific degenerate sequences. These sequences can be modeled using position weight matrices, which can be used to predict TF binding sites (Stormo and Fields 1998). In mammalian systems, few of these predicted sites are actually bound by TFs and fewer still regulate transcription. As such, we lack a general method to quantitatively predict expression level driven by TFs. There are a few mechanisms that may affect TF activity at predicted binding sites, including chromatin accessibility, DNA methylation, and combinatorial interactions between TFs. Combinatorial regulation has been shown to be important in specifying *cis*-regulation in a number of systems (Gertz et al. 2009; Smith et al. 2013a; Thanos and Maniatis 1995; Kim and Maniatis 1997; Yáñez-Cuna et al. 2012; Sharon et al. 2012). However, interactions only occur between certain pairs of TFs, and its unclear what combinatorial interactions exist in which systems. A better understanding of combinatorial regulation would allow us to better predict which *cis*-regulatory elements (CREs) are active, what expression level they drive, and the effect of perturbations on their expression.

TFs play an essential role in the specification of cell fate through regulation of developmental genes. Pluripotency, the property that allows embryonic stem (ES) cells to differentiate into any of the three primary germ layers, is maintained by a core set of TFs including Oct4, Sox2, Klf2, Klf4, Klf5, c-Myc, Nanog, and Esrrb (Chen et al. 2008c; Boyer et al. 2005; Loh et al. 2006; Ivanova et al. 2006). The ectopic expression of certain combinations of these TFs is able to alter the state of differentiated cells back to pluripotency, a process known as induced pluripotency (Takahashi and Yamanaka 2006). This demonstrates the importance of these TFs to pluripotency. There is also evidence that the pluripotency TFs act cooperatively in ES cells. They often bind in clusters at common genomic loci to regulate a core set of genes, and a set of 25 tested regions bound by three of these TFs all have enhancer activity in ES cells (Chen et al. 2008c; Kim et al. 2008). However, the expression level driven by these regions varied over a 25-fold range, suggesting that there may be some unknown combinatorial rules dictating the expression level of these regions. In fact, some of these TFs, most notably Oct4 and Sox2, have been shown to physically interact to regulate transcription (Chew et al. 2005; Kuroda et al. 2005; Rodda et al. 2005). It's likely that combinatorial regulation plays a role in the pluripotency network, but the rules of this regulation need to be determined.

Synthetic *cis*-regulatory elements (CREs) are useful tools for investigating *cis*-regulatory mechanisms, especially when coupled with high-throughput expression assays (Gertz et al. 2009; Smith et al. 2013b; Sharon et al. 2012; Mogno et al. 2013). Controlling the arrangement and numbers of TF binding sites in a controlled sequence background provides substantial power to discover interactions between TFs. In addition, massively parallel reporter assays using next-generation sequencing can be used to determine the *cis*-regulatory potential of many sequences in a number of systems (Kinney et al. 2010; Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012b; Sharon et al. 2012; Arnold et al. 2013). Here we demonstrate the use of a massively parallel reporter assay, CRE-seq, to measure the expression of hundreds of synthetic CREs in mouse ES cells. We then used this expression data along with a statistical thermodynamic model to discover interactions between pluripotency TFs and their effect on transcriptional regulation.

Results

We first designed a synthetic CRE library to investigate combinatorial *cis*-regulation by pluripotency TFs. Our goal was to test two possible *cis*-regulatory scenarios: 1) TFs each act independently to regulate transcription, or 2) interactions between TFs help guide regulation. We used synthetic CREs comprised of between one and four high-affinity binding sites for the following TFs central to pluripotency: Oct4, Sox2, Klf4, and Esrrb. Each binding site is in a 20bp building block with at least 8bp of constant sequence surrounding it (Figure 2.S1a). We designed a library (OSKE library) of 599 of these synthetic CREs and measured their expression in ES cells using CRE-seq, a massively parallel reporter assay (Kwasnieski et al. 2012). The synthetic CRE library was cloned into a plasmid vector, upstream of the Pou5f1 basal promoter and the dsRed gene (Figure 2.S1b). Each CRE is associated with 10 unique sequence barcodes (BCs) situated in the 3' UTR of the dsRed gene. We transfected the OSKE plasmid library into mouse ES cells and measured the expression of the CREs 26 hours later using CRE-seq. After filtering, expression data was obtained for 3567 BCs and 415 CREs. The expression measurements were reproducible across three biological replicates (R^2 range=0.88-0.91, Figure 2.S1c).

We first investigated whether simple additivity could explain the expression driven by the CREs. CREs with more transcription factor binding sites (TFBS) tended to be more highly expressed; however, the number of binding sites could not explain all of the expression (R^2=0.14), and the expression of CREs with four TFBS varied over a 13-fold range. This suggests that there is regulation by these TFs that involves more complex effects than simple additivity.

To investigate whether interactions between TFs could explain the expression of the synthetic CRE library, we modeled the expression using a statistical thermodynamic framework (Shea and Ackers 1985; Buchler et al. 2003; Gertz et al. 2009; Kinney et al. 2008; He et al. 2010; Zeigler and Cohen 2014; Brewster et al. 2014). This model framework relates the change in free energy of interactions between proteins and DNA on a CRE to the expression driven by that CRE. Specifically, it includes parameters for the interactions between TFs, RNA Polymerase II (RNAP), and DNA. The probability that RNAP is bound at the CRE is assumed to be proportional to the expression driven by that CRE, connecting the TF-TF interactions and expression. For a specific CRE, the probability that RNAP is bound is calculated by summing the weights of all permutations of bound TFs (Fig 2.S2). This model framework also allows for specific rules about how TFs may interact, and we have used two basic rules for TF-TF interactions: the "neighboring" interactions rule and the "all-across" interactions rule (Materials and Methods). By finding a set of parameters that best predicts expression, we can determine which TFs may interact in this system.

We tested whether a thermodynamic model incorporating interactions between TFs could explain the expression of the CREs better than a model without any TF interactions. We first used a baseline thermodynamic model including only four TF-RNAP interaction parameters, one for each TF, and no TF-TF interaction parameters. This model predicted expression with an R^2 of 0.50 (Figure 2.1a). In order to determine whether any TF interactions contribute to expression in this system, we added TF-TF interaction parameters to the model to see if they could improve the predictive power of the model. To guard against overfitting, we monitored the AIC and sensitivity of the parameter values and used five-fold cross validation. The best model with TF-TF interactions ('Full Model') includes nine total parameters and explains the expression with an

R^2 of 0.72 (Figure 2.1b). The Full Model includes the four TF-RNAP interaction parameters, as well as five parameters for interactions between TFs: Sox2-Oct4, Klf4-Sox2, Klf4-Esrrb, homotypic (an interaction between any TF and another copy of itself) on the same strand, and homotypic on the opposite strand (Fig 2.1b, Table 2.S1). Analysis of genomic ChIP-seq binding data (Chen et al. 2008c) for the Klf4-Esrrb pair and the Klf4-Sox2 pair suggests that they bind cooperatively in the genome (Figure 2.S3). Notably, the homotypic interaction parameters are strongly unfavorable, with some of the largest parameter values in the model (Table 2.S1). Despite this, the other TF-TF interaction parameters still significantly contribute to the model. This is best demonstrated by the fact that the Full Model explains 36% of the variance in expression of the 20 CREs that have one binding site for each of the four TFs, despite the fact that these CREs all have the same total TF-RNAP interactions and do not utilize the homotypic interaction parameters. Thermodynamic modeling has shown that interactions between TFs help determine the expression driven by CREs in the pluripotency network.

The unfavorable homotypic interaction parameters suggest that CREs regulated by a variety of TFs, rather than multiple copies of the same TF, are best able to activate transcription. Indeed, among synthetic CREs with four total TFBS, higher expression is strongly associated with the number of unique types of binding sites in the CRE (Figure 2.2). In other words, CREs with one binding site for each of the four TFs in the library drive much higher expression than CREs with four copies of a binding site for a single TF. In addition, of those CREs with binding sites for two TFs and four total binding sites, the expression is higher when there are two binding sites for each TF than when there are three sites for one TF and one site for the other TF (Figure 2.S4). Interestingly, the model also finds that two copies of the same binding site in the opposite orientation produce a more unfavorable interaction than two copies in the same orientation

(Table 2.S1). While it is clear that reduced expression of homotypic chains plays a role in *cis*-regulation, the mechanism by which this occurs is unclear.

We hypothesized that the expression driven by homotypic chains of TFBS is regulated by competition and interactions between TFs that bind the same binding site. To test this hypothesis, we assayed the expression of CREs comprised of chains of binding sites for Klf4, the site most associated with activation in the original CRE library. Klf2, Klf4, and Klf5 all regulate pluripotency in ES cells and are known to bind to the Klf4 binding site (Jiang et al. 2008). We created a small CRE library (KBS library) with 7 synthetic CREs, comprised of between zero and six Klf4 sites. We measured the expression of the KBS library in ES cells in the context of overexpression of each of the three Klf TFs as well as a control gene (GFP).

The expression profile of the CREs in the KBS library was unique in each TF overexpression condition (Figure 2.3). Each CRE with 3 or more binding sites drives higher expression in the Klf2 overexpression condition than in the control overexpression condition. Interestingly, Klf4 and Klf5 overexpression leads to lower expression from CREs with 5 or 6 binding sites compared to the control overexpression condition. Furthermore, under Klf5 overexpression, a CRE with 6 binding sites drives lower expression than one with 5 binding sites. We also measured the expression level of the Klf genes in the overexpression conditions by qPCR and found that Klf4 was up-regulated in the Klf5 overexpression condition (Fig 2.S5). This data suggests that each Klf TF has a unique effect on the expression of these CREs and that competition between them is an important factor in regulation by these binding sites.

We then used a statistical thermodynamic model to learn what mechanisms could explain the expression driven by chains of Klf4 binding sites. We attempted to fit a model using homotypic and heterotypic interactions between the Klf factors. The model that best explained the data without overfitting used four parameters: one for each TF-RNAP interaction, and a Klf4 homotypic interaction parameter (Table 2.S1). This model could capture most of the overall variation in expression (R^2=0.93) and explained the data significantly better than a model with no TF-TF interactions (by AIC) (Figure 2.S6). All interactions were found to be favorable, but Klf2-RNAP and Klf4-RNAP were very weak and their 95% confidence interval spanned 0 (Table 2.S1). This suggests that Klf4 may be a weak activator with strong self-cooperativity that can prevent Klf2 and Klf5 from binding, especially on longer chains of binding sites and when it is highly expressed.

We next investigated whether clusters of many Klf4 binding sites in the genome are associated with activity and could be regulated by a similar mechanism as our synthetic CREs. We used a sliding window approach to learn the relationship between the number of Klf4 binding motifs in a window and biochemical activity. We found a positive trend between the number of Klf4 binding sites and both DNase hypersensitivity signal and RNA Polymerase II binding signal, up to six binding sites (Figure 2.S7). This effect is not due to GC-content, as demonstrated by permuted Klf4 binding matrices, and is stronger within 10kb of a TSS. This demonstrates that genomic clusters of Klf4 binding sites have biochemical activity that is associated with the number of binding sites, up to six binding sites. Thus, these clusters may regulate transcription in a similar manner as the synthetic CREs, namely through competition and interactions among the Klf factors.

Discussion

Here we describe an investigation into the *cis*-regulatory activity of TFs in the pluripotency network. Using a statistical thermodynamic model, we show that interactions between TFs play a large role in explaining the expression driven by synthetic CREs (22% of the variation in expression after accounting for basic TF-RNAP interactions). We further characterized regulation by homotypic chains of one particular binding site, Klf4, and found that Klf2, Klf4, and Klf5 each have unique effects on transcription from these sites.

Two principal observations lead us to believe that there is *cis*-regulatory logic in the pluripotency network. First, the expression of synthetic CREs with one binding site for each of the four TFs varies over a 3-fold range. Second, TF-TF interactions parameters make important contributions to the thermodynamic model. These interaction parameters included three favorable heterotypic interactions and two strong unfavorable homotypic interactions. The Oct4-Sox2 interaction found by the model is a well-known and characterized interaction (Chew et al. 2005; Kuroda et al. 2005; Rodda et al. 2005). There is some previous evidence for the Klf4-Sox2 interaction (Nakatake et al. 2006; Wei et al. 2009; Xie et al. 2008), while the Klf4-Esrrb interaction is mostly novel. The expression of the synthetic CREs shows that these interactions are indeed important to determining the expression driven by binding sites for these TFs. We have provided a framework demonstrating the quantitative contribution of these interactions to expression.

Our work builds on previous studies showing that homotypic clusters of TF binding sites have unique *cis*-regulatory properties. Homotypic clusters have been associated with conservation and function, as well as reduced ability to drive expression compared to heterotypic clusters (Gotea et al. 2010; Ridinger-Saison et al. 2012; Sharon et al. 2012; Smith et al. 2013b). Our results from the OSKE library also show that homotypic chains of binding sites drive lower expression than heterotypic chains (Fig 2.2). This agrees with another finding that ES cell promoters that are bound by only one pluripotency TF tend to be off and those that are bound by multiple TFs tend to be on (Kim et al. 2008). Interestingly, homotypic chains of Klf4 binding sites from the KBS library drove higher expression than homotypic chains of other binding sites from the OSKE library. This may be explained by the fact that competition between Klf factors (Klf2, Klf4, and Klf5), each of which has a unique effect on expression, plays a role in regulation from these CREs. Our thermodynamic modeling suggests that Klf4 may cooperatively bind with itself, and in the process work to prevent binding by Klf2 and Klf5, which may be the strongest activator. While this model can explain most of the variation in the expression, it cannot capture all of the trends in the expression (Figure 2.S6). For instance, it was unable to capture the lack of increase in expression from five to six binding sites. These results combined with previous studies show that homotypic chains of binding sites are important to *cis*-regulation but their effect may vary based on the binding site and system.

A network of TFs regulates pluripotency in ES cells, and we show here that interactions between these TFs are necessary to specify their *cis*-regulation. The evidence suggests that knowledge of which binding sites are present is not enough to determine the expression driven by a CRE. The ability to fully understand gene expression, and any perturbations to it, will depend on knowing the effect of interactions between TFs.

Materials and Methods

Design of CREs

For the OSKE library, the binding sites were comprised of consensus binding sites for the following four TFs: Oct4, Sox2, Esrrb, and Klf4 (Supplemental Info). These 12bp binding sites were inserted into a larger 20bp building block sequence, of which the other 8bp were constant. For CREs with multiple binding sites, the building blocks were concatenated together. For the

KBS library, a alternate Klf4 binding site with high affinity was used to facilitate cloning (Supplemental Info).

Cloning of plasmid libraries and overexpression plasmids

Plasmid pCF10 was constructed from pGL4.23 (Promega), first by inserting the dsRed-Express2 gene between the Acc65I and FseI sites. Then, the Pou5f1 basal promoter was inserted between the NcoI and HindIII sites. pCF10 served as the basic plasmid backbone. Array synthesized oligos (6,500 unique sequences of 150bp long) were ordered from Agilent through a limited licensing agreement. The oligos were comprised of two primer sequences, a CRE, a 9bp barcode (BC), and multiple restriction enzyme sites (see Supplemental Information). The OSKE library comprised of 599 CREs, each associated with 10 BCs, and the basal promoter alone, associated with 30 BCs. The rest of the array contained CREs not used in this study. The array synthesized oligos were prepared as previously described (Kwasnieski et al. 2012), except using primers CF159 and CF160 with an annealing temperature of 55C for the initial PCR step, and then purified from a polyacrylamide gel as described previously (White et al. 2013). These were cloned into plasmid pCF10 at the ApaI and ScaI sites. The Pou5f1 basal promoter and dsRed were then amplified from pCF10 using primers CF121 and CF122 and then inserted into the plasmid library from the previous step at the XbaI and HindIII sites. Plasmids without the basal promoter and dsRed were filtered out by cutting in the backbone at the SpeI site and gel extracting the band at the appropriate size. This formed the OSKE library.

The Klf4 binding sites (KBS) library was formed by cloning individual CREs into the reporter plasmid. CREs with Klf4 binding sites were ordered from IDT (oligos BS300-BS308, Supplemental Data 2.1). Oligonucleotides (oligos) were cloned into pCF10 at sites HindIII and ApaI, upstream of the basal promoter and dsRed gene in the same location as the CREs in the

OSKE library. Two control plasmids were also constructed, with the hsp68 promoter and the SV40 promoter. pGL-hsp68 was constructed as described previously from pCF10 (Kwasnieski et al. 2014). A plasmid with the SV40 promoter was cloned by inserting the SV40 promoter from the pbDonor-tdTomato plasmid (a gift of the Rob Mitra lab) using primers CF134 and CF135 at the NcoI and HindIII sites of pCF10. Oligos with BCs were then inserted into the plasmids containing the CRE inserts. First, oligos CF48 and CF49, containing random 12bp BCs, were annealed. Next, these annealed oligos were cloned into the plasmids with CRE inserts at the XbaI and SacI sites. 12 colonies containing random BCs for each CRE plasmid were picked and used to comprise the KBS library. The BCs in the plasmids were then sequenced by Sanger sequencing, and only those plasmids with a BC insert were retained.

Overexpression constructs were constructed based on the pCX-OKS-2A plasmid. The individual TF genes for Klf2, Klf4, Klf5, and GFP were inserted between the EcoRI sites of the pCX-OKS-2A plasmid. Klf4 sequence was taken from the pCX-OKS-2A plasmid, Klf2 was taken from the pMXs-ms-Klf2 plasmid, and Klf5 was taken from the pMXs-ms-Klf5 plasmid. pCX-OKS-2A (Addgene plasmid #19771), pMXs-ms-Klf2 (Addgene plasmid #50786), and pMXs-ms-Klf5 (Addgene plasmid #50787) were gifts from Shinya Yamanaka.

Cell culture and transfection

RW4 ES cells were cultured as described previously (Chen et al. 2008b; Xian et al. 2005), on gelatin coated plates and in media comprised of: DMEM, 10% fetal bovine serum, 10% newborn calf serum, nucleoside supplement, 1000 U/ml leukemia inhibitory factor (LIF), and 0.1uM β-mercaptoethanol. For transfection of OSKE library, ES cells in a 6-well plate were transfected using 10ul Lipofectamine 2000 (Life Technologies), 3ug plasmid library and 0.3ug CF128 (GFP plasmid control) per well. For transfection of KBS library and the Klf overexpression plasmids, ES cells in a 6-well plate were transfected with 10ul Lipofectamine 2000 (Life Technologies), 2.25ug Klf overexpression plasmid (CF127, CF128, CF131, or CF136), 0.75ug KBS library, and 0.3ug CF128. The cells were passaged 6 hours post-transfection, and RNA was extracted 26 hours post-transfection using the PureLink RNA mini kit (Life Technologies). Three replicates of each sample (OSKE library and the KBS library in each overexpression condition) were transfected and processed.

CRE-seq

Expression measurements of each CRE were determined using CRE-seq as described previously, using Illumina sequencing of both the RNA and original plasmid DNA pool (Kwasnieski et al. 2014). Briefly, excess DNA was removed from the RNA using the TURBO DNA-free kit (Applied Biosystems). cDNA was then prepared using SuperScript RT II (Life Technologies) with oligo dT primers. Both the cDNA and the plasmid DNA pool were amplified using primers CF150 and CF151b, using 21 cycles. The PCR amplification products were digested using XbaI and XhoI (NEB), and the resulting digestion products were ligated to custom Illumina adapter sequences P1_XbaI_BCX (where X is 7 through 15) and P2_XhoI, each of which is comprised of a forward (F) and reverse (R) strand that were annealed. An enrichment PCR step of 20 cycles with primers CF52 and CF53 was then used, and the resulting product was sequenced on one lane of the Illumina HiSeq for the OSKE library, and part of a lane on the Illumina MiSeq for the KBS library.

Sequencing reads were filtered to ensure that the first 13 nucleotides perfectly matched the expected sequence. For the OSKE library, this resulted in 64.3 million reads combined for the three RNA samples, and 24.5 million reads for the DNA sample. For the KBS library, this resulted in 1.79 million reads combined for the 12 RNA samples and 181,000 reads for the DNA sample. The expression of each barcode (BC) in each sample was calculated as (RNA read count)/(DNA read count), and only BCs passing a read count threshold were included for further analysis. The read count thresholds were 2000 reads in the OSKE library DNA sample, 50 reads in the KBS library DNA sample, and 3 reads in the RNA samples. The expression of each CRE in each replicate was calculated as the mean of the expression of each BC it was associated with, and only CREs with at least 3 BCs passing the read count filter in each replicate were included in the analysis. The overall expression of each CRE was the mean of its expression in each replicate. For the KBS library, the expression of each CRE in each overexpression condition was normalized to the expression of the SV40 CRE in that condition.

qPCR of Klf genes

For quantification of gene expression level of Klf2, Klf4, and Klf5 in overexpression conditions, RW4 cells were transfected as before but using 2.25ug Klf overexpression plasmid and 0.75ug CF128 (GFP reporter plasmid). 26 hours post-transfection, the cells were resuspended in PBS and sorted on GFP signal using the BD FACSArias III machine into RNAprotect (Qiagen). Cells with the GFP plasmid (CF128) mimic the cells with the KBS library in the previous transfections, as they were each transfected with the same amount of either the GFP plasmid or the KBS library. RNA was extracted using RNeasy Mini Kit (Qiagen), excess DNA removed using TURBO DNA-free kit (Applied Biosystems), and cDNA synthesized using SuperScript RT II (Life Technologies). qPCR was performed using Absolute SYBR Green, low ROX (Life Technologies), with primers listed in Supplemental Table 2.2.

Thermodynamic modeling

The statistical thermodynamic model was implemented as described previously (Buchler et al. 2003; Gertz et al. 2009; Gertz and Cohen 2009; Zeigler and Cohen 2014). The model incorporates parameters for interactions between TFs, RNAP, and binding sites on the DNA. These parameters are proportional to the free energy of interaction. The probability that RNAP is bound at the promoter is assumed to be proportional to the expression driven by that promoter/CRE. For a given CRE, the statistical weight of each possible binding configuration is calculated, and the probability of RNAP being bound is the sum of the weights of all configurations in which RNAP is bound over the sum of the weights of all configurations. See supplementary information for more details.

TF interaction rules dictate when two TFs are allowed to interact. Only when TFs are allowed to interact does the interaction parameter contribute to the statistical weight of a given binding configuration. We have used two basic rules for TF-TF interactions based on whether two TFs can interact if another protein in bound in between. The “neighboring” interaction rule only allows TFs to interact if no other TFs are bound in between them in a particular binding state. The “all-across” interaction rule allows TFs to interact with any other TF bound in that particular binding state (Figure 2.S2). The functional consequence of the neighboring interaction rule is to impart dependence on the order of the binding sites in the CRE. The all-across interaction rule does not distinguish between different orders of binding sites for a particular combination of sites on the CRE.

The model was fit with custom Python scripts (see supplemental information) using SciPy. The fitting routines used for the OSKE library minimized the sum of squared error of the expression measurements, using log-transformed expression (both observed and predicted). The initial starting values for each parameter were set to 0, but the fit was robust to different starting parameter values. Five-fold cross validation was used by splitting the data into a training set of 4/5 of the data and a test set of 1/5 of the data. Each partition of the data was used in the test set exactly once. 95% confidence intervals for the parameter values were calculated using a sensitivity measure based on the asymptotic normal distribution for the parameter estimate. See supplemental information for further details.

For the KBS library, the initial starting parameter values were based on an initial screen of many possible parameter values. The predicted expression of each CRE in the KBS library was calculated using a model with each of 2 million sets of random parameter values, taken from a normal distribution with mean of zero and standard deviation of 1. Each set of parameter values consisted of a value for each of the three TF-RNAP parameters and three of the six possible TF-TF interactions (three homotypic and three heterotypic), with all other parameter values set to zero. Each of the 20 possible configurations of the three TF-TF interaction parameters was sampled one hundred thousand times. The parameter set that predicted expression with the lowest error (using a modified objective function, see below) was used as the starting value to a fitting routine, with only those parameters with non-zero values allowed to be fit. After this fitting routine, insignificant parameters were removed from the model, and a final fitting routine was run.

For the KBS library, a modified objective function was used to take into account the relative expression in each condition as well as the overall expression. The sum of squared error of overall expression for each CRE was calculated as usual, as well as the sum of squared error of the fraction of the maximal expression in the given condition. The geometric mean of both sources of error was calculated as the objective function for fitting. This allows for better predictions of the patterns of expression.

Other statistical analysis and data sources

RNA Polymerase II ChIP-seq and DNaseI hypersensitivity signal data for the mouse genome are from the ENCODE Consortium and were downloaded from the ENCODE UCSC web portal (http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixMouse.html). TF ChIP-seq data is from Chen et al. 2008. All genome coordinates were converted to mm9. Binding matrices were taken from Jaspar (Mathelier et al. 2014). The Sox2 binding matrix was trimmed after the 8th position to exclude the part corresponding to the Oct4 binding site. Similarly, the Oct4 (Pou5f1) binding matrix was trimmed before the 8th position to exclude the part corresponding to the Sox2 binding site. Permuted PWMs were created by randomly permuting the positions in the matrix (thus retaining the nucleotide content). FIMO (Grant et al. 2011) was used for to find predicted binding sites using default options with *P*-value threshold of 10^{-4}. Bedtools (Quinlan and Hall 2010) was used for manipulations and analysis of bed files. Custom scripts were used for other analysis. The sliding window approach used 200bp windows in the genome with a 100bp step size. For each window we counted the number of Klf4 predicted binding sites by matches to the Klf4 PWM and the mean signals of RNA Polymerase II binding and DNase I hypersensitivity.

Acknowledgements

We thank members of the Cohen lab for critical reading of the manuscript, Xuhua Chen and other members of the Mitra lab for cell culture assistance, Joe Planer in the lab of Jeff Gordon for assistance with FACS sorting. We further thank Jamie Kwasnieski and Ilaria Mogno for help with the development of CRE-seq, and Robert Zeigler for assistance with the thermodynamic modeling code.

Supplemental information

Thermodynamic modeling of transcription

Statistical thermodynamic models of transcription are based on the statistical weights of the binding configurations of *cis*-regulatory elements (CREs). They are calculated as described previously (Buchler et al. 2003; Gertz et al. 2009), and are additionally described here. The weight (W) of any given binding configuration is given as:

$W = e^{-(\Sigma q+\Sigma\omega)}$ (1)

ω is the interaction parameter between any two proteins (either two TFs or a TF and RNAP) bound in that binding configuration and allowed to interact. q is the interaction of a TF or RNAP and DNA. It incorporates both affinity and concentration, and is:

$q = k - \ln[TF]$ (2)

where k is a constant equal to $\Delta G^o/RT$. q was fixed at 0 for each TF in the reference condition (no overexpression of any TF for the OSKE library, or overexpression of GFP condition for the KBS library). The relative concentrations of the TFs in the TF overexpression conditions are used to calculate relative q values and these are fixed. The weight of the empty DNA binding state (no TFs or RNAP bound) is set as 1.

For a given CRE, the weights for all possible binding configurations are calculated, and the probability that RNAP is bound is:

$P_{bound} = \frac{\Sigma W_{bound}}{\Sigma W_{bound}+\Sigma W_{unbound}}$ (3)

where W_{bound} is the weight of states in which RNAP is bound, and $W_{unbound}$ is the weight of states in which RNAP is unbound. The probability that RNAP is bound is converted to an expression measurement by scaling it so that the mean expression is the same as in the observed expression measurements.

Competition between Klf TFs in the KBS library was modeling by assuming that all three TFs could bind the Klf4 binding site. A further assumption for simplicity was made that all three TFs bound the site with the same affinity (somewhat supported by (Jiang et al. 2008)) and are present at the same concentration in the cell. While these assumptions may be violated, in practice the model would predict the same trends regardless. The relative concentration of each TF in each overexpression condition was set based on the qRT-PCR measurements in the overexpression conditions. When modeling competition, the possible binding states include each possible Klf TF bound to each Klf4 binding site.

Thermodynamic model fits were done with custom Python scripts using SciPy (scipy.optimize.minimize function). The parameters were fit to minimize the objective function using the L-BFGS-B and SLSQP optimization algorithms in alternating fashion until the parameter values converged. The objective function used was the sum of squared errors of the expression measurements, using the log of the observed expression and the log of the predicted expression. Five-fold cross-validation was used, and the predicted expression value of each CRE in the test sets was used to determine the cross-validation R^2. 95% confidence intervals for the parameter values were calculated using the asympototic normal distribution for the parameter estimate. More details can be found in (Bates and Watts 1988). A set of parameter values was deemed significant if the confidence intervals for all of the parameters were significantly different than zero.

Array sequence and binding sites used in synthetic CREs

The array ordered from Agilent through a limited licensing agreement consisted of 150bp oligos with the following sequence:

ACTACAAGGGCCCA[CRE]AAGCTTCT[FILL]CGTCTAGAC[BC]TGAGCTCTGCAACTCCTACG

Where [CRE] is the CRE comprised of concatenated building blocks of binding sites described below, [FILL] is random filler sequence to bring the length of the sequence up to 150bp (the filler is of variable length depending on the length of the CRE), and [BC] is a random 9bp barcode.

Each building block consisted of 20bp with a binding site sequence in the middle. The binding site sequences, described below, consist of a 12bp sequence. The central 10bp sequences (underlined below) are based on binding sites from the literature or ChIP-seq data. The first position is set as a 'G' in every binding site for consistency, and the last position is set as a 'C' in all binding sites except for Klf4, in which it is set as a 'G' to avoid a restriction site needed in the cloning.

Building Block: AGCTACXXXXXXXXXXXXGT. The 12 Xs are where the binding site sequence goes.

Sox2: GCTCATTGTTTC. Based on the canonical binding site "CATTGTT" (Chen 2008), with "CT" added before from the UTF1 promoter and the "T" added after from the FGF4 promoter (Remenyi 2004).

Oct4: GGGATGCTAATC. Based on the canonical binding site "ATGCTAAT" (Chen 2008), with the “GG” added before from the FGF4 promoter (Remenyi 2004).

Esrrb: GTTCAAGGTCAC. Based on consensus binding sites “TCAAGGTC”A (van den Berg 2008), with the second position (‘T’) from the P2 binding site in the Pou5f1 promoter (X. Zhang 2008).

Klf4 (OSKE library): GGGGCGGGGCCG. Based on the most common binding site matching the Klf4 PWM from Klf4 ChIP-seq peaks (Chen 2008).

Klf4 (KBS library): GGGGTGGGGCCG. Same as above, but with the fifth position changed to ‘T’ to facilitate cloning. The fifth position can be either a ‘T’ or a ‘C’ according to the PWM.

Figures

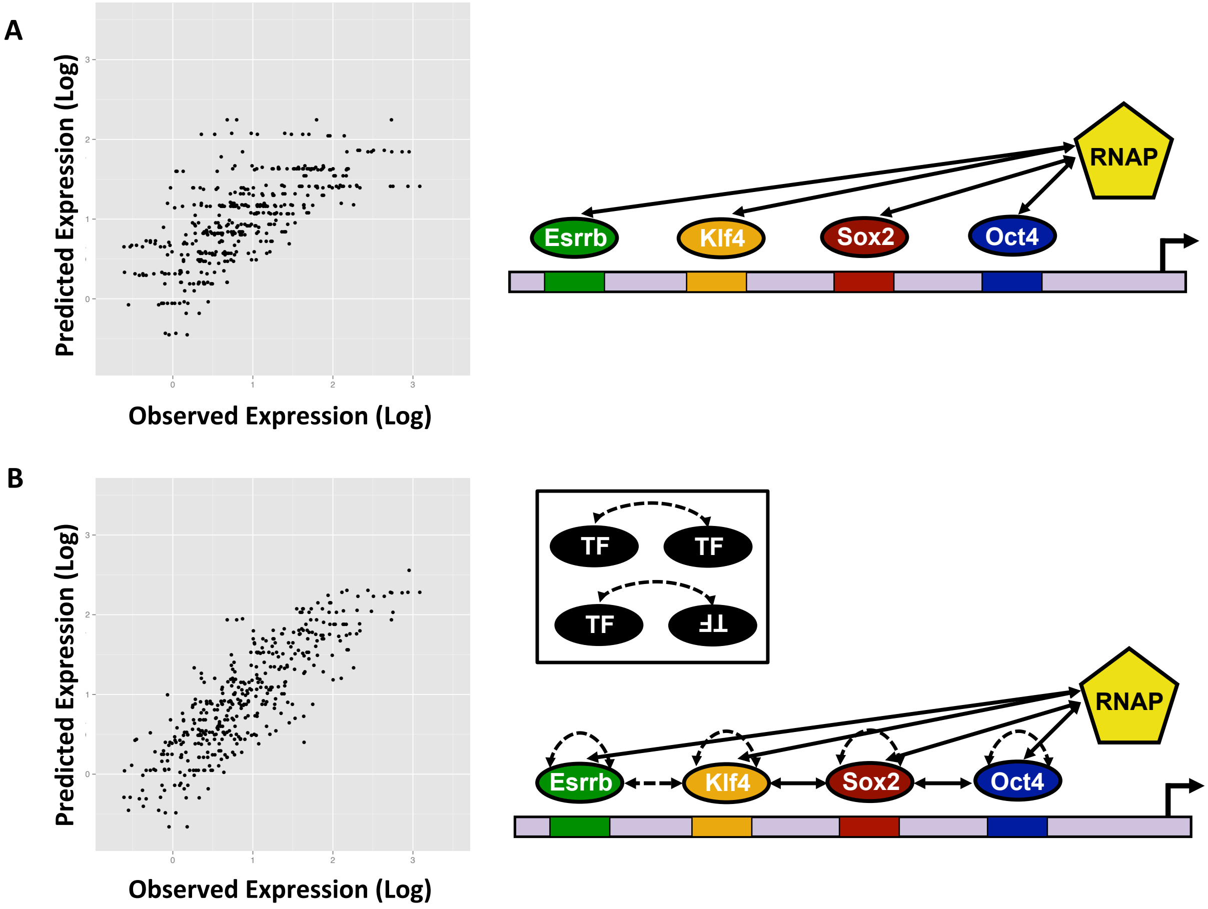


Figure 2.1. Thermodynamic model of OSKE library. In dot plots, observed expression of each CRE from CRE-seq experiment is plotted on the x-axis, and the predicted expression of each CRE by the model is on the y-axis. In depictions of models, solid lines represent interactions only between TFs that are neighboring in a given binding state, and dashed lines represent interactions occurring between any TFs bound on the CRE. A) Model with only four TF-RNAP interaction parameters predicts expression with R^2 of 0.39. B) Full model with five TF-TF interaction parameters in addition to four TF-RNAP interaction parameters predicts expression with R^2 of 0.70.

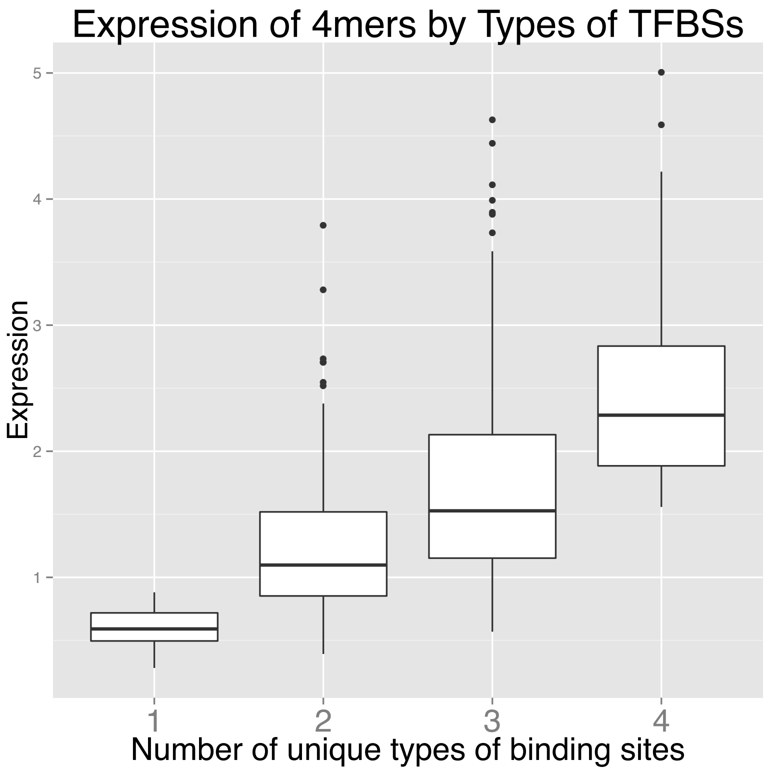


Figure 2.2. Expression by unique types of binding sites. Expression of CREs in the OSKE library with four total TFBS by the number of unique types of binding sites in the CRE.

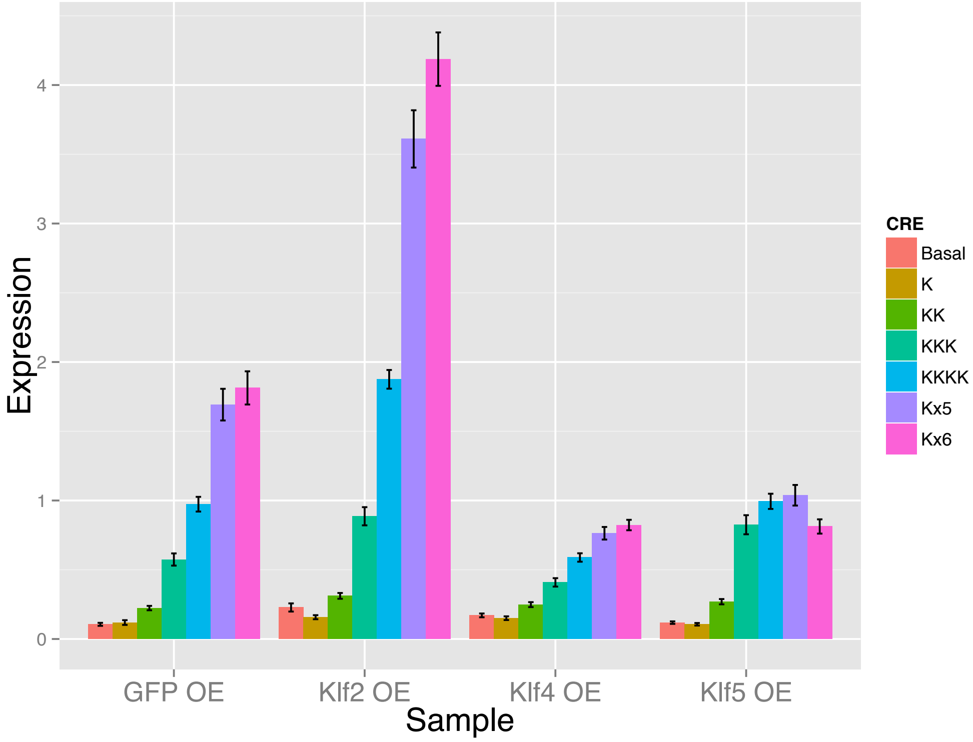


Figure 2.3. Expression of CREs with only Klf4 binding sites. Expression of the basal promoter and CREs with one to six Klf4 binding sites in each of the four overexpression conditions.

Supplemental Tables

OSKE library

Parameter	**Value**	**95% C.I.**
Esrrb-RNAP	0.8915	0.7788, 1.004
Klf4-RNAP	1.06	0.9294, 1.19
Oct4-RNAP	0.5807	0.4729, 0.6884
Sox2-RNAP	0.366	0.2515, 0.4805
Homotypic same orientation (A)	-1.132	-1.476, -0.7874
Homotypic opposite orientation (A)	-2.468	-3.29, -1.65
Klf4-Esrrb (A)	1.119	0.739, 1.499
Oct4-Sox2, only with Oct4 closer to TSS (N)	0.981	0.1595, 1.802
Klf4-Sox2 (N)	1.336	0.8047, 1.867

KBS Library

Parameter	**Value**	**95% C.I.**
Klf2-RNAP	0.1593	-0.2326, 0.5511
Klf4-RNAP	0.2306	-0.05295, 0.5142
Klf5-RNAP	2.169	1.063, 3.274
Klf4-Klf4 (N)	2.369	1.344, 3.393

Table 2.S1: Fit parameter values from thermodynamic models. (N) indicates an interaction with the neighboring interactions rule, (A) indicates an interaction with an all across interaction rule. Positive values are favorable, negative values are unfavorable.

Name	**Sequence**
BS300_KK_upper	CAAGCTACGGGGTGGGGCCGCTAGCTACGGGGTGGGGCCGCT
BS301_KK_lower	AGCTAGCGGCCCCACCCCGTAGCTAGCGGCCCCACCCCGTAGCTTGGGCC
BS302_K_lower	AGCTAgcggccccacccCGTAGCTTGGGCC
BS302_K_upper	caagctacgGGGTGGGGCCGCt
BS303_KKKK_lower	agctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagcttgggcc
BS303_KKKK_upper	caagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCt
BS306_KKK_lower	agctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagcttgggcc
BS306_KKK_upper	caagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCt
BS307_Kx5_lower	agctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagcttgggcc
BS307_Kx5_upper	caagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCt
BS308_Kx6_lower	agctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagctaGCGGCCCCACCCcgtagcttgggcc
BS308_Kx6_upper	caagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCtagctacgGGGTGGGGCCGCt
CF48	CTAGACTNNNNNNNNNNNNCCGAGCT

CF49	CGGNNNNNNNNNNNNAGT
CF52	AATGATACGGCGACCACCGAG
CF53	CAAGCAGAAGACGGCATACGA
CF84	CGAAGTCTGAAGCCAGGTGT
CF90	TCGACGTCaagcttATTGGCACACGAACATTCAA
CF121	TAGCGTCGAGGACATCAAGA
CF122	TGGTTTGTCCAAACTCATCAA
CF134	TCATGTATaagcttTAATGCATGGCGGTAATACG
CF135	TTAGTTtcatgaTGATCAGATCCGAAAATGGA
CF150	TACACCGTGGTGGAGCAGTA
CF151b	AGCGTActcgagTTGTTAACTTGTTTATTGCAGCTT
CF159	ACTACAAGGGCCCAAGC
CF160	CGTAGGAGTTGCAGAGCTC
P1_XbaI_BC7_R	/5Phos/C*TAGAGACTGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC8_R	/5Phos/C*TAGCTTGGAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC9_R	/5Phos/C*TAGCCGATTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC10_R	/5Phos/C*TAGGGCAGCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC11_R	/5Phos/C*TAGCCATCATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC12_R	/5Phos/C*TAGTAACAAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC13_R	/5Phos/C*TAGTCGTAACTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

P1_XbaI_BC14_R	/5Phos/C*TAGGCAGCTATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC15_R	/5Phos/C*TAGCAATCAAGTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
P1_XbaI_BC7_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAGTCT
P1_XbaI_BC8_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCCAAG
P1_XbaI_BC9_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAATCGG
P1_XbaI_BC10_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCTGCC
P1_XbaI_BC11_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGATGG
P1_XbaI_BC12_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGTTA
P1_XbaI_BC13_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTACGA
P1_XbaI_BC14_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATAGCTGC
P1_XbaI_BC15_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGACTTGATTG
P2_XhoI_F	/5Phos/T*CGAAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
P2_XhoI_R	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT

Table 2.S2: Sequences used. Sequences for primers and other oligos used in this study.

Supplemental Figures

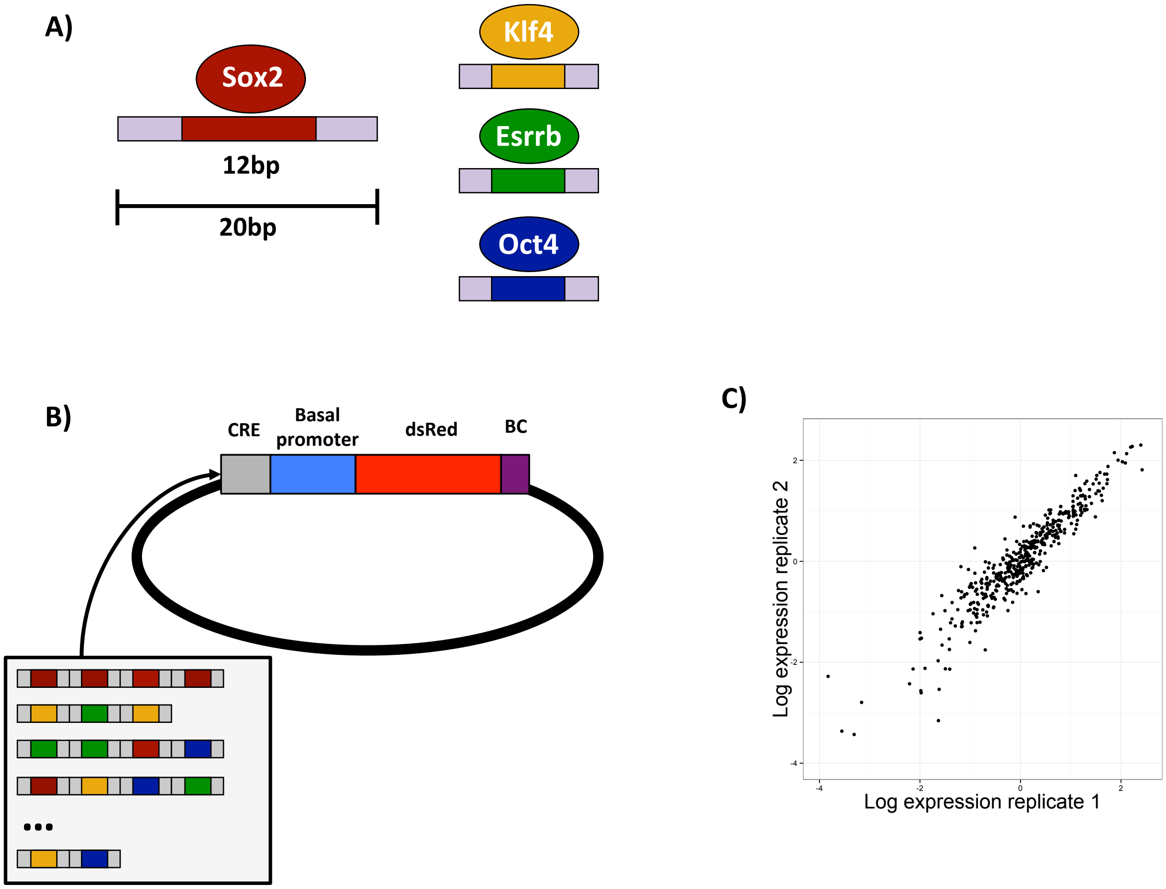


Figure 2.S1: OSKE CRE-seq library. A) Layout of the 20bp sequence building blocks containing the TF binding sites. B) Schematic of the members of the CRE-seq plasmid library. C) Representative dot plot of the reproducibility of the CRE-seq expression between two biological replicates.

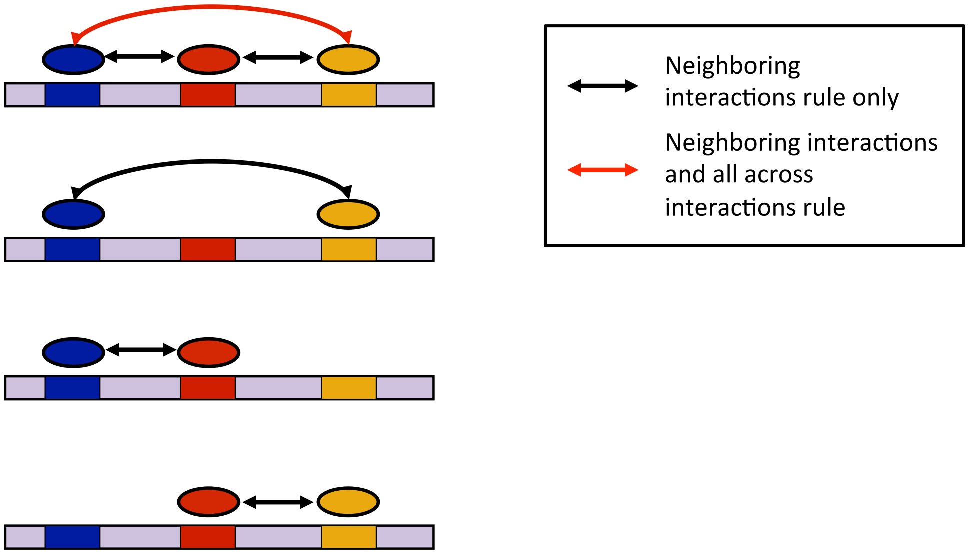
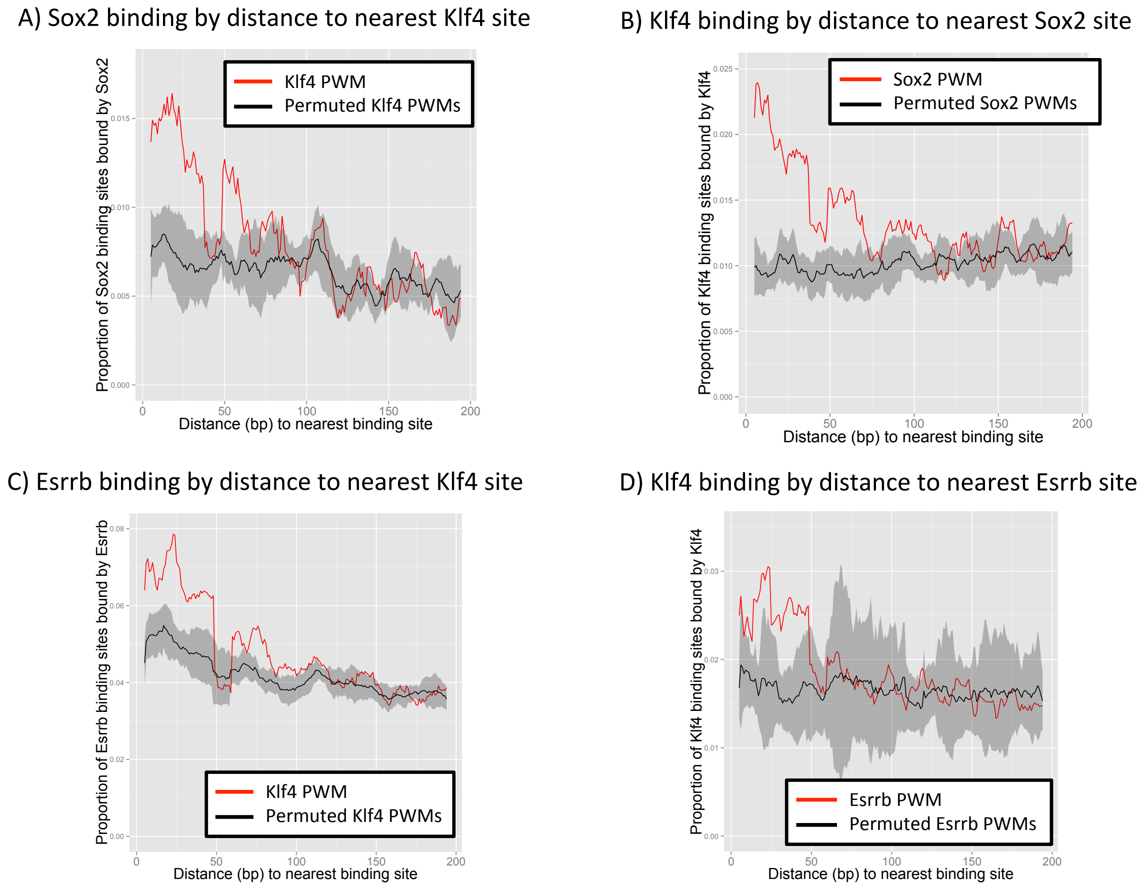

Neighboring interactions rule only
Neighboring interactions and all across interactions rule

Figure 2.S2: Interaction rules in statistical thermodynamic model of transcription. The weights of each possible binding state of a CRE are based on the possible interactions between TFs. Interaction rules dictate which TFs are allowed to interact, as shown here for a CRE with three TFBS and a subset of possible binding states. Each oval is one TF.


A) Sox2 binding by distance to nearest Klf4 site

B) Klf4 binding by distance to nearest Sox2 site

C) Esrrb binding by distance to nearest Klf4 site

D) Klf4 binding by distance to nearest Esrrb site

Figure 2.S3: Enriched TF binding near interacting partners. Binding of TFs determined by genomic ChIP-seq data (Chen et al 2008) are likely to be close to predicted binding sites based on a PWM for interacting partners. The red lines represent binding based on distance to nearest PWM for each TF. The black lines represent binding based on distance to nearest permuted PWMs for reach TF, with the line representing the mean over 10 shuffled PWMS and the ribbon representing the standard deviation. Shown for A) Sox2 binding by Klf4 sites, B) Klf4 binding by Sox2 sites, C) Esrrb binding by Klf4 sites, and D) Klf4 binding by Esrrb sites.

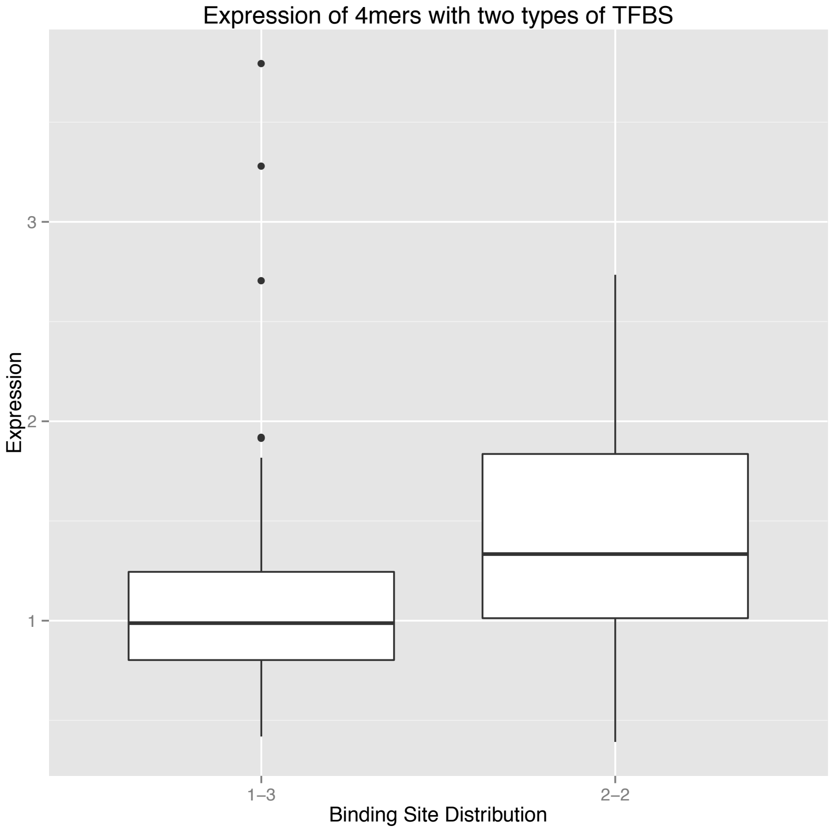


Figure 2.S4: Expression of CREs with two types of binding sites. The expression of CREs with four total binding sites and two types of binding site varies based on whether the binding sites are split 2-2 or 3-1.

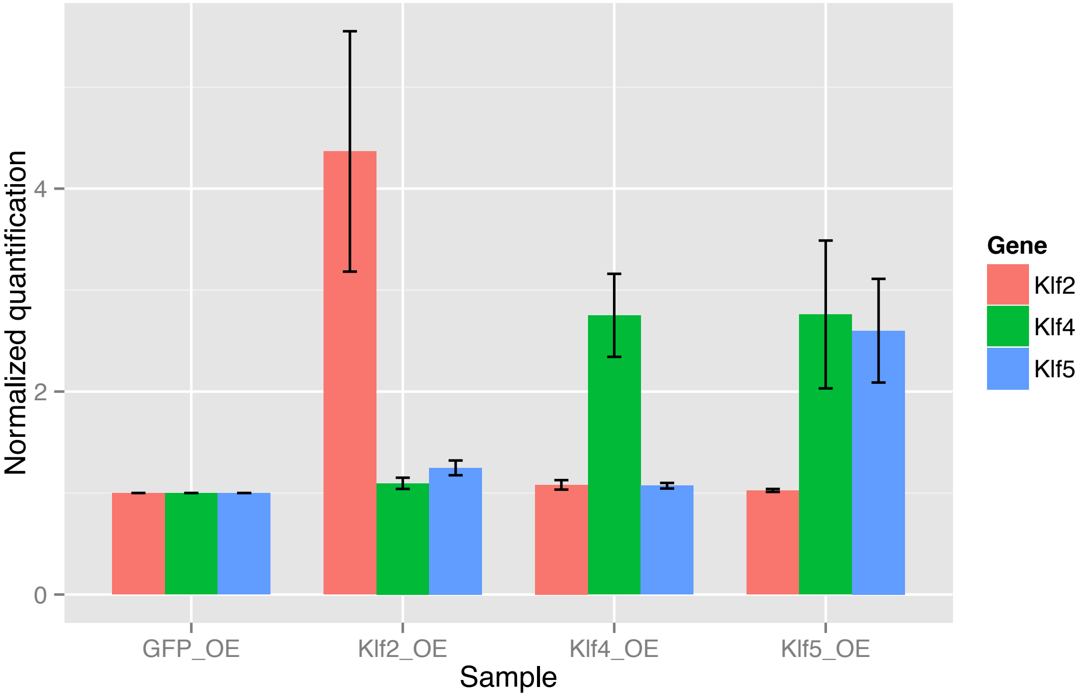


Figure 2.S5: Expression of Klf genes in overexpression conditions. The expression level of Klf2, Klf4, and Klf5 was measured by qRT-PCR in four overexpression conditions.

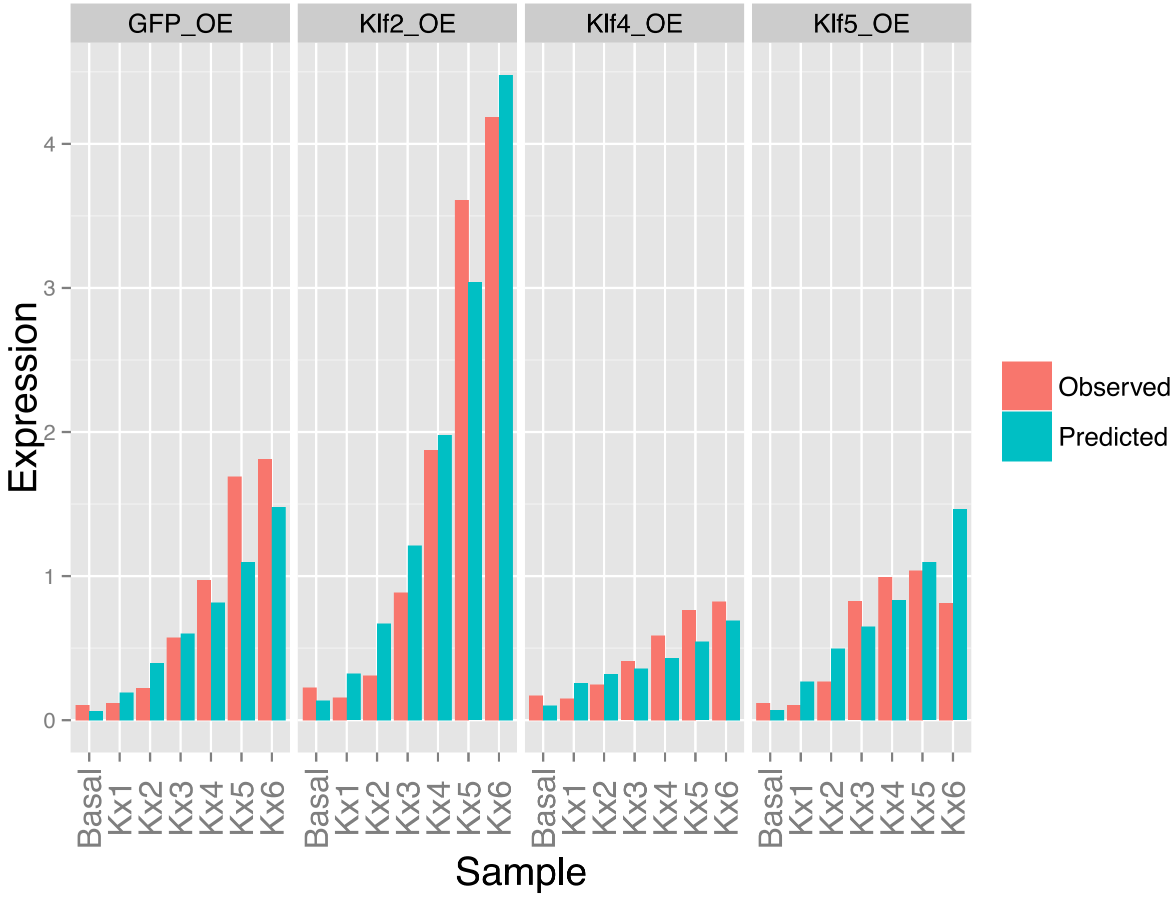


Figure 2.S6: Predicted expression of KBS library. The observed CRE-seq expression of each CRE in the KBS library in each overexpression condition, along with the expression predicted by the thermodynamic model.

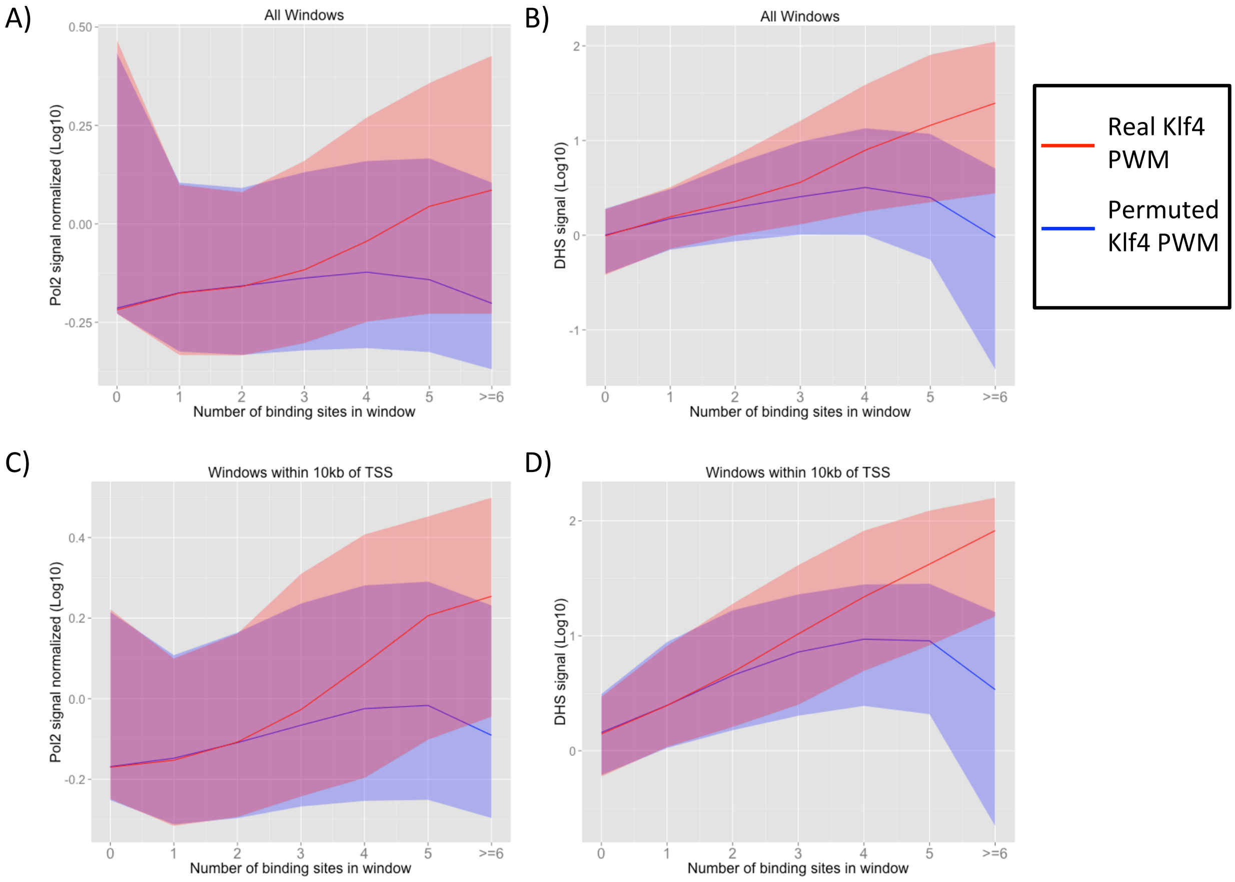


Figure 2.S7: Clusters of Klf4 binding sites function in the genome. The number of Klf4 binding motifs was measured in 200bp genomic windows. The mean RNAP binding signal (A) and DHS signal (B) by number of Klf4 sites in all genomic windows, and in only those genomic windows within 10kb of a TSS (C) and (D). The red line indicates the number of sites found using the Klf4 PWM, and the blue line indicates the number of sites found using 10 permuted Klf4 PWMs. For each color, the middle line is the median and the ribbon represents the upper and lower quartile (the top and bottom of the box in a boxplot).

Chapter 3: Changes in *cis*-regulatory rules during differentiation

Abstract

Transcriptional regulation plays an important role in development. In Chapter 2 we showed that interactions between TFs help dictate cis-regulation in embryonic stem (ES) cells. In this chapter we investigated how these rules of interaction change as ES cells are differentiated down the neural lineage using retinoic acid (RA). We measured the expression of a library of synthetic cis-regulatory elements (CREs) comprised of transcription factor (TF) binding sites for Oct4, Sox2, Klf4, and Esrrb in ES cells treated with RA. Comparing this expression to the expression driven by these CREs in ES cells, we found that the CRE library as a whole drives very different expression in the two cell types. However, the subset of CREs without any repeat TF binding sites drove similar expression in ES cells and RA-treated cells. Furthermore, a statistical thermodynamic model of transcription trained on expression data from ES cells predicts the expression in RA-treated cells only for CREs without repeat binding sites. This suggests that heterotypic interactions between TFs are similar in both cell types, but the effect of binding sites for individual TFs and homotypic interactions may vary.

Precise regulation of transcription is key to the proper development and maintenance of cell fate. Each cell type has a set of transcription factors (TFs) that play a key role in maintaining cell fate and regulating the genes in that cell. Embryonic stem (ES) cells are regulated by a core set of TFs that work to maintain the pluripotent state through transcriptional regulation. These TFs often work in concert by binding similar locations in the genome and in some cases working cooperatively to regulate transcription (Chen et al. 2008c; Kim et al. 2008; Boyer et al. 2005; Loh et al. 2006). For instance, Oct4 and Sox2 physically interact to regulate transcription in ES cells (Ambrosetti et al. 2000; Chew et al. 2005; Kuroda et al. 2005; Ng et al. 2012; Rodda et al. 2005). Through a synthetic *cis*-regulatory element (CRE) approach, we found cooperative interactions between Klf4 and Sox2, Klf4 and Esrrb, and Oct4 and Sox2 in ES cells (Chapter 2). We also found a negative homotypic interaction. Thus, interactions between TFs play an important role in the expression driven by binding sites for pluripotency TFs in ES cells.

While we have shown the *cis*-regulation of binding sites for pluripotency TFs in ES cells, it is unknown how this regulation changes as ES cells differentiate. In addition to their roles in maintaining pluripotency, some of the pluripotency TFs also regulate differentiation into other cell types (Thomson et al. 2011; Teo et al. 2011). In particular, Oct4 can promote the mesendoderm state while repressing the neural ectoderm state; and Sox2 can promote the neural ectoderm state while repressing the mesendoderm state (Thomson et al. 2011). Thus, these two TFs, which work together in ES cells to promote pluripotency, have opposing roles in other lineages. To complicate things further, most of these TFs, including Oct4, Sox2, and Klf4 are part of TF families in which other members can bind very similar binding sites (Ferraris et al. 2011; Jiang et al. 2008; Ng et al. 2012). Therefore, it is unclear how *cis*-regulation will change as ES cells differentiate. The interactions and rules could be constant, or they could change completely. The real answer is likely a mix of both options, and our goal was to determine which interactions change and which stay mostly the same during neural differentiation.

We measured the expression driven by a set of synthetic CREs in ES cells differentiated down the neural pathway to determine if the principles of *cis*-regulation in ES cells are maintained. We used a library of synthetic CREs comprised of binding sites for the pluripotency

TFs Oct4, Sox2, Esrrb, and Klf4 that was previously built in Chapter 2. For this new experiment, we first transfected the CRE library into ES cells and then differentiated the ES cells to measure the expression of the synthetic CRE library in this new condition. We treated the transfected ES cells with media without leukemia inhibitory factor (LIF), a pluripotency promoting factor, and with retinoic acid (RA). RA is required *in vivo* for normal neurogenesis, and RA treatment of ES cells can lead to differentiation into a number of cell types, including neurons and glial cells (Jacobs et al. 2006; Soprano et al. 2007). Thus, RA treatment of these transfected cells leads them away from pluripotency and towards the neural pathway. 41 hours after the change to RA media, the expression of the CREs was assayed and reproducible expression was detected (381 CREs with quality expression measurements, range of R^2 between biological replicates: 0.61-0.67, Fig 3.1). This is lower expression quality than we see in ES cells in pluripotency media, perhaps due to the heterogeneity of the differentiated cell population.

We next investigated how the expression of the CREs in RA media compared to the expression of the ES cells in LIF media, originally measured in Chapter 2. As the expression of the pluripotency TFs studied in the library changes during this treatment (Ivanova et al. 2006), we expected the expression driven by binding sites for these TFs to change as well. The expression could change in two ways: 1) by a simple scaling factor in which the relative expression of the CREs in each condition is very similar, or 2) by complex dynamics in which the relative expression of the CREs in the library changes. We found that there is a poor correlation between the expression in LIF and RA when looking at all CREs (n=304 CREs, R^2=0.24, Fig 3.2a). However, for CREs with no repeat binding sites, the expression between the two conditions is much more similar (n=87, R^2=0.62, Fig 3.2b). This suggests that the similarity of expression between the two conditions is dependent on the complexity of the CRE (i.e. how much of a variety of binding sites there is). In CREs with four total binding sites, the expression in RA media trends with the number of unique types of binding sites, as we saw in LIF media (Fig 3). However, the effect is not as strong in RA as LIF, suggesting that the homotypic chains of binding sites, which drive this effect, are not as repressive in RA. Thus, there are likely some changes in the *cis*-regulatory code that drive the differential expression between the two conditions.

To further investigate how *cis*-regulation changed between the LIF and RA conditions, we used a statistical thermodynamic model to learn how the *cis*-regulatory interactions varied between the two cell types. We tested how well the model that fit the expression data in LIF, the LIF Model, explained the RA expression. Over the whole CRE library (n=323), the thermodynamic model explains the RA expression data poorly (R^2=0.178 vs. R^2=0.73 for LIF) (Fig 4A). However, if the measured relative concentration of the pluripotency TFs in RA media is input into the model, the model explains the expression of the CRE library in RA somewhat better (R^2=0.25). This implies that the changing TF concentrations do have an effect on the expression, as one would expect. Furthermore, when only those CREs without repeats of individual binding sites are included (n=90), the model does a much better job at explaining the expression (R^2=0.49) (Fig 4B). This confirms the previous assertion that CREs with repeats of binding sites behave differently in the two treatments. Modeling these CREs does not use the homotypic parameter, as there are no repeats binding sites in them. It is thus likely that the homotypic parameters are behaving differently in the two conditions. The heterotypic interaction parameters, which dominate the predictions from CREs with no repeat binding sites, may be more constant between conditions.

Here we show a comparison of the expression driven by synthetic CREs in two cell types: ES cells in LIF (pluripotency) media, and ES cells differentiated down the neural pathway using RA. Overall, the CREs show different expression between the two cell types, but certain principles seem to be consistent. Interestingly, in RA-treated cells, the expression of CREs with combinations of unique binding sites (i.e. no repeats of binding sites) are well explained by the thermodynamic model found using the LIF-treated cells, whereas the library of CREs as a whole is not explained well. It is known that many of the TFs in ES cells are down regulated upon differentiation and RA treatment (Thomson et al. 2011; Wu et al. 2014; Ivanova et al. 2006). However, it's possible that other family members of these TFs may be up regulated and could bind the same binding sites. This suggests two models for how TFs are acting during RA-induced differentiation: 1) different TF family members bind to these sites and recruit polymerase with different effect but interact with TFs at other sites in a similar manner, or 2) other TF family members are not relevant and the only changes are in the activity level of the pluripotency TFs, which alters the effect of the homotypic interactions. The fact that the simple adjustment of TF concentrations in the thermodynamic model improves its ability to explain expression in RA suggests that the second option is at least partly true. Either way, it is apparent that the manner in which binding sites drive expression as cells differentiate is affected by changing *cis*-regulatory rules.

Methods

Cell culture

ES cell line RW4 was cultured as described previously (Chapter 2). To differentiate the cells and perform transfections of the CRE-seq library, ES cells in a 6-well plate were first transfected with 3ug of the OSKE plasmid library and 0.3ug of GFP reporter plasmid CF128. 6 hours post-transfection, the ES cells were passaged into media without LIF but with retinoic acid, with half of the transfected cells being passaged. RNA was harvested from the cells 48 hours post-transfection, using the PureLink RNA mini kit (Life Technologies). RNA was prepared for Illumina sequencing using the CRE-seq protocol as described previously (Chapter 2).

Statistical analysis

The CRE-seq data from RA was processed as described previously (Chapter 2), except using a DNA read filter of 50. The data was further filtered for those CREs with a standard error of the mean (SEM) less than 0.4 of the mean expression of the CRE. The CRE-seq expression from the ES cell conditions are from Chapter 2, using a DNA read filter of 2000 reads. All statistical analysis and thermodynamic modeling was performed as described previously (Chapter 2). All expression comparisons were done using the log expression values.

Figures

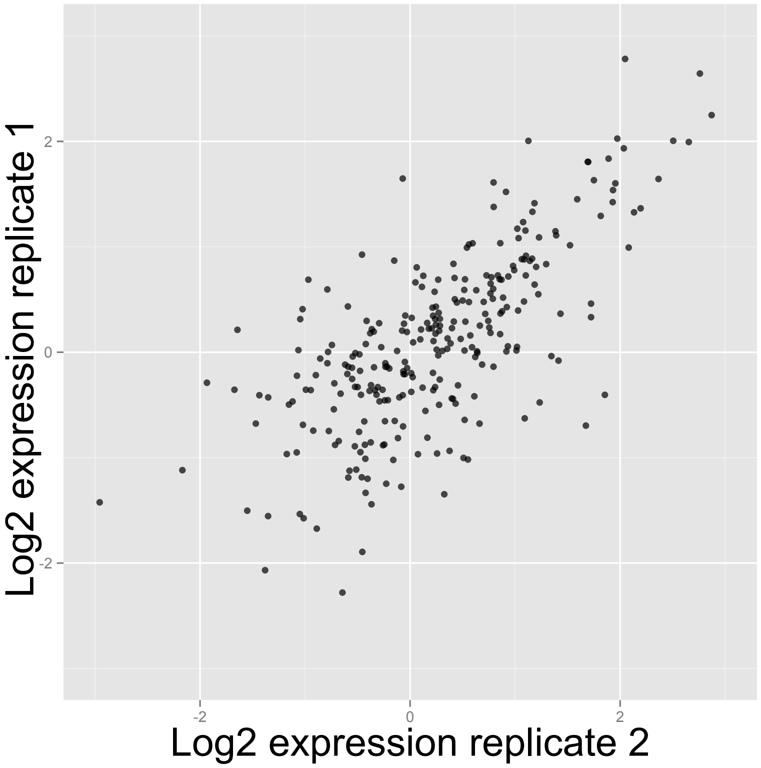


Figure 3.1: Reproducibility of expression measurements in RA. The expression of CREs by CRE-seq from two representative replicates in cells treated with RA.

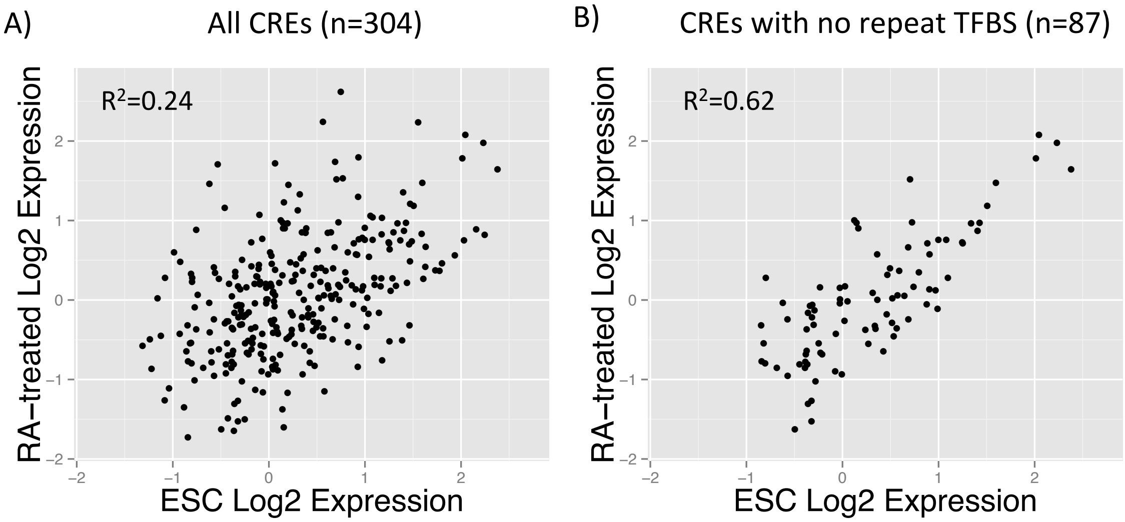


Figure 3.2: Expression of CREs in LIF vs. RA media. The expression of CREs is compared between ES cells in LIF media and in ES cells treated with RA. A) Expression of all CREs. B) Expression of CREs with no repeat TFBS.

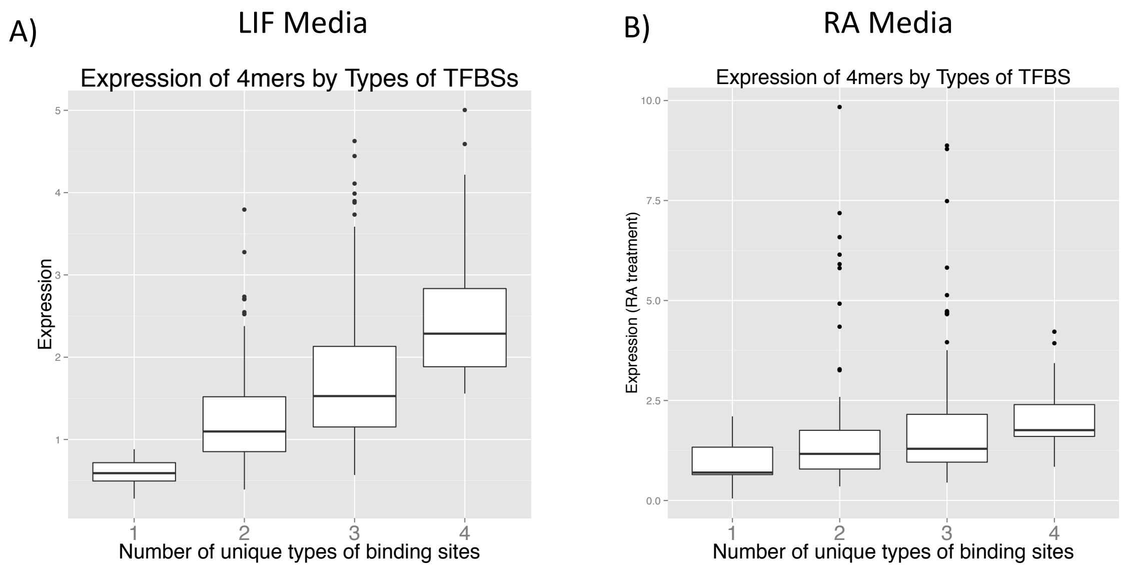


Figure 3.3: Expression of CREs by unique sites. The expression of CREs with exactly four total TF binding sites by number of unique types of binding sites, in A) ES cells in LIF media, B) ES cells treated with RA.

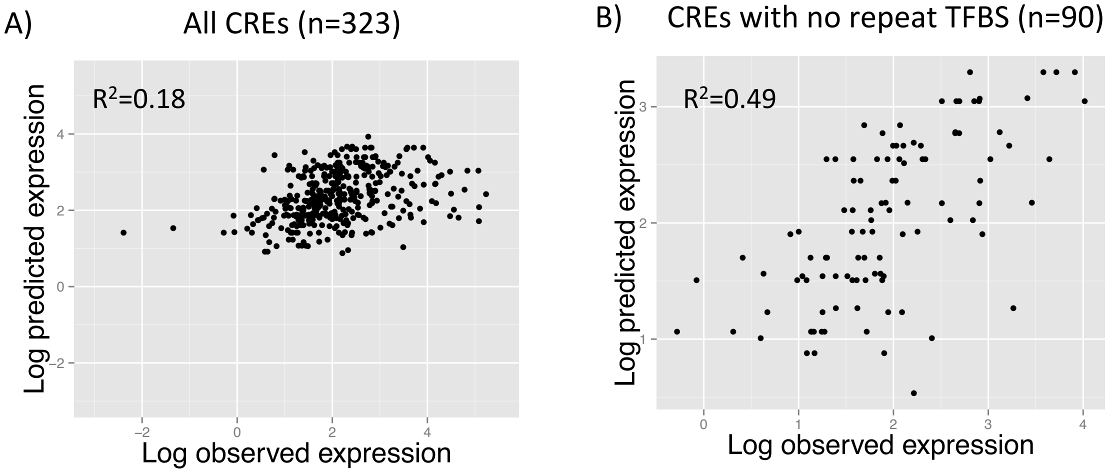


Figure 3.4: Thermodynamic model predictions of RA expression. The thermodynamic model fit on CRE expression in ES cells was used to predict the expression of CREs in RA. A) All CREs. B) CREs with no repeat TFBS.

Chapter 4: High-throughput functional testing of ENCODE segmentation predictions

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 these data to segment the genome into regions with different predicted regulatory activities (Dunham et al. 2012; Hoffman et al. 2012b). 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 line (H1-hESC). 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 show that a significant fraction (~26%) of the ENCODE enhancer predictions have regulatory activity suggesting that histone modification states can reflect the *cis*-regulatory activity of sequences in the genome, but that specific sequence preferences, such as transcription factor binding sites, are the causal determinants of *cis*-regulatory activity.

This chapter was written as a paper, *High-throughput functional testing of ENCODE segmentation predictions*, with Jamie Kwasnieski, Hemangi Chaudhari, and Barak Cohen that was published in *Genome Research* (2014, Volume 24: 1595-1602). Jamie and I were joint first authors and both contributed to the writing of the paper along with Barak Cohen. Jamie, Hemangi, and I performed the experiments and analyzed the data. It is available under a Creative Commons License (Attribution-NonCommercial 4.0 International).

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. 2012a) and ChromHMM (Ernst and Kellis 2010, 2012), take functional genomics data as input (DNase-seq; FAIRE-seq; and ChIP-seq of histone modifications, RNA Polymerase II large subunit (POLR2A), 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. 2012b). 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 the accepted operational definition of enhancer activity as the ability to modulate expression of a reporter gene under control of a basal promoter. 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 levels 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. 2012b; Sharon 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 (Supplemental Data 4.1).

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) (Supplemental Data 4.2, Supplemental Data 4.3). We performed four independent transfections in K562 cells and found that our expression measurements are precise, displaying high reproducibility between biological replicates (Figure 4.1A, R^2 range: 0.95-0.97). To test the robustness of our measurements, we used a luciferase assay to measure expression driven by twelve individual CREs upstream of the minP basal promoter. Expression in the luciferase assay exhibits strong agreement with the batch CRE-seq expression measurements upstream of the hsp68 promoter (R^2=0.70, Figure 4.1B, Figure 4.S1), demonstrating that our assay accurately measures *cis*-regulatory activity and that our results have little dependence on the choice of minimal promoter.

Expression of Segmentation Classes

We compared the activity of each class of segmentation prediction to the activity of its corresponding negative control distribution of scrambled sequences. We used two metrics to classify individual segmentations as "active" or "inactive" with respect to this negative control expression distribution (Table 4.1). 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. We recognized that CREs may be active even if they drive expression below the 95th percentile of the control, so we also used a second metric to capture some of these sequences. We compared the sixteen replicate measurements for each CRE (four barcodes per CRE in four independent experiments) with the distribution of all of the scrambled controls (Wilcoxon Rank Sum Test, one-tailed, $P<0.05$, Bonferroni correction with $N=3236$). We conducted the same test for each scrambled CRE to estimate the fraction of scrambled sequences that drive activity (Table 4.1, square brackets). By both of these metrics, a significant number of Enhancer and Weak Enhancer CRE predictions are active (Figure 4.1C, 4.1D, Table

4.1). 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 4.1E, 1F, Table 4.1). 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 (Table 4.S1). 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 (Figure 4.S2). This result indicates that our expression measurements 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.

Our previous work (White et al. 2013) showed that CRE-seq can detect repression below basal promoter activity, particularly when the minimal promoter has detectable expression on its own. In this experiment we chose the Hsp68 promoter as it drives expression in the 48th percentile of the library of genomic sequences. Many sequences, both segmentation predictions and scrambled sequences, drove expression that was significantly lower than the scrambled distribution, indicating that we can detect repression in this assay. However, we observed no significant increase in the number of sequences with repressive activity in the segmentations as compared to the scrambled sequences suggesting that the segmentations do not repress expression below what is expected by chance (Table 4.S1, Wilcoxon Rank Sum Test, $P<0.05$, Bonferroni correction). We conclude that Enhancer, Weak Enhancer, and Repressed segmentations do not have the ability to repress the Hsp68 promoter.

Unexpectedly, we found that sequences classified as Weak Enhancers drive a higher median level of activity than sequences classified as Enhancers (Figure 4.S3, 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 4.1). Compared to Weak Enhancers, segmentations in the Enhancer class have higher GC content (Figure 4.4B), a sequence feature associated with higher *cis*-regulatory activity (Landolin et al. 2010; 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 (Figure 4.S4A). 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 correlate with 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 4.2B), thought to signify active enhancers, and H3K36me3 (Barski et al. 2007) (Figure 4.2D), 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 4.2A, 4.2C), even within the Enhancer or Weak Enhancer classes (Figure 4.S5). We did not find an association of H3K27ac signal in the larger context (up to 500bp surrounding the selected regions). In one study “dips” in the levels of H3K27ac correlated with enhancer activity (Kheradpour et al. 2013), which is consistent with our observation that lower levels of H3K27ac are more predictive of enhancer activity. However, in our data we did not see correlation between the H3K27ac “dip score” and *cis*-regulatory activity (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. 2012b) (DNase HS from Duke; DNase HS from University of Washington [UW]; Faire-seq; and ChIP-seq of H3K4me1, H3K36me3, and RNA POLR2A) are significantly enriched in sequences that drive high expression in our assay (Table 4.S2, Wilcoxon Rank Sum test, $P<0.05$ Bonferroni correction with $N=16$). 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 4.3A, 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 and all other single features had AUC lower than 0.6 (Table 4.S2). A logistic regression model with the above-mentioned six chromatin features and two primary sequence features (PSF), improves the classification of active sequences (Figure 4.3A, AUC=0.733), but only marginally above that of DNase HS alone. However, even amongst those CREs with a high DNase HS score (UW DNase HS score>5, 685/2096 CREs pass this threshold), the active CREs are enriched for seven chromatin features, suggesting that there is some additional information in the histone modifications beyond DNase HS despite the fact that DNase HS is by far the most predictive feature (Table 4.S3). As chromatin and primary sequence features can only classify active sequences to a moderate level, we 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. 2014), we searched for motifs enriched or depleted in activated CREs and found 50 significant, non-redundant motifs (Table S4). A logistic regression model that incorporated these binding models performs better at distinguishing active sequences than the chromatin and PSF model (Figure 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 (Table 4.S5). 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. In addition, the most significant motif found in a discriminative *de novo* motif analysis (Bailey 2011) was highly similar to the AP1 motif (E=0.0041) (Gupta et al. 2007). Amongst segmentations with a predicted AP1 motif, DNase HS (Duke) is the only chromatin feature significantly enriched in those that are active

(Table 4.S3), suggesting that DNase HS provides some additional information beyond the presence of the AP1 motif. The expression driven by CREs with predicted AP1 motifs is significantly higher than the expression driven by sequences without the motif (Figure 4.3C, $\log_2$ 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 (Dunham et al. 2012) (Figure 4.3D; $p=8.8x10^{-10}$ 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. 2011b; Arvey et al. 2012; Gorkin et al. 2012; Smith et al. 2013b). 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 the lack of observed repression. 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.

Only a small fraction (~26%) of predicted enhancer sequences had activity in this assay. It is therefore possible that a large fraction of the predictions in ChromHMM/Segway are false positives. Alternatively, many sequences might score as false negatives in this assay. The short length and episomal nature of the expression assay could contribute to false negatives, although we emphasize that the accepted operational definition of an enhancer is a sequence that modulates the activity of an episomal reporter gene. In addition, our comparison of segmentations to scrambled controls does not allow us to find active sequences that express at low levels. Finally, it is possible that some sequences might only be active in the context of the genome or when paired with a different minimal promoter sequence. While the relative number of active sequences between classes in our assay should be accurate, as the same experimental design was utilized for all sequences, our estimates should be taken as a lower bound of the number of active sequences.

Finally, we conclude that combinations of histone modifications often 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.

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 Figure 4.S6.

The plasmid library was prepared as described (Kwasnieski et al. 2012), except using primers CF166 and CF167 (Table 4.S6) 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 (Table 4.S6). 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 (Table 4.S6). 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. 2012b) were downloaded from the Ensembl genome browser (Flicek et al. 2013) 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

RepeatMasker 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. Genomic and scrambled CRE sequences are listed in Supplemental Data 1. All genomic coordinates used are from hg19.

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 (Table 4.S6). 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 4.6) 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 (Supplemental Data 4.2). The expression of each CRE in each biological replicate was calculated as the mean of the expression of each BC associated with it, and the overall expression of each CRE was calculated as the mean of its expression in each biological replicate. The standard error of the mean (SEM) was calculated as described previously (Kwasnieski et al. 2012) (Supplemental Data 4.3).

Luciferase assays

Plasmid pGL-CBR was created by inserting the click-beetle red (CBR) luciferase gene (from pCBR-Control Vector [Accession Number AY258592], Promega) into pGL4.23 (Promega) at the XbaI and NcoI sites. pGL-CBR contains the minP basal promoter from pGL4.23. 12 individual CREs from the oligo library were amplified by PCR and inserted into pGL-CBR at the NheI and HindIII sites to form individual pGL-CBR-CRE plasmids. The 46-bp *cis*-regulatory element containing the HS II enhancer from Ney et. al.(1990) was also cloned into pGL-CBR using annealed oligos POS1 and POS2 (Table 4.S6), also at the NheI and HindIII sites of pGL-CBR, to create a positive control pGL-CBR-CRE plasmid. Each pGL-CBR-CRE plasmid, along with the original pGL-CBR, was then transfected into K562 cells individually in triplicate using the Neon transfection system. Each transfection used 4ug pGL-CBR-CRE plasmid with 0.4ug Renilla control plasmid (pRL-CMV, Promega) and 2 x 10^5 cells. Transfected cells were then seeded into 12-well plates with 1ml of growth media. 26 hours later, each well was split into two wells, each in a separate 24-well plate (Krystal 24 Well Black Assay Plate, MidSci). These were then immediately imaged using I IVIS 50 (Caliper, Hopkinton, MA;

exposure time 10-60 seconds, binning 8, field of view 12, f/stop 1, open filter), with one plate imaged for CBR-luciferase using 150 µg/mL D-luciferin (Gold Biotech), and one plate imaged for Renilla using 1 µg/ml Coelenterazine (Biotium Inc.). The CBR-luciferase signal of each transfection sample was normalized by the corresponding Renilla signal, and the expression of each CRE was determined by the mean of the three transfections (Supplemental Data 4.4).

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. 2012b), 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, RNA 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. 2014), 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 AP1 binding matrix that was enriched in highly expressed sequences in our assay was from JASPAR (MA0099.2). DREME (Bailey 2011) was used for discriminative motif finding, using the sequences activated over the 90th percentile of the scrambled distribution as the positive group and all other sequences as the negative group, with the maximum motif length set at 12bp and all other default options. The TOMTOM web module (http://meme.nbcr.net/meme/cgi-bin/tomtom.cgi) (Gupta et al. 2007) was used to find similar motifs, using default options. 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 90th 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 (two-tailed, $P<0.05$, 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 (Table 4.S5).

Tables

Segmentation Prediction	**Active over 95% scrambled**	**Active by Wilcoxon**
K562 Enhancer	11.3% [5.30%]	26.0% [12.68%]
K562 Weak Enhancer	25.7% [5.32%]	39.17% [15.1%]
K562 Repressed	5.35% [4.98%]	7.00% [7.39%]
H1-hESC Enhancer	4.34% [5.30%]	11.33% [14.1%]

Table 4.1. 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 95th 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 $N=3236$).

Figures

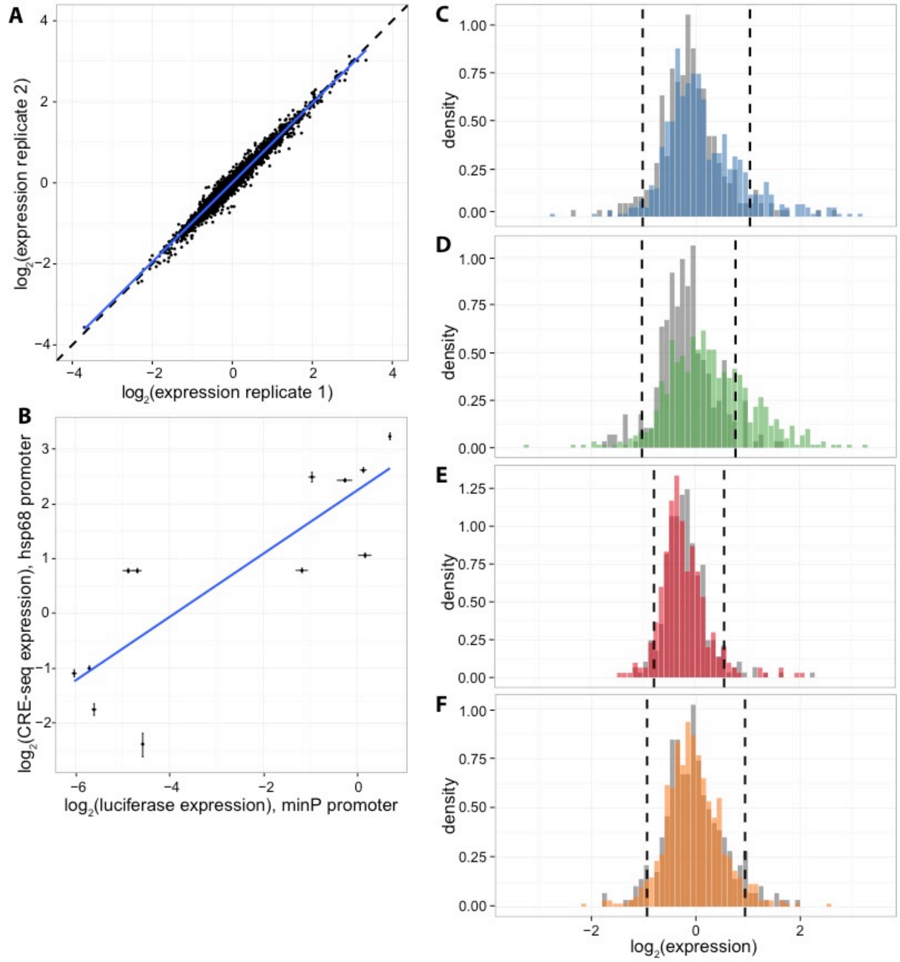


Figure 4.1. Reproducible expression measurements show differences in expression by segmentation class. A) Representative scatterplot showing expression of each CRE in two biological replicates (R^2=0.95, range of R^2 between all replicates: 0.95-0.97). Dashed black line is line of equality and blue line is best fit. **B)** Correlation between CRE-seq and luciferase assays. Twelve CREs measured in individual luciferase assay (upstream of minP promoter, x-axis) and batch CRE-seq assay (upstream of hsp68 promoter, y-axis). Error bars represent the standard error of the mean. Blue line is best fit. R^2=0.70 **C-F)** 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. **C)** K562 Enhancer class (blue), **D)** K562 Weak Enhancer class (green), **E)** K562 Repressed class (red), **F)** H1-hESC Enhancer class (orange).

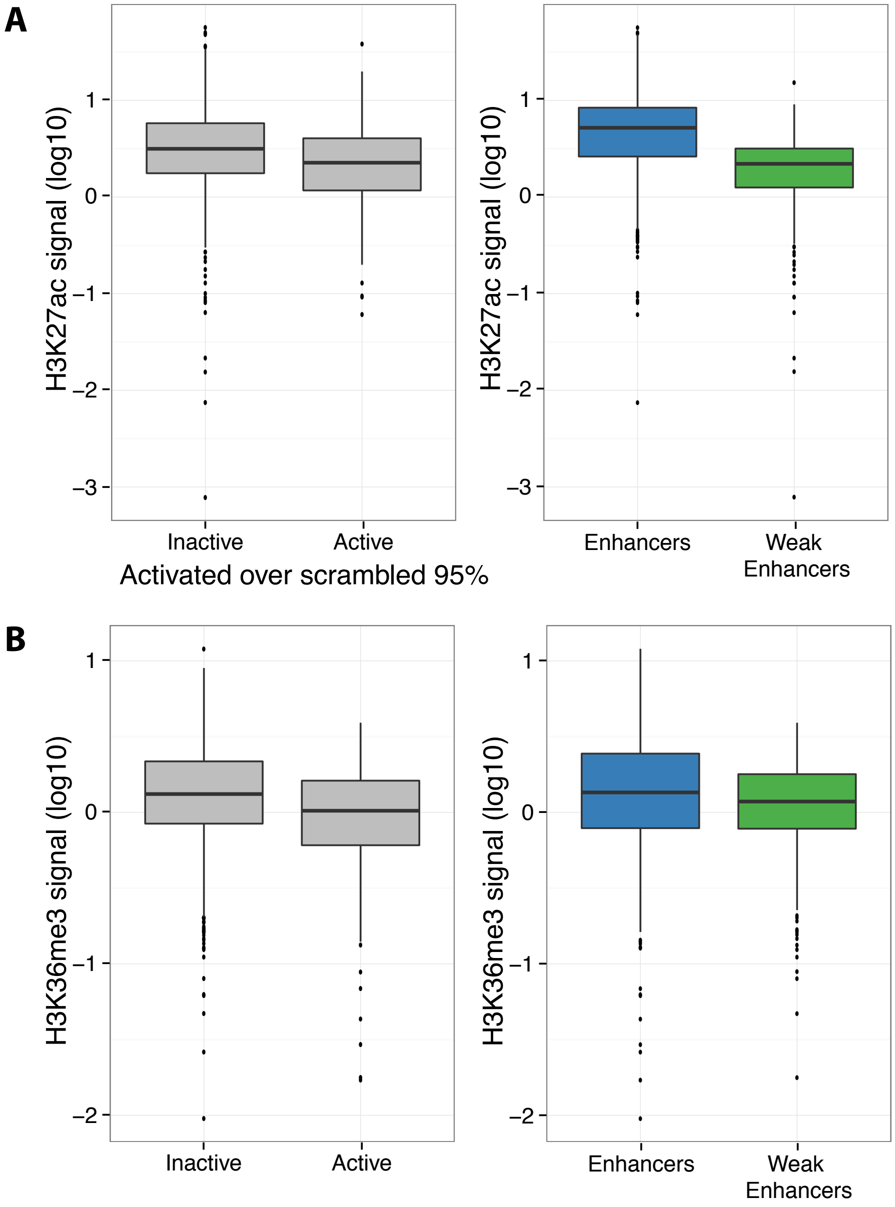


Figure 4.2. Lower H3K27ac and H3K36me3 signal are associated with higher Weak Enhancer expression. Boxplots showing that **(A)** H3K27ac signal and **(C)** H3K36me3 signal are depleted in active CREs compared to inactive CREs. **B)** H3K27ac signal and **(D)** H3K36me3 signal are also depleted in Weak Enhancers (green) compared to Enhancers (blue). Active CREs are those above 95th percentile of scrambled distribution (Table 4.1).

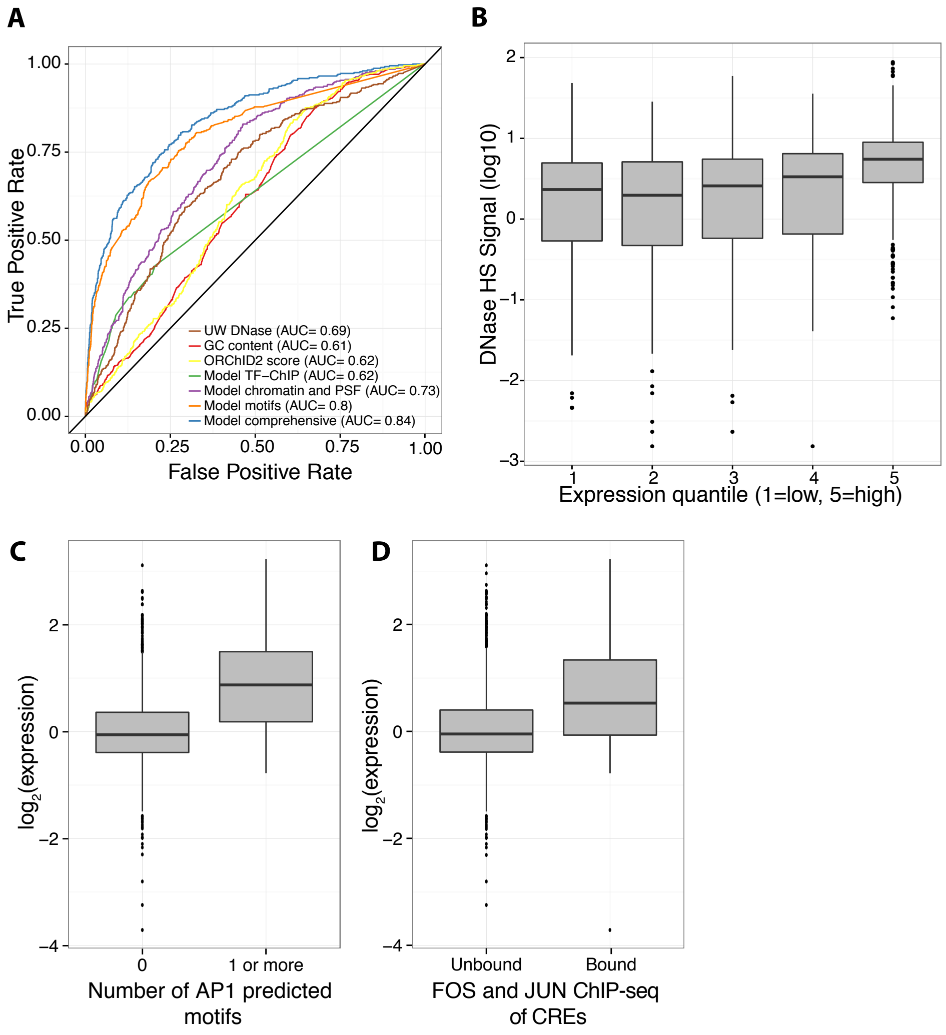


Figure 4.3. Chromatin features and sequence-specific binding identify active sequences.

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.

Supplementary Figures and Tables

Segmentation Prediction	Repressed under 5% scrambled	Repressed by Wilcoxon
K562 Enhancer	6.02% [5.30%]	6.83% [9.86%]
K562 Weak Enhancer	4.00% [5.32%]	7.33% [9.51%]
K562 Repressed	3.52% [4.98%]	7.67% [6.34%]
H1-hESC Enhancer	2.50% [5.30%]	10% [10.21%]

Table 4.S1. Percentage of repressed CREs by segmentation class. For each ENCODE segmentation class, the table shows the percentage of all genomic CREs that are repressed, with the percentage of matched scrambled controls that are repressed in square brackets. Repression was determined by comparing CRE expression to the 5th percentile of matched scrambled controls (Repressed under 5% scrambled) or by statistically comparing replicate measurements of expression to matched scrambled control distribution (Repressed by Wilcoxon, Wilcoxon Rank Sum test, one-tailed, $P<0.05$, corrected using Bonferroni method with $N=3236$).

Chromatin Modification or Dataset	**AUC**	**Wilcoxon p-value**
Control	0.523	0.135
CTCF	0.504	0.807
Duke DNase	0.680	$7.93\text{x}10^{-30}$
FAIRE	0.591	$9.84\text{x}10^{-9}$
H3K27ac	0.523	0.150
H3K27me3	0.523	0.142
H3K36me3	0.553	$8.34\text{x}10^{-4}$
H3K4me1	0.551	$1.32\text{x}10^{-3}$
H3K4me2	0.536	0.0239
H3K4me3	0.507	0.670
H3K9ac	0.517	0.274
H3K20me1	0.531	0.0436
Pol2	0.596	$1.28\text{x}10^{-9}$
UW DNase	0.685	$2.27\text{x}10^{-31}$

Table 4.S2: Predictive capability of single chromatin features. The AUC was calculated for ROC curves created to classify active CREs vs. inactive CREs (see methods for details) using individual chromatin features. A Wilcoxon rank sum test (two-tailed, P<0.05, Bonferroni correction with N=14 for p-value threshold of 0.00357) was used to test for differences in the chromatin feature scores between activate CREs and inactive CREs.

Chromatin Modification or Dataset	p-value in high DNase segments	p-value in segments with predicted AP1 motif
Control	2.70×10^{-5}	0.819
CTCF	9.49×10^{-3}	0.400
Duke DNase	1.24×10^{-6}	0.00258
FAIRE	0.0392	0.669
H3K27ac	1.18×10^{-4}	0.827
H3K27me3	0.131	0.0606
H3K36me3	8.74×10^{-5}	0.0163
H3K4me1	0.0124	0.0890
H3K4me2	5.58×10^{-3}	0.189
H3K4me3	2.83×10^{-5}	0.879
H3K9ac	1.86×10^{-7}	0.598
H3K20me1	1.48×10^{-4}	0.160
Pol2	0.236	0.107
UW DNase	0.134	0.0241

Table 4.S3: Enrichment of single chromatin features in subsets of CREs. Differences in the chromatin feature scores were calculated between active CREs and inactive CREs within subsets of the CREs (Wilcoxon Rank Sum Test, two-tailed, $P<0.05$, Bonferroni correction with $N=14$ for p-value threshold of 0.00357). CREs with high DNase HS scores (UW DNase HS score>5) or with a predicted AP1 motif were analyzed.

Motif	**Database**	**p-value**
AP1 (MA0099.2)	Jaspar	7.68E-39
Myc (MA0147.1)	Jaspar	1.36E-07
SPI1 (MA0080.2)	Jaspar	1.14E-07
Mycn (MA0104.2)	Jaspar	1.19E-05
ELF5 (MA0136.1)	Jaspar	3.31E-06
NFE2L2 (MA0150.1)	Jaspar	5.99E-15
Klf4 (MA0039.2)	Jaspar	5.44E-12
GABPA (MA0062.2)	Jaspar	1.25E-15
FEV (MA0156.1)	Jaspar	9.86E-09
Tal1:Gata1 (MA0140.1)	Jaspar	1.12E-13
SP1 (MA0079.2)	Jaspar	7.04E-12
Zfx (MA0146.1)	Jaspar	7.17E-06
ELF3	High-throughput SELEX	3.67E-09
ZNF740_v2	High-throughput SELEX	8.38E-08
SP8	High-throughput SELEX	1.29E-06
SP1	High-throughput SELEX	1.88E-12
SP3	High-throughput SELEX	1.06E-11
SP4	High-throughput SELEX	6.36E-10
EHF^	High-throughput SELEX	2.83E-11
JDP2_v3	High-throughput SELEX	5.88E-15
ELF5_v2	High-throughput SELEX	5.69E-12
ERG	High-throughput SELEX	2.89E-07
ERF	High-throughput SELEX	3.04E-07

Jdp2	High-throughput SELEX	8.40E-21
Elk3	High-throughput SELEX	8.05E-07
ELK3^	High-throughput SELEX	4.31E-07
ELK4^	High-throughput SELEX	1.23E-07
ELK1^	High-throughput SELEX	1.75E-06
ETS1_v3^	High-throughput SELEX	3.51E-08
ERG_v3^	High-throughput SELEX	2.38E-10
ETS1^	High-throughput SELEX	6.28E-09
ELF3_v2	High-throughput SELEX	1.53E-07
Klf12	High-throughput SELEX	1.00E-05
ELK1_v2^	High-throughput SELEX	1.90E-07
EGR3	High-throughput SELEX	2.59E-05
NFE2	High-throughput SELEX	2.50E-25
GABPA	High-throughput SELEX	5.65E-09
KLF16	High-throughput SELEX	3.58E-11
FLI1_v3^	High-throughput SELEX	2.64E-07
FLI1^	High-throughput SELEX	1.03E-06
ETV6_v2	High-throughput SELEX	3.48E-10
Elf5^	High-throughput SELEX	4.16E-08
ETV1^	High-throughput SELEX	2.43E-05
ETV2	High-throughput SELEX	7.74E-10
ETV4^	High-throughput SELEX	3.51E-07
ELF1_v2	High-throughput SELEX	5.34E-07
FEV^	High-throughput SELEX	2.55E-05

JDP2	High-throughput SELEX	1.22E-20
ELF1	High-throughput SELEX	1.60E-11
ELF5	High-throughput SELEX	5.79E-12
ELF4	High-throughput SELEX	6.94E-09
Ehf	UniPROBE	5.30E-06
Sp4	UniPROBE	1.15E-05
Klf7	UniPROBE	1.56E-12
Ehf_v2^	UniPROBE	9.53E-06
Jundm2_v3	UniPROBE	3.75E-30
Jundm2_v4^	UniPROBE	3.75E-30
Gabpa	UniPROBE	5.49E-08
Pho4.primary	UniPROBE	2.12E-06
Gcn4.DBD.primary	UniPROBE	2.17E-27
HLH.27	UniPROBE	1.28E-05
Zfp281	UniPROBE	1.11E-07
Ascl2_v3	UniPROBE	3.85E-09
Ascl2_v4^	UniPROBE	3.85E-09
Sp4_v2^	UniPROBE	2.39E-05
Egr1_v2	UniPROBE	2.60E-08
Sfpi1	UniPROBE	1.82E-05
Egr1^	UniPROBE	2.99E-08
Klf7_v2^	UniPROBE	5.47E-13
MXL.3	UniPROBE	2.86E-07
Gabpa_v2^	UniPROBE	2.82E-09

Zfp281_v2^	UniPROBE	1.09E-07
MXL.3_v2^	UniPROBE	2.86E-07

Table 4.S4: Enriched Motifs in Activated Sequences. These predicted motifs (see methods for sources) were found to have a significantly different distribution of appearances in the active CREs vs. inactive CREs by Wilcoxon test ($P<0.05$, two-tailed, Bonferroni correction with $N=1687$). ^ symbol indicates that the motif was subsequently removed for further analysis because it was highly redundant with another motif.

Model	**Number of Parameters**	**AIC**	**AUC**	**Mean AUC from 5x cross-validation**
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

Table 4.S5: Logistic Regression Models. The four logistic regression models used to predict active CREs from our assay: “Histone and primary sequence features (PSF)” includes histone features, GC-content, and ORChID2 score; “TF ChIP” includes peaks from TF ChIP-seq; “Binding” includes predicted binding models for TFs; and “Comprehensive” includes 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) are listed for each model.

Name	Sequence
Primer CF121	TAGCGTCGAGGACATCAAGA
Primer CF150	TACACCGTGGTGGAGCAGTA
Primer CF151b	AGCGTACTCGAGTTGTTAACTTGTTTATTGCAGCTT
Primer CF168	ATGCATGCCTAGAATTACTACTGGAACA
Primer JKO25F	CATCAAGCTTCTCCTCCGGCTCGCT
Primer JKO25R	CGTTGTAAAACGACGGGATC
Adapter MO576	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTCGATCATG
Adapter MO577	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGACTATCATG
Adapter MO578	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCTACCCTCATG
Adapter MO579	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATAGTGGACACATG
Adapter MO582	ATCGAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
Adapter MO583	ATAGTCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
Adapter MO584	AGGGTAGCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
Adapter MO585	TGTCCACTATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
POS1	CTAGCCTCAAGCACAGCAATGCTGAGTCATGATGAGTCATGCTGAGGCTTAA
POS2	AGCTTTAAGCCTCAGCATGACTCATCATGACTCAGCATTGCTGTGCTTGAGG

Table 4.S6: Oligonucleotide sequences used in this study.

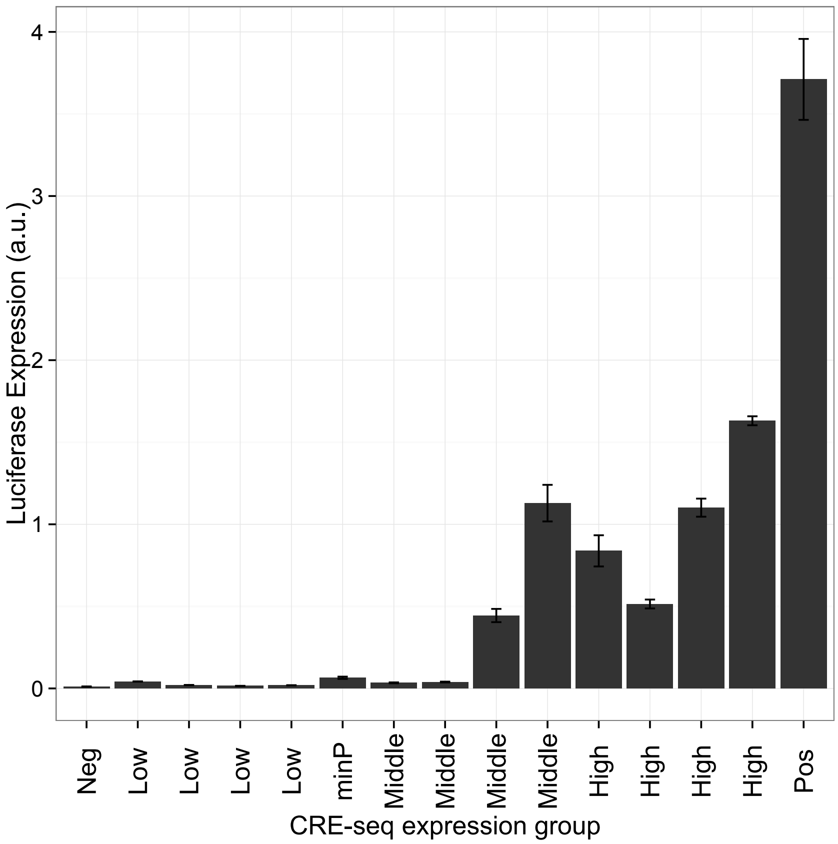


Figure 4.S1: Luciferase assay expression measurements. Barplot showing the expression by luciferase assay of CREs from the CRE-seq library and controls. The 12 CREs from the library are grouped by CRE-seq expression (Low, Middle, High). "Pos" is the positive control from the HS II enhancer; "minP" is the minimal promoter with no enhancer; and "Neg" is the no vector control. The x-axis is sorted by CRE-seq expression, with the exception of "Pos", which was not measured by CRE-seq. Error bars represent the standard error of the mean of up to three transfections per CRE.

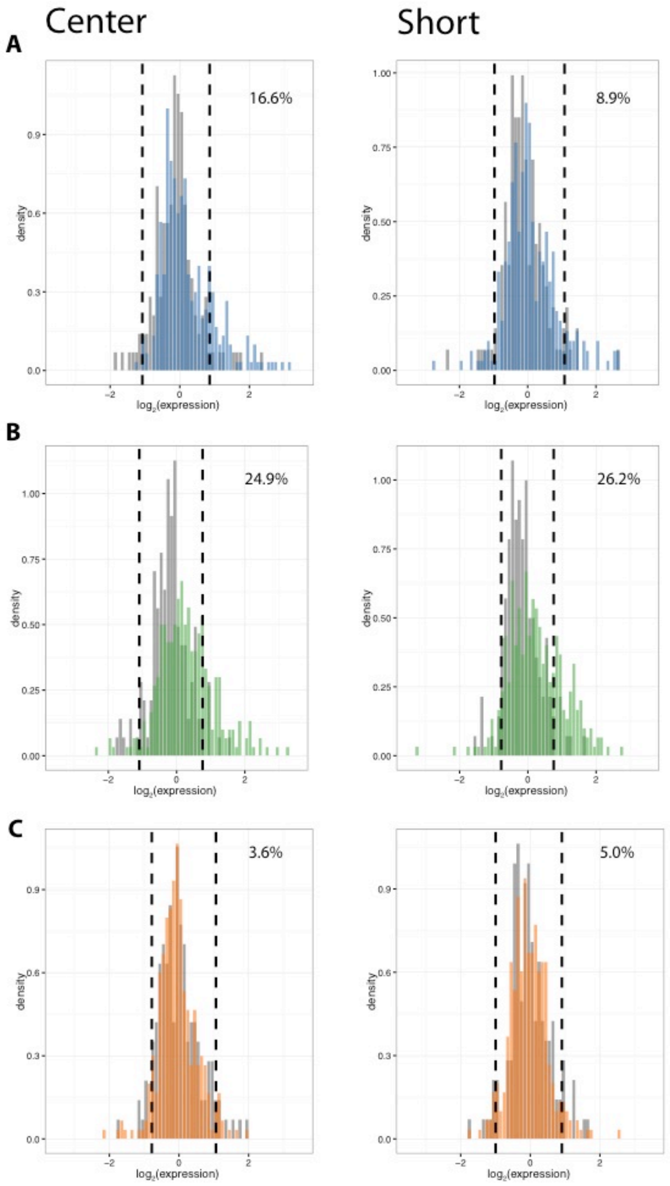


Figure 4.S2. Computing the fraction of active sequences does not depend on the method of choosing short segments. Histograms showing the distribution of expression for each class; **A)** Enhancers (blue), **B)** Weak Enhancers (green), **C)** H1-hESC Enhancers (orange); either for the sequences from 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.

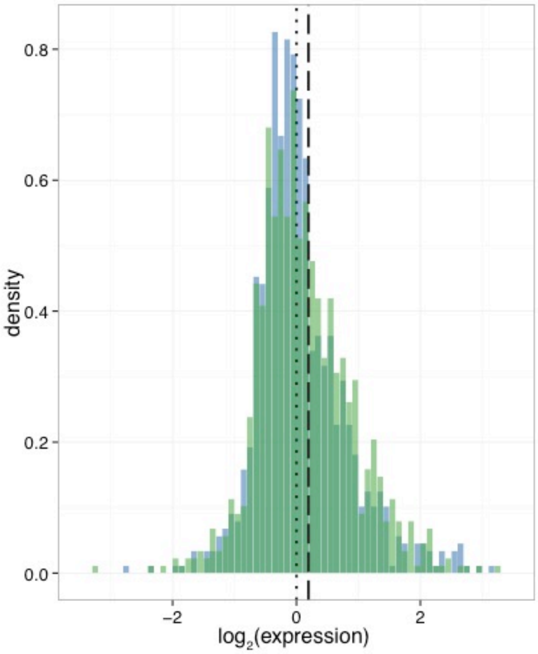


Figure 4.S3. Weak Enhancers control higher median expression than Enhancers. Histogram of expression measurements, showing the distribution of Weak Enhancers (green) shifted to the right of that of Enhancers (blue). Lines show median expression for Enhancers (dotted) and Weak Enhancers (dashed).

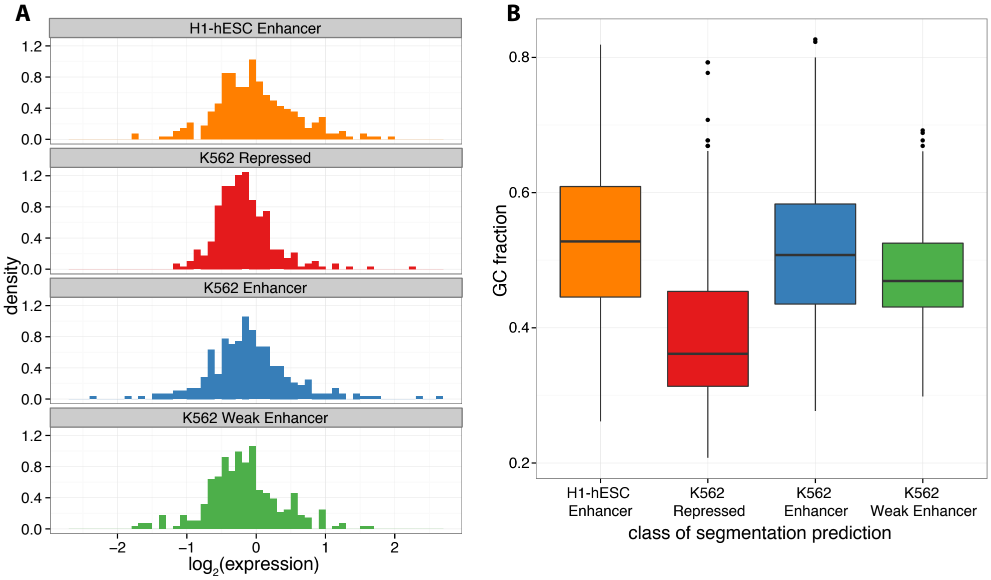


Figure 4.S4. Expression and GC Fraction of Scrambled CREs. A) Histograms showing the expression controlled by each set of scrambled sequences. **B)** Boxplots show the distribution of GC fraction for each category of scrambled sequences.

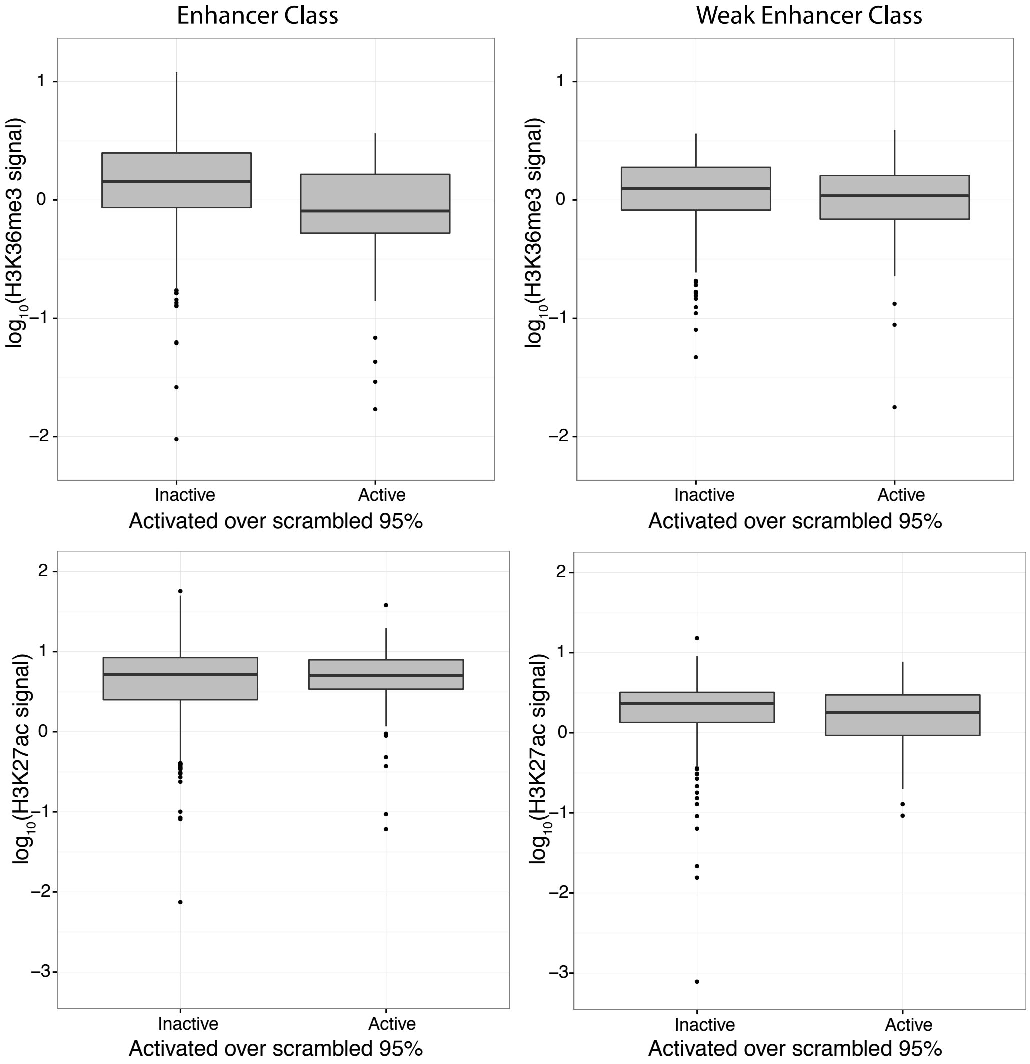


Figure 4.S5. Active Weak Enhancer and Enhancer CREs have lower levels of H3K36me3 and H3K27ac. Boxplots showing that H3K27ac signal and H3K36me3 signal are depleted in active CREs compared to inactive CREs. Plots are similar to 2A and 2C except data is separated by segmentation class. Active CREs are those above the 95th percentile of the scrambled distribution (Table 4.1).

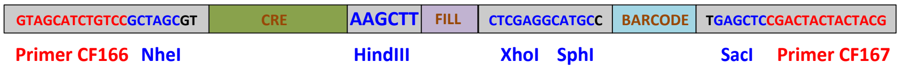


Figure 4.S6. Diagram of 200-mer Oligos Used to Construct the CRE-seq Library. Red sequences are used for PCR priming and blue sequences are restriction enzyme sites. The *CRE* is the 121-130bp sequence from the genomic predictions or scrambled controls. The *FILL* is 0-9bp of random sequence to bring 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.

Chapter 5: Discussion

My thesis has focused on the determinants of *cis*-regulation. A better understanding of the biology underlying transcriptional regulation will allow us to predict the expression driven by potential regulatory elements across conditions and cell types. My work in Chapters 2 and 3 focused on interactions between TFs in the pluripotency network in ES cells and cells undergoing the early stages of differentiation. Using the expression driven by synthetic *cis*-regulatory elements (CREs), I was able to show that interactions and competition between TFs help dictate this expression. Chapter 4 demonstrated that chromatin modifications and sequence features can distinguish genomic regions with *cis*-regulatory potential, and that combining both types of data is necessary for the best predictions. This work shows that progress can be made in predicting expression from DNA regulatory features and that mechanisms of TF activity will continue to be imperative for this understanding.

Synthetic CREs and TF interactions

I have extended the use of synthetic CREs, a tool for studying TF interactions, to a mammalian system. Combining synthetic CREs with a massively parallel reporter assay, such as CRE-seq, provides the means to assay the expression of hundreds to thousands of patterns of binding sites in any transfectable cell type. I have used this system in mouse embryonic stem (ES) cells, a previously intractable system for high-throughput testing of CRE expression. Previous groups have measured the expression of many promoters in human cell types, but this was done using one-at-a-time luciferase assays that are extremely labor and time intensive (Landolin et al. 2010). Previous work in the Cohen lab has tested the expression driven by many synthetic CREs in yeast (Gertz et al. 2009; Gertz and Cohen 2009; Mogno et al. 2010; Zeigler and Cohen 2014), which is easier to work with than mammalian systems. This work also required expression measurements in separate cell populations, a method that does not scale well. Furthermore, while measuring gene expression in yeast can take advantage of different growth conditions (Gertz and Cohen 2009), gene expression in mammalian cells can shed light on the regulatory mechanisms important to development. For these reasons, extending the assay of synthetic CREs to mammalian cells using CRE-seq represents significant progress.

I have demonstrated the continued utility of thermodynamic models in learning about the mechanistic basis of *cis*-regulation. Using thermodynamic models with synthetic promoters in yeast can provide for high predictive power, explaining 60-75% of variation in expression (Gertz et al. 2009; Gertz and Cohen 2009; Mogno et al. 2010). Other groups have used thermodynamic models in mammalian systems. These studies usually attempt to model gene expression of endogenous genes based on large intergenic sequences for the purpose of identifying gene regulatory networks, which rarely provide the power to learn rules of *cis*-regulation (Chen and Zhong 2008; Chen et al. 2008a). The thermodynamic modeling described in my dissertation both obtained very high predictive power (72%) and demonstrated the importance of TF interactions to a mammalian developmental system, adding to the body of literature on interactions between TFs (Gertz and Cohen 2009; Parker et al. 2011; Erceg et al. 2014; Smith et al. 2013b; Beer and Tavazoie 2004; Kaplan et al. 2011; Segal et al. 2008). The TFs I worked with in ES cells are important for pluripotency and self-renewal, and many of them also contribute to cancer, as proliferation is important to both ES cells and cancer cells. The interactions between pluripotency TFs will improve our ability to predict the impact of changes to *cis*-regulatory sequence or TF activity, as changes that affect a TF interaction will result in a different effect size than changes that do not. Furthermore, this could help researchers determine which of the large intergenic regions of mammalian genomes are functional, as those with clusters of binding sites for known interacting TFs are more likely to regulate expression.

My findings have shown that some, but not all, interactions between TFs help specify *cis*-regulation. Within the four pluripotency TFs I studied, there are six possible pairwise heterotypic interactions. I found that half of these interactions (three) contribute significantly to expression in the thermodynamic model. While this is a small sample size, it demonstrates that a significant number of possible interactions between TFs known to regulate similar genes may be important. However, these interactions follow additional rules. I found two possible mechanisms for how these interactions could operate based on the arrangement of the binding sites. One of the interactions worked regardless of how the binding sites were arranged, whereas the other two interactions only operated when the TFs were bound with no other TFs bound in between. Additionally, one of the interactions depended on the order in which the TFs were bound in relation to the basal promoter. This suggests that there may be even more interactions between TFs that are only relevant in contexts that I did not assay. I also found that as cells differentiate, some of the interactions likely remain constant, whereas others seem to change. Thus, the *cis*-regulatory picture that is emerging will be a complex one in which certain TFs interact in certain contexts and certain cell types.

One of the most interesting TF interactions to come out of the thermodynamic modeling was the negative homotypic interaction that applies to any TF. This interaction is consistent with the findings of other groups showing that homotypic chains of TF binding sites saturate in expression (Smith et al. 2013b; Sharon et al. 2012). I found this effect to be strongest for Oct4, Sox2, and Esrrb, in which four sites for any one factor drives expression barely over basal. However, when I investigated the expression driven by many Klf4 sites, I saw that these could drive high expression, possibly due to the fact that multiple Klf factors can bind to the site. This is more consistent with others observations that certain TFs show self-cooperativity and that homotypic clusters of some binding sites are important to enhancers (Pan and Nussinov 2009; Segal et al. 2008; Gotea et al. 2010). From these results, I can conclude that homotypic chains of binding sites possess unique *cis*-regulatory properties that depend on the binding site.

Predicting active genomic regulatory elements

In Chapter 4 I have shown that prediction of which genomic sequences show *cis*-regulatory potential based on chromatin and sequence features do have some accuracy, but also need to be improved. We showed that given the best segmentation algorithms developed by other groups based on chromatin features (Ernst and Kellis 2012; Hoffman et al. 2012a, 2012b), some active *cis*-regulatory elements can be found above background. Furthermore, logistic regression models I used suggest that regions with binding sites for known activators, along with some chromatin signals that correspond with activation (such as DNase HS), are most likely to have *cis*-regulatory function. However, two main observations suggest that our understanding of the mechanisms of histone modifications needs refinement. First, H3K27ac, a mark commonly association with activation, was not associated with active regulatory regions, and may even be weakly associated with inactive regions. Second, sequences classified by the other groups as Weak Enhancers based on chromatin features actually drove higher expression than Enhancers. The continued development of methods to predict *cis*-regulatory potential from genomic features should focus on elements whose activity in our assay did not match the predicted activity.

This work may also improve predictions of which genomic regions have regulatory function that could impact human disease. The ability to find active regions in mammalian genomes is very important, as only a subset of non-coding regions have important function.

Despite this, there is a lot of genomic space that does have function, as evidenced both by selection on non-coding regions and the fact that many areas associated with disease in GWAS fall in non-coding regions (Dunham et al. 2012; Meader et al. 2010; Rands et al. 2014; Cooper et al. 2005; Maurano et al. 2012). Better predictions of active regulatory sequences may improve our ability to find variants that are casual in disease.

The expression data of these genomic regions also makes suggestions on how much of the intergenic region of the human genome is functional. We found around 26% of predicted enhancer regions drive expression over that of random DNA in K562 cells. About 1.2% of the genome was predicted to be an enhancer (Hoffman et al. 2012b) in K562 cells, suggesting that only about 0.3% of the intergenic region of the human genome can activate transcription in K562 cells. However, there are many cell types and tissues in the human genome, each of which use unique *cis*-regulatory regions. It's also likely that some of the repressed elements may have function, working to keep the region off and expression low. Regardless, this suggests that in any given cell type, the amount of the genome working to actively regulate transcription is rather small.

Future Directions

While I have made progress in understanding *cis*-regulation, there is significant room for improvement. My thermodynamic model of expression in ES cells achieved very high predictive power with an R^2 of 0.72 between observed and predicted expression levels. This may be close to the upper limit of the ability of the model to explain expression without exhausting all possible measurements. However, I believe there is still potential for better understanding of the expression driven by CREs with all four TF binding sites, of which there were only 20 in my original library. Since these CREs contain exactly one binding site for each factor, differences in the relative ability of each binding site to recruit polymerase cannot explain the variation in expression driven by these CREs, nor can any homotypic interaction parameters, which were a major component of my model. Regardless, the thermodynamic model explains 36% of the variation in expression of these CREs using two TF interactions. This is good, but still substantially below the level seen in the whole library. Thus, more specific rules incorporating order or orientation of the binding sites may improve the predictions and our understanding of *cis*-regulation.

The fact that interactions between TFs play an important role in dictating the expression level driven by CREs means that more in-depth studies will be needed. My synthetic CREs used a fixed sequence background with fixed spacing between TF binding sites and only one consensus site per TF. It's possible that changes in the sequence background, spacing, and affinity would alter the TF interactions or *cis*-regulatory dynamics. Furthermore, I only looked at four TFs, a tiny fraction of the number of TFs in a mammalian genome. More work on evaluating other TF binding sites in different combinations and contexts would help shed light on what TF interactions exist and their importance in predicting expression. Testing short genomic regions with binding sites could help determine how robust TF interactions are to new contexts. While the thermodynamic model based on expression of my synthetic CREs is able to predict which genomic windows (with at least one binding site for one of the four TFs in my library) have a RNA Polymerase II (RNAP) binding peak with a reasonable degree of accuracy (AUC=0.74), there is still substantial room for improvement. The model is unable to quantitatively predict either RNAP binding or expression of endogenous genomic genes. A natural next step would be to expand this model to genomic CREs.

My work into the *cis*-regulatory principles of RA-treated cells leaves substantial room for progress. I showed that as cells differentiate, the relative effect of a set of binding sites changes. This means that studies into a number of relevant cell types will need to be done, as each cell type has a different set of TFs with their own interactions. Despite learning this general principle, I was unable to model with much accuracy the expression of CREs in RA-treated cells using a thermodynamic model. There are a few possible reasons to explain this. RA-treated cells display much more heterogeneity in their morphology than ES cells, which would lead to more heterogeneity in their gene expression and lower reproducibility of CRE-seq data. This could possibly be overcome by enriching the cell population for cells in a certain state, thus reducing the heterogeneity and improving the CRE-seq data quality. It's also possible that the plasmids with the CREs are lost during differentiation. Integrating the CREs into the genome of ES cells, especially at a fixed locus, and then differentiating them may increase the signal. Lastly, these CREs may have lower absolute expression in the RA-treated cells, as the CREs are comprised of binding sites known to be important in ES cells and not necessarily as important in RA-treated cells. The lower expression could lead to lower expression quality simply due to lower signal. Using a library with new binding sites thought to be more active in RA-treated cells could help with this problem. Alternatively, it is possible that the *cis*-regulation in RA-treated cells is controlled by processes that are simply not captured by my thermodynamic modeling. More work into learning the rules of *cis*-regulation during the process of differentiation is warranted, and solving some of these problems would help towards that aim.

Despite the gains made in identifying active *cis*-regulatory elements in the genome, there is still a lot of progress to be made before accurate predictions can be made. We found that at best about 26% of predicted enhancers from the ENCODE group (Hoffman et al. 2012b) were active over background. Using my best logistic regression model to identify active CREs, I found an area under the receiver operating characteristic curve (AUC) with cross-validation of 0.8, which is good but still leaves substantial predictive power left. It also does a poor job of predicting quantitative expression level (rather than just classifying active sequences). In order to improve this predictive ability, expression measurements of more sequences will be needed. These sequences could be selected to test more specific hypothesis (such as whether dips in H3K27ac signal are predictive of expression) or general predictions of enhancer regions. Additionally, it will help to measure the chromatin signals in the same context as the expression measurements are made. In my case, the chromatin signal was measured in a genomic context, and the expression measurements were made in a plasmid-based transient assay in a different sequence context. One option would be to measure the chromatin signal from the plasmid reporter gene. Alternatively, a reporter assay could be performed in an integrated genomic context at a fixed location and the chromatin signal measured from there. This is likely to be the most relevant measure, as its unclear how well chromatin on a plasmid recapitulates the chromatin in the genome. We found sequence features to be the best predictor of expression, and these should be easier to work with using the sequence data currently available. In this work, I used known binding motifs for TFs to identify sequences with enhancer activity, and others have used k-mers to identify sequences bound by enhancer proteins (P300) (Lee et al. 2011b). Both types of features point to TF binding as the determinant of expression. More work into *cis*-regulatory logic, such as my work in Chapter 2, will help to elucidate how these sequence features control expression.

Conclusions

It has become clear through my dissertation work that predictions of expression from sequence are possible but that more understanding of the underlying processes are needed. Transcriptional regulation is a vital process that if disrupted, can have major consequences for disease and development. It involves the work of many proteins interacting with each other and with DNA. As such, it is a complicated process to work out. I have helped show that it is possible to ascertain some rules for transcriptional regulation, and that mechanistic models can help us learn these rules. It is imperative that work continues in this area, so that we can gain a better understanding of this process and how it affects larger biological processes.

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Autom Contr* **19**: 716–723.

Ambrosetti DC, Schöler HR, Dailey L, Basilico C. 2000. Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the Fibroblast growth factor-4 enhancer. *J Biol Chem* **275**: 23387–23397.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **42**: 255–9.

Arnosti DN. 2003. Analysis and function of transcriptional regulatory elements: insights from Drosophila. *Annu Rev Entomol* **48**: 579–602.

Arnosti DN, Barolo S, Levine M, Small S. 1996. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**: 205–14.

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **22**: 1723–34.

Bailey TL. 2011. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**: 1653–1659.

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**: 823–837.

Bates DM, Watts DG. 1988. *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons.

Beer M a, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* **117**: 185–98.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–5.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Bishop EP, Rohs R, Parker SCJ, 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**: 1314–20.

Boyer L a, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**: 947–56.

Brent R, Ptashne M. 1985. A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* **43**: 729–36.

Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. 2014. The transcription factor titration effect dictates level of gene expression. *Cell* **156**: 1312–23.

Buchler NE, Gerland U, Hwa T. 2003. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* **100**: 5136–41.

Bussemaker HJ, Li H, Siggia ED. 2001. Regulatory element detection using correlation with expression. *Nat Genet* **27**: 167–71.

Chambers I, Silva J, Colby D, Nichols J, Nijmeijer B, Robertson M, Vrana J, Jones K, Grotewold L, Smith A. 2007. Nanog safeguards pluripotency and mediates germline development. *Nature* **450**: 1230–4.

Chambers I, Tomlinson SR. 2009. The transcriptional foundation of pluripotency. *Development* **136**: 2311–22.

Chantalat S, Depaux A, Héry P, Barral S, Thuret JY, Dimitrov S, Gérard M. 2011. Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res* **21**: 1426–1437.

Chen C-C, Zhong S. 2008. Inferring gene regulatory networks by thermodynamic modeling. *BMC Genomics* **9 Suppl 2**: S19.

Chen C-C, Zhu X-G, Zhong S. 2008a. Selection of thermodynamic models for combinatorial control of multiple transcription factors in early differentiation of embryonic stem cells. *BMC Genomics* **9**: S18.

Chen CTL, Gottlieb DI, Cohen B a. 2008b. Ultraconserved elements in the Olig2 promoter. *PLoS One* **3**: e3946.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008c. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**: 1106–17.

Chew J-L, Loh Y-H, Zhang W, Chen X, Tam W-L, Yeap L-S, Li P, Ang Y-S, Lim B, Robson P, et al. 2005. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol* **25**: 6031–46.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**: 901–13.

Cosma MP, Tanaka T, Nasmyth K. 1999. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell* **97**: 299–311.

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**: 21931–21936.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, Minokawa T, Amore G, Hinman V, Arenas-Mena C, et al. 2002. A genomic regulatory network for development. *Science* **295**: 1669–78.

Dröge P, Müller-Hill B. 2001. High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. *Bioessays* **23**: 179–83.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a., 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**: 57–74.

Ema M, Mori D, Niwa H, Hasegawa Y, Yamanaka Y, Hitoshi S, Mimura J, Kawabe YI, Hosoya T, Morita M, et al. 2008. Krüppel-like factor 5 Is Essential for Blastocyst Development and the Normal Self-Renewal of Mouse ESCs. *Cell Stem Cell* **3**: 555–567.

Eminli S, Foudi A, Stadtfeld M, Maherali N, Ahfeldt T, Mostoslavsky G, Hock H, Hochedlinger K. 2009. Differentiation stage determines potential of hematopoietic cells for reprogramming into induced pluripotent stem cells. *Nat Genet* **41**: 968–976.

Erceg J, Saunders TE, Girardot C, Devos DP, Hufnagel L, Furlong EEM. 2014. Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. *PLoS Genet* **10**: e1004060.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–6.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**: 817–25.

Evans MJ, Kaufman MH. 1981. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**: 154–156.

Feng B, Jiang J, Kraus P, Ng J-H, Heng J-CD, Chan Y-S, Yaw L-P, Zhang W, Loh Y-H, Han J, et al. 2009. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* **11**: 197–203.

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**: 1055–64.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41**: D48–55.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed Y Bin, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**: 58–64.

Gagliardi A, Mullin NP, Ying Tan Z, Colby D, Kousa AI, Halbritter F, Weiss JT, Felker A, Bezstarosti K, Favaro R, et al. 2013. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J* 1–17.

Gertz J, Cohen BA. 2009. Environment-specific combinatorial cis-regulation in synthetic promoters. *Mol Syst Biol* **5**: 244.

Gertz J, Siggia ED, Cohen B a. 2009. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**: 215–218.

Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, Loftus SK, Beer M a, 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**: 2290–301.

Gotea V, Visel A, Westlund JM, Nobrega M a, Pennacchio L a, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**: 565–77.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**: R24.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**: 99–104.

He X, Samee MAH, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. ed. U. Ohler. *PLoS Comput Biol* **6**: e1000935.

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**: 576–89.

Hess J, Angel P, Schorpp-Kistner M. 2004. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci* **117**: 5965–73.

Hochschild A, Irwin N, Ptashne M. 1983. Repressor structure and the mechanism of positive control. *Cell* **32**: 319–325.

Hochschild A, Ptashne M. 1986. Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. *Cell* **44**: 681–7.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012a. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Meth* **9**: 473–476.

Hoffman MM, Ernst J, Wilder SP, Kundaje a., Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes J a., Birney E, et al. 2012b. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 1–15.

Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR. 2006. Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**: 533–8.

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318–356.

Jacobs S, Lie DC, DeCicco KL, Shi Y, DeLuca LM, Gage FH, Evans RM. 2006. Retinoic acid is required early during adult neurogenesis in the dentate gyrus. *Proc Natl Acad Sci U S A* **103**: 3902–3907.

Jiang J, Chan Y-S, Loh Y-H, Cai J, Tong G-Q, Lim C-A, Robson P, Zhong S, Ng H-H. 2008. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* **10**: 353–60.

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**: 327–339.

Kaplan T, Li X-Y, Sabo PJ, Thomas S, Stamatoyannopoulos J a., Biggin MD, Eisen MB. 2011. Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development ed. G.S. Barsh. *PLoS Genet* **7**: e1001290.

Kazemian M, Blatti C, Richards A, McCutchan M, Wakabayashi-Ito N, Hammonds AS, Celniker SE, Kumar S, Wolfe S a, Brodsky MH, et al. 2010. Quantitative analysis of the

Drosophila segmentation regulatory network using pattern generating potentials. *PLoS Biol* **8**.

Keegan L, Gill G, Ptashne M. 1986. Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* **231**: 699–704.

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 2,000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*.

Kim HD, O'Shea EK. 2008. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* **15**: 1192–8.

Kim J, Chu J, Shen X, Wang J, Orkin SH. 2008. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**: 1049–61.

Kim TK, Maniatis T. 1997. The mechanism of transcriptional synergy of an in vitro assembled interferon-beta enhanceosome. *Mol Cell* **1**: 119–29.

Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A*.

Kinney JB, Murugan A, Jr CGC, Cox EC. 2008. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence : SI Appendix. **10**: 1–24.

Kranz A-L, Eils R, König R. 2011. Enhancers regulate progression of development in mammalian cells. *Nucleic Acids Res* **39**: 8689–8702.

Kulkarni MM, Arnosti DN. 2003. Information display by transcriptional enhancers. *Development* **130**: 6569–75.

Kuroda T, Tada M, Kubota H, Kimura H, Hatano S, Suemori H, Nakatsuji N, Tada T. 2005. Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. *Mol Cell Biol* **25**: 2475–85.

Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res*.

Kwasnieski JC, Mogno I, Myers C a, Corbo JC, Cohen B a. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**: 19498–503.

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* 890–898.

Lee D, Karchin R, Beer M a. 2011a. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* 2167–2180.

Lee D, Karchin R, Beer M a. 2011b. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21**: 2167–80.

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**: 2522–2540.

Li J, Biggin MD. 2015. Gene expression. Statistics requantitates the central dogma. *Science (80-)* **2018**: 2013–2015.

Li XY, MacArthur S, Bourgon R, Nix D, Pollard D a., Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**: 0365–0388.

Li Y, McClintick J, Zhong L, Edenberg HJ, Yoder MC, Chan RJ. 2005. Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. *Blood* **105**: 635–637.

Li Y, Zhang Q, Yin X, Yang W, Du Y, Hou P, Ge J, Liu C, Zhang W, Zhang X, et al. 2010. Generation of iPSCs from mouse fibroblasts with a single gene, Oct4, and small molecules. *Cell Res* **0325901**: 1–9.

Liu X, Huang J, Chen T, Wang Y, Xin S, Li J, Pei G, Kang J. 2008. Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Res* **18**: 1177–89.

Loh Y-H, Wu Q, Chew J-L, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, et al. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**: 431–40.

Malik S, Roeder RG. 2000. Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. *Trends Biochem Sci* **25**: 277–83.

Martin GR. 1981. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A* **78**: 7634–7638.

Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, Okochi H, Okuda A, Matoba R, Sharov A a, et al. 2007. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol* **9**: 625–35.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, et al. 2014. JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**.

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**: 1190–5.

Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* **20**: 1335–43.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 1–9.

Mogno I, Kwasnieski JC, Cohen B a. 2013. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* **23**: 1908–1915.

Mogno I, Vallania FLM, Mitra RD, Cohen B a. 2010. TATA is a modular component of synthetic promoters. *Genome Res*.

Moses AM, Pollard D a., Nix D a., Iyer VN, Li XY, Biggin MD, Eisen MB. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* **2**: 1219–1231.

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**: 63–72.

Nagy a, Rossant J, Nagy R, Abramow-Newerly W, Roder JC. 1993. Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. *Proc Natl Acad Sci U S A* **90**: 8424–8428.

Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S. 2008. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol* **26**: 101–6.

Nakatake Y, Fukui N, Iwamatsu Y, Masui S, Takahashi K, Yagi R, Yagi K, Miyazaki J-I, Matoba R, Ko MSH, et al. 2006. Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. *Mol Cell Biol* **26**: 7772–82.

Newburger DE, Bulyk ML. 2009. UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **37**.

Ng CKL, Li NX, Chee S, Prabhakar S, Kolatkar PR, Jauch R. 2012. Deciphering the Sox-Oct partner code by quantitative cooperativity measurements. *Nucleic Acids Res* 1–9.

Nili EL, 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**: 1361–1368.

Nishiyama A, Sharov A a, Piao Y, Amano M, Amano T, Hoang HG, Binder BY, Tapnio R, Bassey U, Malinou JN, et al. 2013. Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Sci Rep* **3**: 1390.

Nishiyama A, Xin L, Sharov A a, Thomas M, Mowrer G, Meyers E, Piao Y, Mehta S, Yee S, Nakatake Y, et al. 2009. Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell* **5**: 420–33.

Niwa H, Miyazaki J, Smith a G. 2000. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* **24**: 372–376.

Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, Habib N, Yosef N, Chang CY, Shay T, et al. 2011. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**: 296–309.

Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, et al. 2004. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**: 1378–1381.

Oliveri P, Tu Q, Davidson EH. 2008. Global regulatory logic for specification of an embryonic cell lineage. *Proc Natl Acad Sci U S A* **105**: 5955–62.

Pan Y, Nussinov R. 2009. Cooperativity dominates the genomic organization of p53-response elements: a mechanistic view. *PLoS Comput Biol* **5**: e1000448.

Parker DS, White M a, Ramos AI, Cohen B a, Barolo S. 2011. The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Sci Signal* **4**: ra38.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012a. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 1–9.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012b. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*.

Ptashne M. 2005. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci* **30**: 275–9.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet* **10**: e1004525.

Raveh-Sadka T, Levo M, Segal E. 2009. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* **19**: 1480–96.

Ridinger-Saison M, Boeva V, Rimmelé P, Kulakovskiy I, Gallais I, Levavasseur B, Paccard C, Legoix-Né P, Morlé F, Nicolas A, et al. 2012. Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Res* **40**: 8927–8941.

Rodda DJ, Chew J-L, Lim L-H, Loh Y-H, Wang B, Ng H-H, Robson P. 2005. Transcriptional regulation of nanog by OCT4 and SOX2. *J Biol Chem* **280**: 24731–7.

Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. 2009. The role of DNA shape in protein-DNA recognition. *Nature* **461**: 1248–1253.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**: 535–40.

Setty Y, Mayo a E, Surette MG, Alon U. 2003. Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci U S A* **100**: 7702–7.

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**: 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**: 211–30.

Smith RP, Riesenfeld SJ, Holloway AK, Li Q, Murphy KK, Feliciano NM, Orecchia L, Oksenberg N, Pollard KS, Ahituv N. 2013a. A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design. *Genome Biol* **14**: R72.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013b. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* 1–10.

Solter D, Solter D. 2006. From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. *Nat Rev Genet* **7**: 319–27.

Soprano DR, Teets BW, Soprano KJ. 2007. Role of Retinoic Acid in the Differentiation of Embryonal Carcinoma and Embryonic Stem Cells. *Vitam Horm* **75**: 69–95.

Stadtfeld M, Brennand K, Hochedlinger K. 2008. Reprogramming of Pancreatic β Cells into Induced Pluripotent Stem Cells. *Curr Biol* **18**: 890–894.

Stormo GD, Fields DS. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* **23**: 109–113.

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**: 41–45.

Struhl K, Segal E. 2013. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**: 267–73.

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**: 861–72.

Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**: 663–76.

Teo AKK, Arnold SJ, Trotter MWB, Brown S, Ang LT, Chng Z, Robertson EJ, Dunn NR, Vallier L. 2011. Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev* **25**: 238–50.

Thanos D, Maniatis T. 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**: 1091–100.

Thomson M, Liu SJ, Zou L-N, Smith Z, Meissner A, Ramanathan S. 2011. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* **145**: 875–89.

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**: 75–82.

Van den Berg DLC, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I, Poot R a. 2010. An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* **6**: 369–81.

Van den Berg DLC, Zhang W, Yates A, Engelen E, Takacs K, Bezstarosti K, Demmers J, Chambers I, Poot R a. 2008. Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. *Mol Cell Biol* **28**: 5986–95.

Von Hippel PH, Berg OG. 1986. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A* **83**: 1608–1612.

Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W. 2013. Klf4 organizes long-range chromosomal interactions with the OCT4 locus inreprogramming andpluripotency. *Cell Stem Cell* **13**: 36–47.

Wei Z, Yang Y, Zhang P, Andrianakos R, Hasegawa K, Lyu J, Chen X, Bai G, Liu C, Pera M, et al. 2009. Klf4 interacts directly with Oct4 and Sox2 to promote reprogramming. *Stem Cells* **27**: 2969–78.

White M a, Myers C a, Corbo JC, Cohen B a. 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**: 11952–7.

Whiteld TW, Wang J, Collins PJ, Partridge EC, Aldred S, Trinklein ND, Myers RM, Weng Z. 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol* **13**: R50.

Wu C-Y, Feng X, Wei L-N. 2014. Coordinated repressive chromatin-remodeling of Oct4 and Nanog genes in RA-induced differentiation of embryonic stem cells involves RIP140. *Nucleic Acids Res* 1–12.

Xian HQ, Werth K, Gottlieb DI. 2005. Promoter analysis in ES cell-derived neural cells. *Biochem Biophys Res Commun* **327**: 155–162.

Xie D, Cai J, Chia N-Y, Ng HH, Zhong S. 2008. Cross-species de novo identification of cis-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells. *Genome Res* **18**: 1325–35.

Yáñez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A. 2012. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res* **22**: 2018–30.

Yeo J-C, Jiang J, Tan Z-Y, Yim G-R, Ng J-H, Göke J, Kraus P, Liang H, Gonzales KAU, Chong H-C, et al. 2014. Klf2 is an essential factor that sustains ground state pluripotency. *Cell Stem Cell* **14**: 864–72.

Yuan X, Wan H, Zhao X, Zhu S, Zhou Q, Ding S. 2011. Combined Chemical Treatment Enables Oct4-Induced Reprogramming from Mouse Embryonic Fibroblasts. *Stem Cells*.

Zeigler RD, Cohen B a. 2014. Discrimination between thermodynamic models of cis-regulation using transcription factor occupancy data. *Nucleic Acids Res* **42**: 2224–34.

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

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**: 65–70.