Combinatorial gene regulation by transcription factors

by

Sharon R. Grossman

B.A. Chemical and Physical Biology and Mathematics
Harvard University

Submitted to the Department of Biology
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Signature redacted

Signature of Author .....
Sharon R. Grossman
Department of Biology

Signature redacted

Certified by .....
Eric S. Lander
Professor of Biology
Thesis Supervisor

Signature redacted

Accepted by .....
Amy Keating
Professor of Biology
Co-Chair, Graduate Committee

Combinatorial gene regulation by transcription factors

by Sharon R. Grossman

Submitted to the Department of Biology
on April 2nd, 2019 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ABSTRACT

Combinatorial gene regulation is encoded in enhancers and promoters in the form of binding sites for transcription factors (TFs), which collaboratively recruit the transcriptional machinery and drive gene expression. Using high-throughput and quantitative technologies developed by our lab and others, we studied TF binding sites in enhancers from numerous different cell types and regulatory systems, shedding light general principles of motif composition and organization in typical cellular regulatory elements.

We find extensive synergy between TF binding sites, some with organizational constraints and some with flexible positioning. We demonstrate that different TFs bind at distinct positions within regulatory elements, suggesting a new type of architectural constraint in enhancers. Importantly, our analysis of both TF organization and cooperativity revealed distinctive patterns that separates TFs into potential functional classes.

Together, our results suggest a structure of the regulatory code at the level of TF function and generate new hypotheses about regiospecific binding patterns and functions of TF classes within enhancers.

Thesis Supervisor: Eric S. Lander
Title: Professor of Biology

Acknowledgements

Graduate school has been a period of enormous scientific and personal growth for me. I could not have done it without the support of my colleagues, mentors, friends, and family, who were there for me through the highs and lows. In particular, I would like to thank the following people:

- I cannot express enough gratitude for my advisor, Eric Lander, who inspired me with his love of biology and genuine curiosity about life, taught me both scientific rigor and vision, and supported and believed in me through some of the most difficult personal challenges of my life. His confidence in my ability to grow and succeed made me into the scientist and person I am today.

- My thesis committee, Chris Burge, Aviv Regev, and Brad Bernstein, for their scientific guidance and insight.

- Jesse Engreitz, who has been my colleague and friend since the very beginning of our rotations, and spent countless hours discussing science and life. Having a comrade through graduate school meant more than I can say.

- Tim Wang, Charlie Fulco, and Brian Cleary for scientific discussion and commiseration throughout our graduate careers.

- Nicole Brellethin and Kate Mulherin, for making everything run smoothly and always being there for a smile or cup of coffee.

- Pardis Sabeti, whose advice shaped my scientific path, and who had been the best mentor and friend anyone could hope for.

- My parents, sister, and grandparents, whose unwavering love and support inspired me to follow my passions and gave me the confidence and strength to get to where I am today.

- Kristin Knouse, Eric Bent, Jeremiah Wala, Will Gibson, Xenos Mason, Jennifer Lo, Belinda Wang, Dave and Yakir Reshef, and Hilary Finucane, who friendship and laughter made graduate school fun and inspiring, and taught me so much about life.

Table of Contents

Chapter 1

Introduction

SECTION I. Overview

Control of gene expression plays a key role in a multitude of biological processes, from organismal development and cell differentiation to response to environmental stresses. Since the discovery of regulatory DNA in the 1960s, deciphering the cis-regulatory "code" that specifies how gene expression is encoded in regulatory sequences has been a major research focus. In addition to shedding light on core biological process, the ability to "read" and "write" the regulatory code has broad implications for human health: deviations from the appropriate expression patterns due to mutations in regulatory sequences or transcriptional regulators underlie the formation of many cancers and other diseases (1-3), and the ability to design novel regulatory sequences with desired properties opens the way to targeted, specific treatments. Furthermore, the majority of evolutionary changes in animal morphology and other traits stem from mutations in developmental regulatory sequences (4-7).

Information about when, where, and to what level each gene should be expressed is encoded in cis-regulatory sequences such as promoters and enhancers, in the form of organized arrays of binding sites for sequence-specific transcription factors (TFs). Cellular differentiation and cues from the environment trigger lead to the activation of specific sets of TFs, which bind to enhancers and interact with cofactors to stabilize the transcription-initiation machinery to enable gene expression (8, 9).

Compared to bacteria, metazoan systems employ vastly more elaborate protein machineries to carry out transcription (10). TFs almost never function alone, instead relying on combinatorial input from multiple partner TFs to drive robust expression. In addition, the general protein machineries required for core promoter recognition and basal transcription in metazoans include numerous diverse and functionally specialized components. Complex systems of cofactors are involved in mediating signals between TFs and the transcriptional machinery and remodeling local chromatin structure, adding a final layer of regulation (11).

Over the past four decades, the molecular players in transcription have largely been identified, but the detailed mechanistic steps involved in enhancer-mediated transcriptional activation remain unclear. The vast majority interactions between TFs, cofactors, and basal transcriptional machinery are not known. Without knowing the biochemical activities of each bound TF, it is hard to interpret dependencies observed between TF binding sites in functional characterizations of regulatory sequences. Basic mechanistic questions about enhancer structure and function remain, such as: how many distinct biochemical activities are required for transcription? Do individual TF binding sites contribute to different functional steps? If so, is there a universal formula at the level of TF function?

From a systems perspective, cells "compute" information from input TFs via regulatory sequences to determine the output expression pattern. In keeping with the intricate molecular underpinnings of transcription, the computations performed by enhancers are remarkably complex, generating many different relationships between the combination and concentration of input TFs and the output gene expression (12,

13). Many aspects of the enhancer sequence can affect transcriptional output, including the identity, number and affinity of TF binding sites, their arrangement and spacing within the enhancer, and potential synergies between the bound TFs (14). With only a handful of well-characterized enhancer sequences and little mechanistic understanding of combinatorial TF function, a global cis-regulatory code has remained elusive (15-17).

A major breakthrough in our ability to investigate regulatory sequences and how they encode gene expression came with the development of high-throughput methods to identify enhancers by mapping genome-wide chromatin states (18-20) and TF occupancy (21, 22). However, until recently these predicted enhancers could only be functionally characterized using low-throughput methods, precluding the systematic dissection of functional elements within regulatory sequences and the generation of large datasets for computational modeling. A second breakthrough occurred shortly before I began my Ph.D., when our lab and others developed high-throughput parallel enhancer assays enabling measurements of thousands of regulatory elements in one experiment (23-25).

In this thesis, I present our contribution towards deciphering regulatory DNA and uncovering a general cis-regulatory code using the unprecedented ability to comprehensively identify native genomic enhancers, rapidly characterize their regulatory activity, and dissect their functional properties. In Chapter 1, we develop a generalizable framework for deciphering cis-regulatory grammar through (i) systematic identification of functional TF binding sites, (ii) experimental quantification of the activity of sites corresponding to each TF, and (iii) analysis of patterns of cooperativity between TFs and architectural constraints. We used this approach to comprehensively

characterize PPARγ-response elements (PPREs) in adipocytes. Comparing of the patterns of cooperativity between different factors revealed clusters of TFs marked by similar interaction patterns with other TFs. This "modular epistasis" suggests the intriguing possibility of functional "equivalency" classes of TFs, which might contribute different regulatory functions. Consistent with this idea, the classes showed similar effects on expression when perturbed and were associated with known biological processes. In Chapter 2, we dig further into functional distinctions between TFs using a different feature of enhancer architecture: binding site position within nucleosome-depleted regions (NDRs). We examine the positional distribution of binding sites in NDRs for 103 TFs across 47 cell types, and show that TFs fall into six distinct classes with strikingly different spatial localization within enhancers. These positional classes bring together factors that have a number of similar properties, and suggests specific functional roles for each class consistent with their NDR localization. This suggests that architectural constraints in enhancers can arise not only from well-known cooperative binding interactions, but also to optimally position a TF to carry out its function. Finally, in Chapter 3, we describe work in progress to experimentally validate the positional classes by showing motif sites in "optimal" positions are more likely to be occupied by their cognate TF and contribute more strongly to expression output in enhancer assays. Moreover, the relative effect sizes of the different classes are consistent with their hypothesized roles.

To provide context for thinking about how gene regulation is encoded in DNA, I first present a brief overview of the eukaryotic transcriptional machinery controlling the activity of RNA polymerase at core promoters and the transcriptional cycle. I review the

discovery and initial characterization of enhancers, and how known properties of enhancers are used today by various methods to predict regulatory elements in the genome. Next, I discuss TF binding and function, examining intrinsic TF DNA-binding specificities, how TFs bind in the genome, and various effector functions of TFs after DNA binding. Finally, I discuss models of enhancer architecture and regulatory grammar, drawing upon deeply characterized test cases as well as recent high-throughput experiments manipulating various features of enhancer architecture.

SECTION II. Molecular mechanisms of transcription at promoters

Expression of protein-coding genes begins with the transcription of the gene's DNA sequence into RNA by RNA polymerase II (RNAPII). Transcription usually begins at a defined location at the 5' end of the gene, the transcription start site (TSS). The TSS is embedded in a core promoter, extending ~50 bp upstream and downstream of the TSS, which constitutes the minimal DNA sequence needed for basal transcription. The core promoter contains recognition sites for the transcriptional machinery, comprising RNAPII and general transcription factors (GTFs). Basal transcription from core promoters is generally low, and can be further suppressed by closed chromatin. Robust transcription requires activation by additional regulatory DNA elements located at varying distances from the TSS called enhancers (Fig. 1A). Enhancers contain binding sites for TFs, which recruit cofactors and can promote or repress transcription.

Transcription from a core promoter is a tightly regulated multistep process resulting in a specified expression output. Since all the processes controlling RNAPII transcription must ultimately lead to the basal transcriptional machinery at the core promoter, understanding the structure and properties of these central elements is

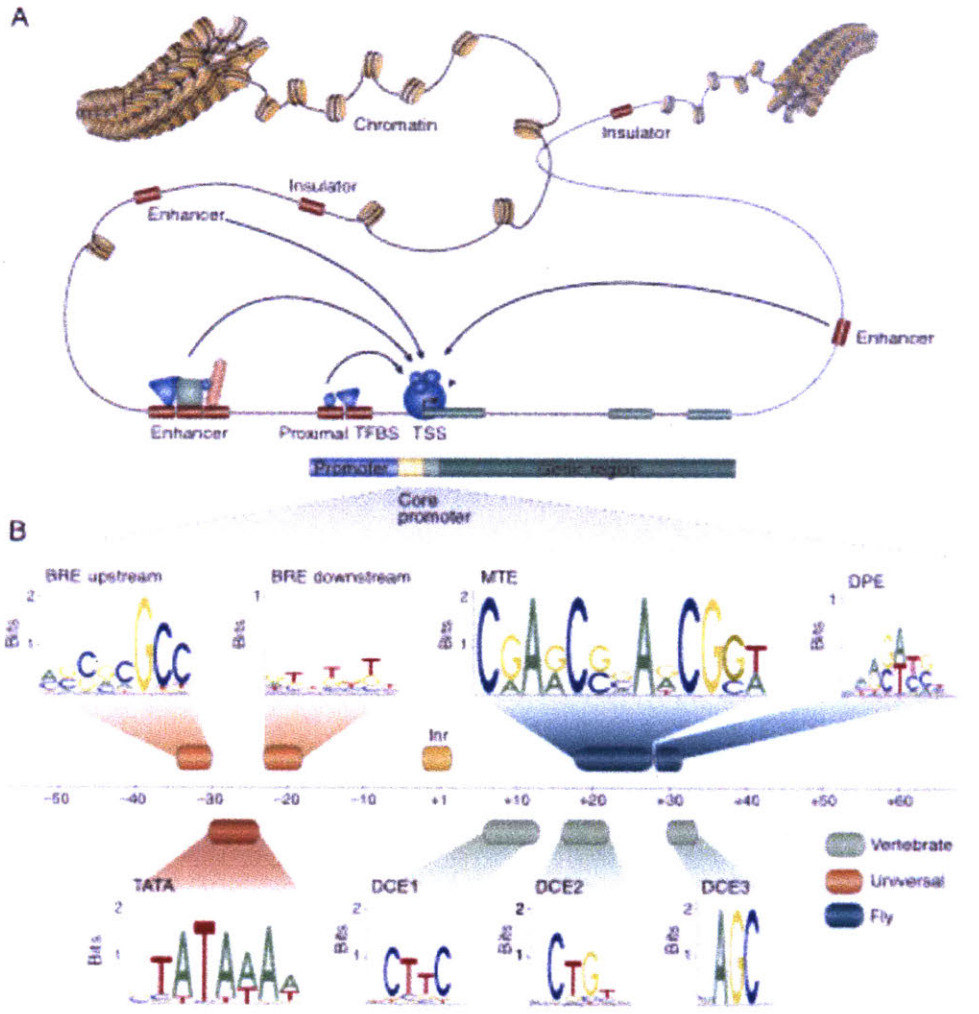
essential to the study of gene regulation. I will therefore begin by reviewing the structure of core promoters and the molecular mechanisms underlying the assembly and activation of the transcriptional machinery. These basic promoter properties of core promoters provide the foundation for understanding how this machinery is regulated by distal enhancers through various combinations of TFs and cofactors.

Properties of gene core promoters

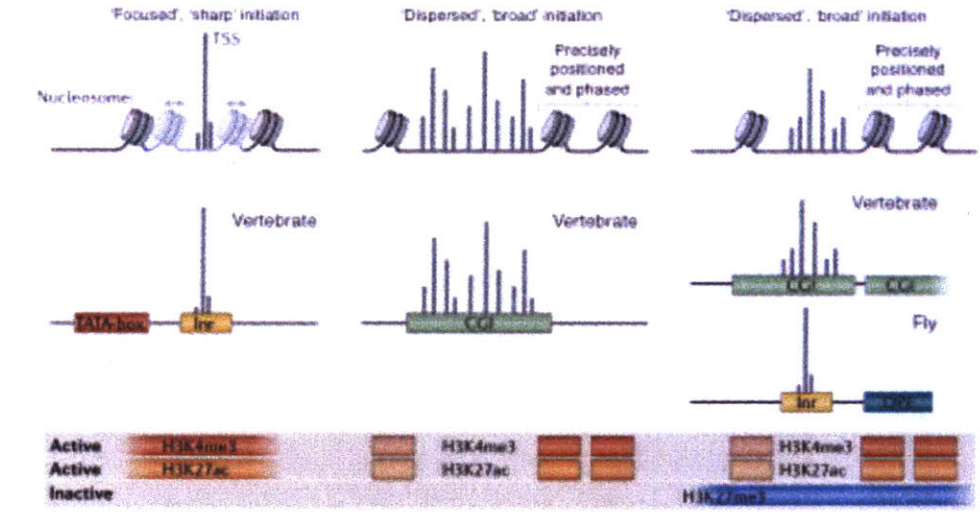
Sequence elements

The core promoter serves as a binding platform to position and support the assembly of the pre-initiation complex (PIC), which includes RNAPII and GTFs (26). Metazoan core promoters contain several interchangeable sequence elements known as core-promoter motifs, which are positioned at fixed distances from the TSS and recruit various GTFs and mediate the assembly of the PIC (Fig. 1B).

The most commonly occurring (though not universal) core-promoter motif that has been identified to date is the initiator (Inr) motif (27). The Inr directly overlaps the TSS and is recognized by several components of the TFIID complex, a GTF involved in recruiting RNAPII and mediating PIC assembly (28, 29). Another well-characterized core-promoter motif is the TATA-box motif, located ~30 bp upstream of the TSS, which is recognized and bound by the TATA-box binding protein (TBP) (30). TBP is a component of the TFIID complex, Although the TATA box is conserved from yeast to humans (30), however, it is found in only a minority of core promoters (e.g. ~5% of core



Adapted from Lenhard et al. 2012



Adapted from Haberle et al. 2018

Figure 1. Regulation of transcription. (A) A summary of promoter elements and regulatory signals. Chromatin is comprised of DNA wrapped around histones to form nucleosomes. The structure of chromatin can be tightly wrapped or accessible to proteins (such as TF). The region around the transcription start site (TSS) is often divided into a larger proximal promoter upstream of the TSS and a smaller core promoter just around the TSS. The exact boundaries vary between studies. To recruit RNAPII and to activate transcription of the gene, TFs bind to specific sequence patterns (TFBSs) that are near to the TSS (proximal elements) or that are far away from it (enhancers). **(B)** Sequence patterns in core promoters. The region around the TSS has several over-represented sequence patterns; the TATA box and initiator (Inr) are the most studied and most prevalent. The location of patterns relative to the TSS and their sequence properties are shown as boxes and as associated sequence logos based on the JASPAR database. The Inr pattern is not shown as it varies considerably between studies. Importantly, most promoters only have one or a few of these patterns, and some patterns are mostly found in certain species. BRE, B recognition element; DCE, downstream core element; DRE, DNA recognition element; MTE, motif ten element. **(C)** Mapping endogenous transcription initiation at single nucleotide resolution revealed striking differences between core promoters, leading to the classification of 'focused' or 'sharp' core promoters, which have a single, well-defined TSS **(left),** and 'dispersed' or 'broad' promoters, which have multiple closely spaced TSSs that are used with similar frequency (middle and right). Figure modified from (31) and (32).

promoters in flies (33, 34) and 10-15% in humans (35-37)). In promoters that lack a TATA-box, the Inr motif is often paired with the downstream promoter element (DPE) instead (38). DPE occurs a fixed distance downstream of the Inr motif, and this spacing facilitates cooperative binding of TFIID to the two elements (29, 39).

In addition to these three relatively abundant core-promoter motifs, several other motifs with fixed positions relative to the TSS have been identified, such as the TFIIB recognition elements (BREs) (40, 41), the motif ten element (MTE) (42), and downstream core elements (DCEs) (43). Like TATA, Inr and DPE, these motifs have been shown to bind specific GTFs in vitro (41, 44), and thus might contribute to PIC positioning and assembly in vivo.

Different combinations of core promoter motifs are associated with different transcription initiation patterns (35, 45, 46). Promoters with TATA boxes and Inr elements are associated with a single, well-defined TSS (35, 47). These promoters are generally associated with highly cell-type specific genes with restricted expression patterns (48). In contrast, core promoters of broadly-expressed genes tend to lack a TATA box and instead contain CpG islands (CGIs), regions with high densities of CpG dinucleotides. These promoters are associated with broad, dispersed transcription initiation from multiple closely spaced TSS (35, 45). Promoters of key developmental genes also show dispersed initiation patterns in mammals, and tend to lack TATA boxes and contain long CGIs (35).

Chromatin features

Active core promoters are marked by distinctive chromatin features that make them accessible to the transcriptional machinery and facilitate the assembly of the PIC. Active core promoters reside in nucleosome-depleted regions (NDRs) flanked by precisely positioned and phased nucleosomes downstream of the TSS (49-51). The first few nucleosomes downstream of the TSS (particularly the +1 nucleosome) are enriched for the histone variant H2A.Z, which has been shown to present a lower barrier to RNAPII transcription than canonical histones (52, 53). Housekeeping gene promoters with broad initiation patterns in particular are associated with clearly defined NDRs and stably positioned flanking nucleosomes (54). Promoters of highly regulated genes with focused initiation patterns are often occluded by nucleosomes before activation, and require nucleosome disassembly by ATP-dependent chromatin remodeling complexes

for activation (55, 56). After activation, focused promoters tend to have more imprecisely positioned nucleosomes (54).

Chromatin surrounding promoters is marked by distinctive patterns of histone modifications. The nucleosomes downstream of active promoters display a high level of tri-methylation on H3 Lys4 (H3K4me3) and acetylation of H3 Lys27 (H3K27ac). The role of these histone modifications in promoter function remains unclear. Histone acetylation can decrease the affinity of DNA for nucleosomes (57), and thus may promote open chromatin (58). Histone modifications may also assist in recruiting elements of the transcriptional machinery (59). For example, a subunit (BPTF) of the nucleosome remodeling factor (NURF) contains two domains that can recognize H3K4me3 and H3K27ac respectively, spanned by a rigid linker that positions them to correctly contact nucleosomes with these modifications (60, 61), and TAF1, a subunit of TFIID, shows increased affinity for acetylated histone H4 (62). However, several studies suggest H3K4me3 and H3K27ac may not be required for transcription. *Drosophila* cells expressing non-methylatable forms of canonical and variant H3 histones can properly regulate transcription (63), and a Lys-to-Arg mutation at position 27 does not reduce transcription (64). Some of these effects may reflect functional redundancy among marks, but further studies of the mechanistic role of chromatin modifications is needed to distinguish causation from correlation and determine their functionality in transcriptional regulation.

RNAPII transcription cycle

PIC assembly and promoter opening

RNAPII transcription begins with the ordered assembly of GTFs at the core promoter to form the PIC (Fig. 2A,B). Over the past two decades, the crystal structures of many of the components of PIC have been resolved, and the PIC structure and assembly has been extensively probed biochemically through mutagenesis and crosslinking mapping (65). Recently, crystallographic and cryo-electron microscopy (EM) studies have elucidated the structures of RNAPII complexes with GTFs (66, 67), verifying the earlier biochemically-derived models (68, 69). Together, these studies provide remarkable structural and mechanistic insights into how RNAPII cooperates with the GTFs to bind to and open promoter DNA, initiate RNA synthesis, and escape the promoter.

The first step in PIC assembly is core promoter recognition and binding by TFIID, which contacts multiple core promoter sequence elements as well as histones with promoter-associated chromatin marks (62, 70, 71). The TFIID complex canonically consists of TBP and 13 TAFs, although cell-type specific variants of TBP and TAFs can form alternate complexes in metazoans. The saddle-shaped TBP binds to the DNA minor groove at the TATA box (if present) and induces a ~90-degree bend in the DNA. The remaining TFIID subunits (TAFs) contact additional core promoter elements (as discussed above), arching from the TBP binding site 30 bp upstream of the TSS to the DPE located 30 bp downstream of the TSS and looping back to contact the Inr at the TSS (29). TFIID functions as a molecular ruler to position RNAPII at a precise position relative to the core promoter elements, and can also regulate TBP activity (72, 73) and

facilitate TFIIF and TFIIE recruitment (29). TFIID subunits are the target of various activating TFs, which can expedite TFIID recruitment to promoters (74, 75).

TFIID binding to the core promoter is stabilized by TFIIA (76), which bind the core promoter upstream of TBP (77-79) and inhibits a repressive domain in TFIID that occludes the TBP DNA-binding surface when TFIID is not bound to promoter DNA (80). Notably, this anti-repression effect can be augmented by certain activating TFs, resulting in a considerable increase in TFIID DNA-binding activity (81).

TFIIB also facilitates TBP binding and DNA bending (82), and is required for RNAPII recruitment to promoters (83, 84). The C-terminal domain of TFIIB interacts with sequences upstream and downstream of the TBP binding site (BREs)(40, 41, 85, 86), orienting the assembly of the initiation complex (87, 88). The N-terminal domain binds to the dock domain of RNAPII to recruit the polymerase (89). TFIIB also functions in establishing the TSS using two distinct functional domains, the B-linker helix and B-reader helix (67). The B-linker helix is positioned above the cleft where DNA opening commences and contributes to the establishment and maintenance of open DNA in the transcriptional bubble (67), and the B-reader helix binds the template strand to position DNA for the initiation of mRNA synthesis (90).

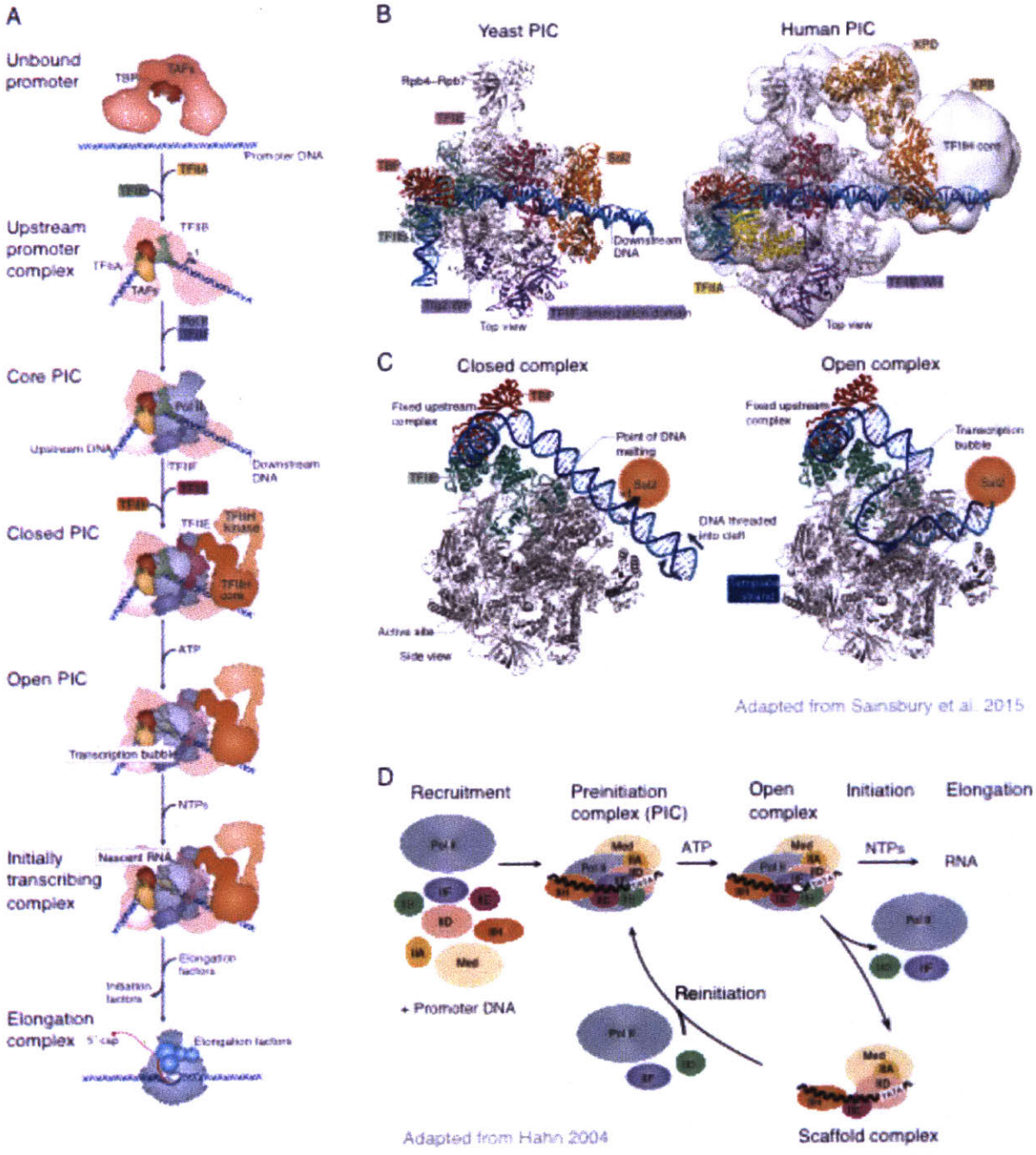
RNAPII binds to the TFIIB-TBP-DNA complex with the assistance of TFIIF to complete the core initiation complex (91). TFIIF interacts with unbound RNAPII to prevent non-specific DNA binding (92), and after binding stabilizes the PIC, particularly TFIIB (93, 94). TFIIF stabilizes DNA downstream of the polymerase active central cleft and inducing a minor opening of the RNAPII clamp domain (66), contributing to the establishment of the TSS (95) and maintenance of the transcriptional bubble (96).

After the assembly of core initiation complex at the promoter, the final two components of the PIC, TFIIE and TFIIH, bind sequentially to form the complete PIC. TFIIE and TFIIH are required for promoter DNA opening. Structural studies showed that TFIIE binds to RNAPII and spans over the polymerase cleft containing loaded DNA, anchoring melted DNA in the polymerase cleft and stabilizing the open promoter complex (66, 97, 98). TFIIE also contains an acidic domain that interacts strongly with TFIIH, facilitating the recruitment of TFIIH to the initiation complex (99) and stimulating its activity (100).

TFIIH possesses DNA-dependent ATPase activity required for transcriptional initiation (101, 102), and functions in promoter opening and escape (103-105) as well as the DNA nucleotide excision repair pathway (106). TFIIH comprises 10 subunits, three of which have catalytic activity, including two ATPases (XPB and XPD) and a kinase (CDK7). The structure of TFIIH-containing PIC has recently become available (66), showing the TFIIH core contacts TFIIE while the XPB domain stretches to contact DNA downstream of the PIC. Transcription initiation requires the complete TFIIH (107), but only XPB catalytic activity is required for promoter opening (105, 108-110), whereas XPD catalytic activity is required for DNA opening in the repair pathway (111). CDK7 phosphorylates the C-terminal domain (CTD) of RNAPII (112, 113), but CTD phosphorylation is not required for transcription initiation *in vitro* (108).

Figure 2. Schematic of Pol II transcription initiation. (A) Depicted is the canonical model for stepwise pre-initiation complex (PIC) assembly from general transcription factors (various colors) and RNAPII (grey) on promoter DNA. The names for the intermediate complexes that form during the initiation-elongation transition are provided to the left of the images. TFIID or its TATA box-binding protein (TBP) subunit binds to promoter DNA, inducing a bend. The TBP-DNA complex is then stabilized by TFIIB and TFIIA, which flank TBP on both sides. The resulting upstream promoter complex is

joined by the RNAPII-TFIIF complex, leading to the formation of the core PIC. Subsequent binding of TFIIE and TFIIH complete the PIC. In the presence of ATP, the DNA is opened (forming the 'transcription bubble') and RNA synthesis commences. Finally, dissociation of initiation factors enables the formation of the Pol II elongation complex, which is associated with transcription elongation factors (blue). TAF, TBP-



associated factor. **(B)** Ribbon models of the yeast PIC (114) and the human PIC (66) (cryo-electron microscopy density is shown as a transparent surface). The TFIIF winged helix (WH) domain is apparently mobile because it is located near the protrusion in the yeast PIC model (114), whereas it is positioned above DNA in the yeast core initiation complex (91). The ATPase subunit of TFIIH (known as XPB in humans and Ssl2 in yeast) is located at downstream DNA. Pol II is depicted in grey. **(C)** Models of closed (left) and open (right) RNAPII-TFIIB-TBP-DNA complexes based on the RNAPII-TFIIB crystal structure (RCSB Protein Data Bank code 3K1F) are shown (67). The location of Ssl2 at downstream DNA and the apparent movement of DNA during promoter opening are indicated by arrows. According to the current model for DNA opening, the ATPase Ssl2 functions as a translocase that threads downstream DNA into the active center cleft of RNAPII while upstream DNA remains fixed. **(D)** The pathway of transcription initiation and reinitiation for RNAPII. Initiation of transcription begins with synthesis of the first phosphodiester bond of RNA. After synthesis of ~30 bases of RNA, RNAPII releases its contacts with the core promoter and the rest of the transcription machinery and enters the stage of transcription elongation. After initiation of transcription by RNAPII, many of the general transcription factors remain behind at the promoter in the scaffold complex. This scaffold complex bypasses the typically slow step of PIC recruitment for subsequent rounds of transcription. Figure modified from (115) and (116).

Initiation and promoter escape

A dramatic conformation change occurs once the complete PIC is bound to the promoter, melting 11-15 bp of DNA surrounding the TSS and positioning the template strand in the active cleft of RNAPII (117). The current model suggests that the TFIIH XPB domain functions as an ATP-dependent translocase, threading the downstream DNA into the active cleft of polymerase while upstream DNA remains fixed, and the resulting strain instigates DNA melting and promoter opening (65).

The opening of the PIC allows polymerase to begin transcribing the first few nucleotides of the nascent transcript. As the nascent RNA grows past 4 nt, it helps to displace TFIIB from the RNAPII exit channel (67, 118, 119). The elongating RNA helps drive conformational changes of the initiating complex, allowing RNAPII to dissociate

from the GTFs and escape the promoter (120). Phosphorylation of the RNAPII CTD by TFIIH is thought to aid this process by destabilizing the interactions between RNAPII and the promoter-bound factors (121). At 8-9 nt, the upstream transcriptional bubble collapses and the nascent transcript reaches the stable full-length RNA-DNA hybrid present in the elongating complex (118, 122). By ~20 nt, the nascent RNA is capped at the 5' end (123), increasing RNA stability and facilitating RNA export and translation (124).

Elongation and pausing

At many genes, the early elongating RNAPII complex pauses after transcribing ~30-50 nt (125, 126). Promoter-proximal pausing was first observed at inactive heat shock response genes (127). The paused RNAPII is rapidly released upon heat shock, resulting in immediate robust gene expression. Pause release thus presents an opportunity for regulation that bypasses the slow steps of PIC recruitment and transcription initiation (128, 129). Subsequent studies found that pause release comprised the rate-limiting step of transcription in other contexts requiring synchronous or rapid changes in gene expression, such as tissue morphogenesis and cell cycle control (130-132).

Promoter-proximal pausing is effected by the negative elongation factor (NELF), which is recruited to RNAPII and the nascent RNA by DRB sensitivity-inducing factor (DSIF) and inhibits transcriptional elongation (133). The promoter-proximal pause site and duration are influenced by the DNA sequence of the gene. RNAPII is prone to transiently pause, especially in A/T rich regions that form less stable DNA-RNA hybrids (134). Instead of inducing novel pauses, NELF and DSIF work by prolonging these

intrinsic, sequence-dependent pauses (133, 135). Several TFs bind NELF, including BRCA1, ERR1 and AP-1, and may help recruit NELF to promoters (136-138).

Paused RNAPII is released by the phosphorylation of NELF and DSIF by positive transcription elongation factor b (P-TEFb), ejecting NELF from the transcriptional complex and relieving elongation inhibition (135, 139-141). P-TEFb also phosphorylates the RNAPII CTD on Ser2 (142), creating a platform for the assembly of factors that travel with the elongating polymerase, including chromatin remodelers and regulators of elongation, RNA processing and termination (142, 143). Like NELF, P-TEFb can be recruited to pause sites by transcriptional activators like NFKB and c-myc (144-146), as well as Bromodomain protein Brd4, which binds acetylated histones (147, 148). Moreover, tethering P-TEFb to the *Drosophila hsp70* gene enhanced expression (149), suggesting that recruitment of P-TEFb may constitute an important function of TFs. Moreover, inhibiting the P-TEFb kinase causes global downregulation of transcription, even at genes with no detectable RNAPII pausing, indicating that a pause-like transition between initiation and elongation is a universal feature of RNAPII-mediated transcription (150-152).

Once RNAPII transitions into productive elongation, additional factors such as FACT and SPT6 regulate its catalytic rate and help it to move through nucleosomes. The histone remodeling complex FACT (facilitates chromatin transcription) travels with RNAPII and functions to disassemble an H2A-H2B dimer from nucleosomes, allowing RNAPII to transcribe through, and redeposit the dimer in its wake (153, 154). SPT6 interacts with histones H3 and H4 and acts as a histone chaperone (155). In addition, SPT6 can increase the elongation rate of RNAPII on naked DNA (156). Recently, global

run-on sequencing (GRO-seq) assays tracking the progress of RNAPII through gene bodies determined that elongation rates between cell types and genes can vary by almost 10-fold, suggesting elongation is another important point of regulation in the transcriptional cycle (157).

Reinitiation using scaffold complex

When RNAPII clears the promoter, Mediator and most of the GTFs remain behind at the core promoter in a scaffold complex (158). This scaffold complex bypasses slow steps of PIC recruitment and allows reinitiation by RNAPII in successive rounds of transcription (Fig. 2D). Eventually, the scaffold complex dissociates from the promoter, preventing further rounds of transcription until the PIC is reassembled. Certain TF activation domains and core promoter elements such as the TATA box have been shown to stabilize the scaffold complex *in vitro* (159-162), resulting in a higher transcriptional output (159, 163).

Single-gene transcriptional dynamics and bursting

The reinitiation cycle results in probabilistic switching of promoters between active and inactive states (164, 165). This process is reflected in transcriptional kinetics: transcription occurs in short, intense bursts comprising clusters of initiation events separated by periods of inactivity (166, 167). Transcriptional output is thus determined by the combination of the amplitude of the burst (i.e. the number of initiation events per burst) and the frequency of the bursts. Studies have shown that burst amplitude is a fixed intrinsic property of the core promoter sequence, which mediates GTF binding (164, 168, 169). Consistent with its role in stabilizing the scaffold complex, the presence of the TATA box motif is associated with larger burst sizes in yeast (168), with a tradeoff

of greater transcriptional noise and cell-to-cell variability (170). In contrast, enhancers appear to modulate transcription by influencing the frequency of transcriptional bursts, without affecting the burst sizes (165, 171).

SECTION III. Transcriptional regulation by enhancers

The first paradigms of gene expression regulation were worked out in bacteria and phage. In these organisms, the core promoter and sequences in the immediate vicinity (50-60 bp) contain all the regulatory elements necessary to determine when a gene is on or off. When scientists began to investigate eukaryotic gene regulation in the 1970s, many believed it would mirror the situation in bacteria, though likely with more complexity in activators and repressors (172). In order to scrutinize an example of a eukaryotic gene promoter, one of the scientists studying gene regulation, William Schaffner, cloned a DNA fragment containing the rabbit β-globin gene and its promoter into two plasmids, one of which contained a partial copy of the SV40 virus genome. Surprisingly, the plasmid containing the SV40 genomic DNA robustly expressed β-globin when introduced into human HeLa cells, while the other plasmid showed no expression. Schaffner and his team narrowed down this effect to a 200 bp segment of the SV40 genome located immediately upstream of the SV40 early promoter, which controls expression of the viral genes required for replication. This segment was shown to enhance the expression of β-globin nearly 100-fold over a distance of 10 kb, longer than the entire SV40 genome (173). Moreover, this "enhancer" worked in either orientation relative to the promoter, and from upstream or downstream of the gene.

A few years after the discovery of the SV40 enhancers, the first cellular enhancer was identified in an intron of the IgH gene (174, 175). Unlike the SV40 enhancer, the

IgH enhancer was cell-type specific, enhancing IgH transcription only in B lymphocytes. This discovery provided one of the first clues that distal cis-regulatory elements might play a central role in controlling differential expression of genes in multicellular organisms across different cell types and environmental conditions. The existence of regulatory DNA located at long genomic distances from their target promoter is a distinctive feature of metazoan genomes, as most regulatory elements in yeast and other simple eukaryotes are located immediately adjacent (100-200 bp) of the promoter (176). In contrast, most metazoan regulatory elements are distal enhancers that regulate their target gene over a distance (177).

The uncoupling of regulatory DNA from promoters allows a gene to be regulated by multiple distal enhancers active in different cell types and conditions, facilitating the complex combinatorial expression patterns necessary to generate a wide array of cellular states. For example, the *Drosophila* segmentation gene *even-skipped* is expressed in seven stripes along the embryo, controlled by five enhancers with distinct spaciotemporal activities (178). Similarly, the Hoxd gene in mice is controlled by dozens of enhancers in two separate topological-associated domains (TADs), with enhancers in the 5' TAD controlling expression in the distal region of the developing limb forming the digits and those in the 3' controlling expression in the proximal region (179). Strikingly, morphological and behavioral complexity in organisms correlates with increased regulatory complexity rather than increased number of genes (180), and deletions or modifications of enhancers underlie many morphological changes between species (5). Thus, the modular organization and distal locations of metazoan enhancers played a key role the development of multiple cell types and the rise of animal diversity.

General properties of enhancers

The ability of the SV40 enhancer to function independently from its distance and orientation to the target gene has turned out to be a general hallmarks of enhancers. They can function at distances up to 2 Mb (181, 182), looping to contact the target promoter (183-185). In addition, enhancers can activate heterologous promoters and function independent of their sequence context, enabling their characterization in reporter constructs. In the genome, they are marked by distinctive chromatin properties, including TF binding, increased DNA accessibility, modifications such as H3K27ac and H3K4me1 on histones surrounding the enhancer, and production of short unstable RNA transcripts, discussed below.

Enhancers comprise clusters of TF binding sites

Functional enhancers contain organized arrays of TF recognition motif sites (Fig. 3A). TFs bound to enhancers in turn recruit non-DNA binding cofactors that regulate transcription through a variety of mechanisms, including modifying and remodeling the local chromatin environment, and recruiting Mediator and the basal transcription machinery. Enhancer activity often requires the binding of multiple TF, including ubiquitously active TFs and lineage- and signal-specific TFs, thus enabling them to serve as information integration hubs. This cooperativity contributes to both DNA binding and post-binding functions (discussed further in section IV).

At the level of binding, cooperativity between TFs can stem from both direct and indirect interactions. TFs can directly interact with each other through protein-protein contacts to increase their mutual DNA affinity and specificity, called "direct cooperativity." Many examples of direct cooperativity have been identified, such as Fos,

which binds DNA either as a homodimer or a heterodimers with Jun (186). Another type of cooperativity stems from the chromatin context of enhancers. Enhancer sequences tend to have a high affinity for nucleosomes, which prevent TFs from accessing the underlying DNA. Binding of multiple TFs in close proximity (e.g. within the same nucleosomal unit) helps to overcome the energy barrier for nucleosome eviction, a process referred to as "indirect cooperativity" or "collaborative competition" (187, 188). Finally, some TFs, called "pioneer factors" possess distinct biochemical properties that allow them to bind to DNA wrapped around nucleosomes. These TFs can bind first to enhancer DNA and facilitate the subsequent binding of additional TFs. Examples of well-characterized pioneer factors include FoxA, which has a domain whose structure mimics the linker histone H1 (189-191), and the *Drosophila* TF Zelda (192, 193).

TFs also function synergistically after DNA binding. An early experiment demonstrating this type of cooperativity showed that TFs from widely divergent organisms (yeast TF GAL4 and mammalian glucocorticoid receptor (GR)) could activate transcription synergistically with the mammalian, despite being unlikely to bind cooperatively, and this synergy persists even when TF binding is saturated (194). The existence of pioneer factors as a special class of TFs suggests there may be other specialized classes of TFs that contribute different molecular functions. Consistent with this idea, studies have shown that combinations of different TFs are more effective at activating transcription than multiple copies of a single TF. Moreover, a recent functional study of ~500 Drosophila TFs with standardized DNA binding showed identified several classes of TFs that were active in different contexts, suggesting they possess qualitatively different regulatory functions (195).

This extensive cooperativity between TFs at the levels of DNA binding and function explain why enhancers tend to contain clusters of multiple TF recognition sites, all required for enhancer activity. The often-complex interplay between different regulators allows for the encoding of intricate patterns of gene expression patterns in enhancers using various combinations of enhancers, as discussed in more detail in section V.

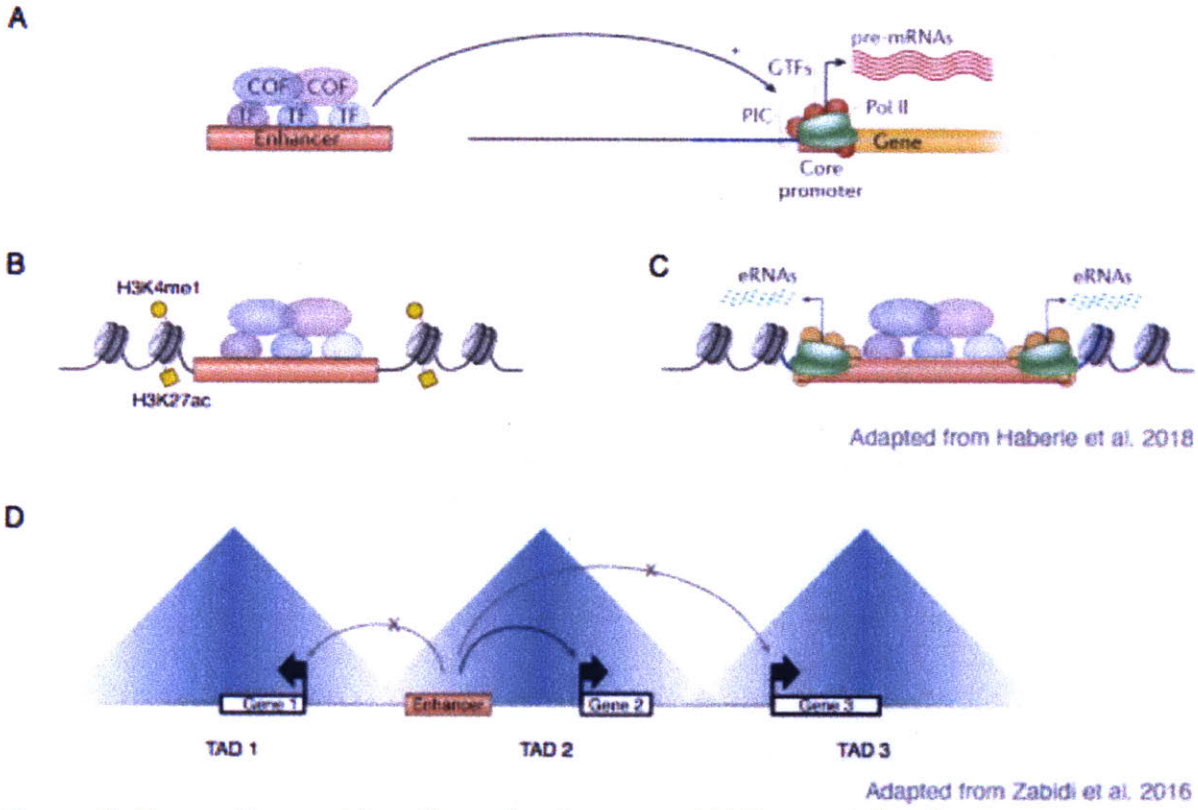


Figure 3. Properties and function of enhancers. (A) Transcription from core promoters is activated by enhancers, which can be located distally and bind sequence-specific transcription factors (TF), which recruit cofactors (COF) that convey the activating cues to the PIC at the core promoter. **(B)** Nucleosomes surrounding active enhancers are marked by characteristic modifications, such as H3K4me1 and H3K27ac. **(C)** Active enhancers exhibit divergent transcription of short, unstable enhancer RNAs (eRNAs) from two separate transcription start sites located at the edges of the NDR where the enhancer resides. **(D)** Enhancer function is typically restricted to activate core promoters within the same TAD, the boundaries of which are enriched in insulator protein binding. Figure modified from (32) and (196).

Chromatin context and enhancer function

As a result of the generally high affinity of enhancer sequences for nucleosomes, inactive enhancers are typically occupied by nucleosomes (197, 198). Nucleosomes block TF binding, maintaining a default "off" state (199-201). Enhancer activation is thought to involve pioneer factors that bind to and displace nucleosomes with the help of ATP-dependent chromatin remodelers (202), and collaborative competition by additional TFs to maintain the open state. Since DNA accessibility is required for TF binding and enhancer activity, it has often been used to predict enhancers. DNA accessibility can be mapped genome-wide using methods that preferentially target accessible chromatin, such as DNaseI treatment or transposon insertion, followed by sequencing (DNase-seq (18) and ATAC-seq (203), respectively). However, enhancer activity is not perfectly correlated with DNA accessibility, as inactive enhancers and promoters as well as other genomic regions such as insulators can also be accessible.

The histones flanking the accessible enhancer region are frequently marked with characteristic post-translational modifications, similar to histones near promoter NDRs (Fig. 3B). Two marks commonly used to identify enhancers are H3K4me1 and H3K27ac, deposited by the enzymes MII3/4 and p300/CBP, respectively. While these and other chromatin modifications are predictive of enhancer activity, their role in enhancer function remains unclear.

Histone modifications could contribute to transcriptional activation in several ways. Modifications of histone residues that contact DNA including H3K27ac have been shown to decrease nucleosome stability (204, 205), potentially assisting the maintenance of the accessible region. Modifications in histone tails can also recruit

"reader" proteins that might enhance the binding of the transcriptional machinery or recruit other chromatin remodelers (61, 206, 207).

Recent data, however, suggests that enhancer-associated histone marks may not be required for enhancer activity. Catalytically inactivating Mll3/4 resulted in a depletion of H3K4me1 but minimally affected gene expression (208, 209), and cells with hyperactive Trr, the fly ortholog of Mll3/4, showed similarly mild expression changes (209). Similarly, replacing the canonical H3 in *Drosophila* with a mutant that could not be acetylated on H3K27 did not cause widespread loss of transcription (64). However, the acetyltransferase activity of p300/CBP is required for transcription activation (210, 211), suggesting p300/CBP might target other histone residues (212) or non-histone proteins such as TFs (213-215). One possible explanation for the seeming dispensibility of enhancer-associated histone marks is that several marks may contribute to activity through parallel mechanisms, such that loss of any one mark has little effect on the overall function. For example, each modification may confer a weak binding affinity for a certain activating complex, but multivalent binding to multiple marks simultaneously results in dramatically increased affinity and specificity (59). It is also possible that some modifications could be functionally irrelevant off-target effects of enzymes. Future studies should help to clarify the role of histone modifications and distinguish causation from correlation (216).

Transcription initiation at enhancers

Widespread transcription initiation at mammalian enhancers has been detected in many cell types (217-219), resulting in the production of short, unstable enhancer RNAs (eRNAs; Fig. 3C). eRNA production correlates with target gene transcription

across different cell types (218) and in inducible systems (220, 221). In some cases, it appears eRNA transcription precedes mRNA production (219, 222), while in others, both transcripts appear at the same time (223, 224).

The functional role of eRNAs has proven difficult to establish, and several hypotheses have been proposed. As discussed in section I, RNAPII can help to displace nucleosomes and establish a NDR, and eRNA might contribute to enhancer activity through this mechanism (225, 226). Alternatively, the nascent eRNA transcripts could play a function role, for example by stabilizing TF binding (227), recruiting cofactors (223, 228, 229), promoting pause release through NELF (221), or fostering enhancer-promoter contacts mediated by cohesion (220, 230, 231). However, none of these RNA-dependent functions appear as yet to occur generally across many eRNA. Most recently, eRNAs have been proposed to mediate the formation of specialized transcriptionally active compartments via phase transition (232, 233), based on RNA-mediated phase transitions seen for nucleolar rRNA and RNAs with tandem repeats (234, 235). Finally, it is possible that enhancer transcription is merely a byproduct of accessible DNA and the presence of activating factors that recruit and stabilize polymerase and the basal machinery (236, 237).

Enhancer targeting to core promoters

A hallmark of enhancers is their ability to regulate promoters over long genomic distances. The dynamic communication of distal enhancers with their target promoters is an area of active research. I first review the current view of the spatial organization of mammalian genomes, and then discuss how enhancers find the correct target promoter within this framework.

Spatial organization of genomes into topological-associated domains

In recent years, the development of high-throughput chromosome conformation capture methods has shed light on the three-dimensional structure and distal interactions of metazoan genomes (238, 239). These methods involve the digestion and re-ligation of chromatin, allowing the read out of DNA sequences in three-dimensional proximity to each other in the nucleus (reviewed in (240)). Applying these methods to Drosophila (241, 242) and mammalian (243, 244) genomes revealed that metazoan chromosomes are organized into self-associating topological-associated domains (TADs). Mammalian TADs comprise a median of 185 kb, containing 1-5 genes and dozens of enhancers (245, 246). Strikingly, TAD boundaries are largely unchanged across cell types and are conserved across a wide swath of evolution, suggesting TADs and their boundaries represent fundamental structural features of chromatin organization (244, 247, 248). TAD boundaries are enriched for insulator-binding proteins such as CTCF and cohesion, and convergent CTCF motif sites are present at >90% of chromatin loop anchors (249, 250). These convergent CTCF-cohesin sites are thought to drive genome segmentation into TADs.

TADs form "regulatory neighborhoods," largely restricting the activity of enhancers and other regulatory elements to genes that fall within the same domain (243, 244). The partitioning of the genome into TADs limits the enhancer search space of possible targets, ensuring high regulatory specificity. Accordingly, disruption of TAD boundary elements can result in the dysregulation of gene expression due to the influence of enhancers in the neighboring TAD (251), and TAD boundary disruptions by

structural variation and aberrant DNA methylation at CTCF binding have been linked to congenital malformations and cancer (252, 253).

Enhancer-promoter communication within TADs

The classic model of enhancer-promoter communication is through the looping of the distal enhancer to the proximal promoter. Several enhancers have been shown to contain promoter-targeting sequences required for enhancer activity at a distance, which likely mediate enhancer-core promoter spatial proximity or contact (254, 255). One of the best characterized examples is the loop formed between the β-globin promoter and the locus control region (LCR), a large distal enhancer that controls temporal switching between embryonic, fetal and adult β-type globin genes. Formation of the LCR-promoter loop is mediated by the TFs GATA1 and LBD1 (256). GATA1 binds to both the LCR and β-globin promoter and forms a complex with the non-DNA binding LBD1, allowing the self-interacting domain of LBD1 to establish and stabilize the loop between the LCR and promoter (257). Additional examples of loops mediated by homotypic TF interactions have been documented, including loops established and stabilized by SP1 in mammals (185, 258) and GAGA in *Drosophila* (259).

Another mode of loop formation involves cell type-specific "genomic organizers" such as SATB1/2 and ARIDA3 that anchor chromatin at specialized A/T rich DNA sequences, organizing it into distinct loops required for proper gene expression (260, 261). SATB1 is the best-characterized tissue specific chromatin organizer. Unstimulated mature T cells do not express SATB1, but its expression is rapidly induced upon T_H2 activation (262). SATB1 forms a transcriptionally active chromatin configuration at the T_H2 cytokine locus, with numerous small loops between the *Il5*, *Il4* and *Il13* promoters

and distal enhancers, and this active chromatin structure is required for cytokine gene expression (262). SATB1-deficient cells show dysregulation of a large number of genes located near anchored loci (261, 263), suggesting it contributes globally to chromatin looping and gene regulation. The cell type-restricted expression of SATB1 and its homologue SATB2 and ARIDA3 may facilitate cell type-specific chromatin organization despite the fixed TAD structure discussed above (263-267).

Despite these well-documented examples of enhancer-promoter loops, emerging evidence suggests that most enhancer-promoter contacts are less stable and more transient. Most active enhancers and promoters are not detected as peaks (i.e. loop anchors), even in high-resolution Hi-C maps (246). Genes within the same TAD are often co-regulated (243, 268), suggesting enhancers move between multiple active promoters in a TAD. Several recent findings support this highly dynamic model. A single enhancer was shown to simultaneously activate transcription from two promoters that were 15 kb apart in a transgenic system in *Drosophila*. Even the supposedly stable LCR-β-globin promoter loops are formed and released with rapid kinetics (171). Furthermore, intronic enhancers can function while RNAPII crosses them during transcription (269, 270), suggesting rigid protein-protein interactions between enhancers and promoters are not required.

Despite the promiscuous contact between enhancers and promoters within a TAD, many genes in the same TAD exhibit distinct expression patterns, raising the question of how cells achieve enhancer-promoter specificity. Promoters may be turned on or off by promoter-specific mechanisms, rendering them impervious to enhancer activity. There is also evidence for different biochemical compatibilities between

enhancers and different classes of promoters. For example, reporter genes with TATA-box containing promoters or with DPE-containing promoters responded differently to the same enhancer landscape (271), suggesting enhancers have specificities or preferences towards certain kinds of promoters. Consistent with this idea, Drosophila core promoters from housekeeping and developmental genes (which generally contain DPE and TATA boxes, respectively) placed into a reporter construct responded to different sets of enhancers (272). These mechanisms may help to effect greater regulatory specificity between enhancers and promoters within the same TAD.

Integrating the activity of multiple enhancers

Most developmental genes that have been studied are regulated by multiple enhancers with specific and overlapping spatiotemporal activities. The number of active enhancers in mammalian cells vastly outnumbers the number of expressed genes, indicating that regulatory interactions between enhancers are common. How the inputs provided by distinct regulatory elements are integrated at promoters remains an area of active investigation.

Modes of regulatory interactions between enhancers

The classic model of enhancers as autonomous and modular predicts that combinations of enhancers will behave additively. Modular and additive behavior has been demonstrated in a number of cases for enhancers controlling the same gene across different tissues and for enhancers in dense clusters, often called superenhancers (273-275).

In other cases, however, the overall regulatory output differs quantitatively or qualitatively from the sum of the individual enhancer activities. For example, detailed

studies of the *endo16* regulatory landscape in sea urchins demonstrated that some enhancers within the region are required for the activity of other enhancers, and observed strong synergic activity among enhancers (276, 277). Synergy between enhancers acting together to amplify expression output has been subsequently observed in many other contexts (for example (278, 279)), and a comparison of enhancer activity in reporter assays and endogenous gene-expression patterns estimated that the activity of more than a third of Drosophila enhancers may be regulated by their native surrounding regulatory context. Notably, antagonistic interactions between enhancers are also commonly observed, often serving to restrict the overall gene expression to the proper spatiotemporal pattern. Examples of cross-repressive interactions include enhancers regulating proper gap-gene-expression patterns in Drosophila (280) and correct developmental and cell-specific expression of Fgf8 in mice (281).

In some cases, interactions between enhancers appear to follow a hierarchical regulatory logic, as was observed for the *endo16* regulatory region. For instance, an enhancer near the PU.1 gene in myeloid cells is required to establish a permissive chromatin environment, allowing for the activation of an adjacent enhancer (282). Repressive hierarchical interactions also occur, often functioning to enforce correct spatiotemporal activity patterns. For example, the Drosophila bithorax complex (BX-C) regulatory region contains nine chromosome domains responsible for expression patterns in different parasegments, each containing several enhancers active only in the appropriate segments (274). Interestingly, only one or two enhancers from each domain, called "initiator elements" showed segment-specific activity when removed from

the native locus, while the rest were active across all segments. In the native context, the initiator elements are sufficient to silence or allow the activity of all the enhancers in the regulatory domain (283, 284).

Enhancer redundancy and competition

Another common mode of non-additive enhancer behavior is enhancer redundancy. Functionally redundant enhancers, often referred to as "shadow enhancers," display overlapping (though not necessarily identical) spatial activity patterns (reviewed in (285)). Shadow enhancers are thought to confer robustness against fluctuating environmental conditions or genetic variability (286, 287). Such robustness is essential for highly deterministic processes such as embryonic development (285).

One mechanism that could underlie this robustness is competition between multiple enhancers for the same promoter, resulting in buffering of the activity of individual enhancers. For example, strongly activated shadow enhancers at the Drosophila *hunchback* and *snai* loci were observed to behave sub-additively, suggesting the regulatory input to the promoter might be saturated. In contrast, weaker activation resulted in additive behavior (288), concordant with a competitive mechanism,. In this model, loss-of-function in one enhancer through mutation or dysregulation would be compensated for by increased promoter contacts of the remaining enhancers, maintaining constant expression levels.

SECTION IV. Transcription factor binding and function

Transcription factors (TFs) lie at the heart of gene regulation, reading the information encoded in regulatory sequences and converting it to biochemical cues that

control gene expression. Some serve as "master regulators" that play a core role in driving cell differentiation and embryonic development, and others effect cellular responses to environmental signals and immune stimuli. In accordance with their essential functions, their protein sequences and regulation are often deeply conserved among metazoans (289, 290), and mutations in TFs and their binding sites cause many human diseases (291). And yet, in other ways TFs appear to be highly dynamic: there is rapid evolutionary turnover of TF binding sites in regulatory sequences (292, 293), and the same TF can regulate different genes in different cell types. Understanding how TFs achieve their tasks of recognizing specific binding sites and controlling transcription is thus a complicated but crucial research goal.

The defining features of TFs are the abilities to (1) bind DNA in a sequence-specific manner and (2) regulate transcription (294). In this section, I first summarize the discovery of sequence-specific TFs and their intrinsic DNA-binding activity. I next review *in vivo* TF binding and additional features that influence genomic binding specificity in cells. Finally, I discuss the current knowledge of TF effector functions and mechanisms of post-binding cooperativity between multiple TFs.

Sequence-specific DNA binding by TFs

Identification of sequence-specific DNA binding factors

Soon after the discovery enhancers in the late 1970s, functional dissection of enhancer and promoter sequences uncovered the first eukaryotic sequence-specific TFs. Early studies used DNaseI footprinting assays (leveraging DNaseI's preferential digestion of naked DNA) to reveal protein-binding sites in enhancer DNA (295), and gel-shift assays (utilizing the difference in gel mobility between unbound and protein-bound

DNA) to assay sequence-specific binding activity of nuclear fractions and isolate the DNA-binding proteins (296). Some of the earliest eukaryotic RNAPII-associated TFs identified using these approaches include the glucocorticoid receptor (GR) and Sp1, observed to bind to sequences in a murine virus (MMTV) and GC box motifs in SV40 respectively (258, 295, 297-299).

Shortly after the first TF were identified, the development of sequence-specific DNA-affinity chromatography enabled the purification of GR and Sp1, as well as numerous other eukaryotic TFs (300-306). The cloning of these TFs soon followed (307-312), opening the door for the functional characterization through perturbations of TF structures and amino acids sequences. New TFs continue to be identified by DNA affinity purification-mass spectrometry (313) and protein microarrays that screen for DNA binding (314), although currently most TFs are identified by sequence homology to known TF domains (discussed below).

TF DNA binding domains

One striking feature of TF function is their modular nature. Most TFs comprise separable DNA binding domains (DBDs) linked to one or more activation or repression domain. This modularity was first demonstrated by fusing the yeast TF GAL4 to the DBD of the bacterial TF LexA. The fused protein could activate transcription for genes with LexA binding sites, showing that the GAL4 activation domain can function autonomously when recruited to DNA by a heterologous DBD (315). Comparisons of the DBDs of different TFs revealed frequent similarities in the amino acid composition and structure. Most eukaryotic DBDs fall into a small set of common structural families, such as basic helix-loop-helix (bHLH), basic leucine zipper (bZIP), C2H2-zinc finger

(ZF), and nuclear hormone receptor (NR). TFs in each DBD family recognize DNA using common structural motifs. Currently there are ~100 known DBD types (catalogued in (316-318)), which are used to scan protein sequences to identify novel DNA-binding proteins and classify TFs (319). All but a handful of functionally-characterized metazoan TFs contain a known DBD (320).

Structures of DBD bound to DNA have been solved for many mammalian TFs (321), shedding light on how TFs recognize specific DNA sequences (322). Two different modes of protein-DNA recognition contribute to TF binding specificity. The first mode, called 'base readout,' is governed by physical interactions such as hydrogen bonds and hydrophobic contacts between the amino acid side chains of the DBD and the base pairs that are recognized (323). Many of these interactions happen in the major groove of DNA, although minor groove and sugar/phosphate backbone interactions also occur (319). TFs can also preferentially bind specific DNA sequences through 'shape readout' (Fig. 4B). This mode of protein-DNA recognition involves the readout of sequence-specific structural features such as DNA bending and unwinding (324-326). Most TF DBD use an interplay of these two modes to recognize their cognate binding sites (327, 328).

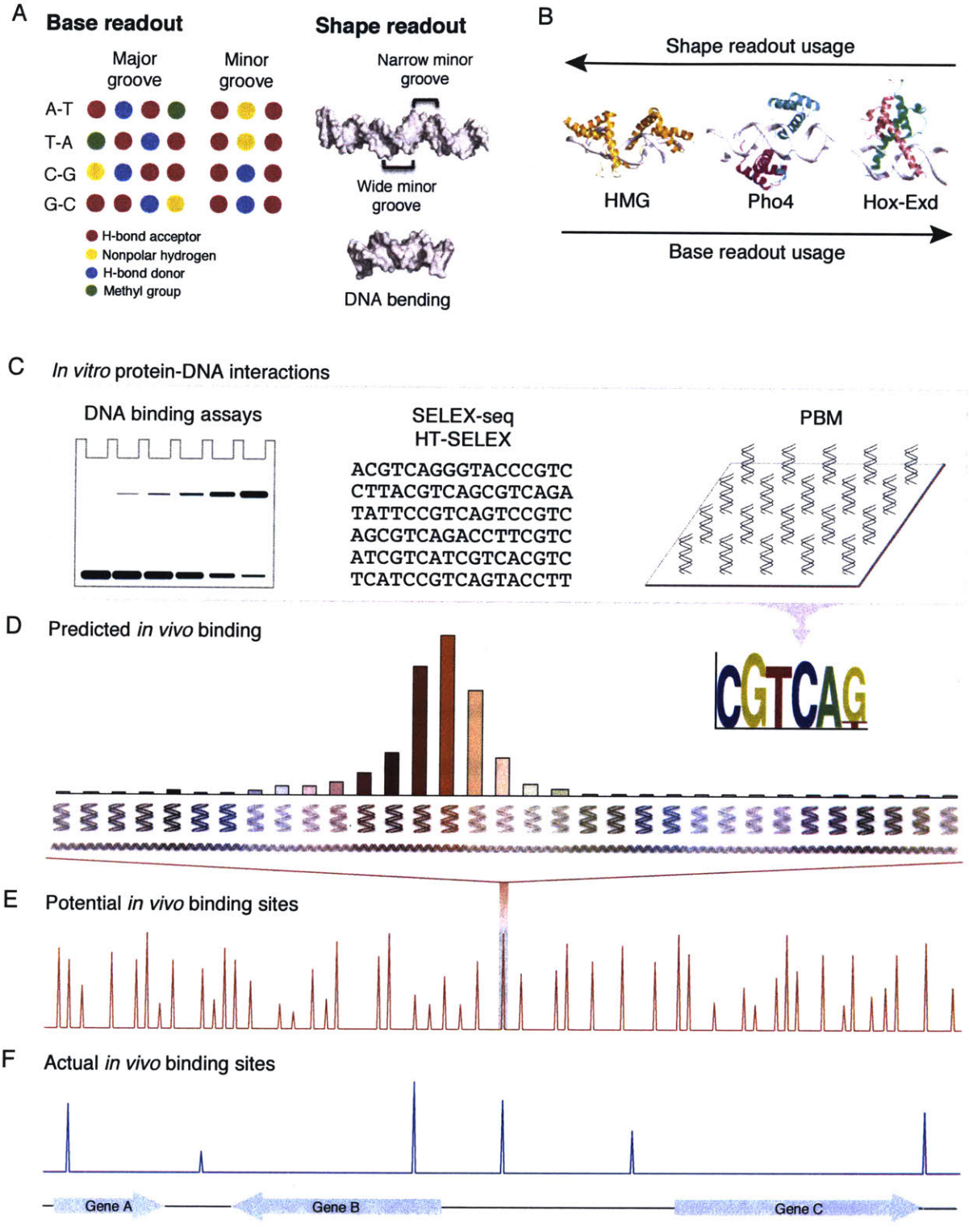
Determining TF DNA-binding specificity in vitro

TFs can have 1,000-fold higher affinity for their cognate binding sequences relative to other sequences. TFs generally have low-affinity non-sequence specific interactions with the DNA backbone (~10^{-3}-10^{-5} M), allowing them to slide along the genome scanning for target sites. At target sites, high affinity sequence specific interactions (~10^{-8}-10^{-12} M) immobilize them to allow for transcriptional regulation.

TF binding specificities are generally summarized by motif models, which are fit using known TF binding sites and predict the relative binding specificity to any new DNA sequence. Early TF motif models were fit by aligning binding sites identified in DNaseI footprinting or gel shift experiments, or analyzing the promoter regions of co-regulated genes (329-331). The development of microarray- and sequencing-based methods in the last 15 years made it possible to generate quantitative affinity measurements of TFs for a large number of sequences (Fig. 4C). For example, protein binding microarrays (PBM) detect the binding of a fluorescently-tagged TF to a microarray spotted with tens of thousands of short DNA sequences (332, 333), and SELEX-based methods enriches TF-bound sequences from a pool of randomized sequences through multiple rounds of selection (334, 335).

Motif models are often represented by a position weight matrices (PWMs) specifying the relative preference of the TF for each base at each position in the binding site (336). PWMs are intuitive and easily visualized as sequence logos. However, they have several limitations: (1) they can only capture base readout by TFs, (2) they assume the contribution of each position within a binding site is independent, and (3) they assume each TF has only one mode of binding DNA. These assumptions do not always hold (337-340), and more intricate models that account for these complexities have been developed (337, 341, 342).

The deep datasets produced by high-throughput methods have dramatically improved our ability to characterize TF binding specificities. It is now possible to determine the binding affinity landscapes of TFs to all possible specifically bound sequences, enabling the characterization of low-affinity binding sites that are not well



Adapted from Slattery et al. 2014

Figure 4. *In vitro* and *in vivo* TF–DNA interactions. (A) Base readout describes direct interactions between amino acids and the functional groups of the bases. Whereas the pattern of hydrogen bond acceptors (red) and donors (blue), heterocyclic hydrogen atoms (yellow) and the hydrophobic methyl group (green) is base pair-specific in the major groove, the pattern is degenerate in the minor groove. **(B)** Shape readout includes any form of structural readout based on global and local DNA shape features, including conformational flexibility and shape-dependent electrostatic potential. The DNA target of the IFNβ enhanceosome (PDB ID 1t2k; top) varies in minor groove shape. The human papillomavirus E2 protein binds to a DNA binding site (PDB ID 1jj4; bottom) with intrinsic curvature. **(C)** Standard and high-throughput in vitro DNA-binding assays provide a motif or model representing TF DNA-binding preferences. **(D)** Genomic DNA sequences matching an in vitro-derived motif represent potential TFBSs. **(E)** Potential in vivo binding sites determined from a TF in vitro- derived motif far outnumber the actual number of in vivo binding sites as measured by ChIP-seq. In general, <5% of potential binding sites are identified as being bound in vivo (**F**). In addition, in vivo binding strength does not always correlate with motif strength, and not all in vivo binding sites contain the expected motif. Non-DNA variables, such as nucleosomes and cofactor interactions, explain part of the difference between predicted and actual binding. Figure modified from (343).

captured by traditional PWMs (344, 345). Many TFs utilize such low-affinity binding sites, such as developmental TFs that use low-affinity sites to read out spatiotemporal TF gradients (346, 347). Furthermore, the datasets generated with high-throughput methods have revealed many TFs bind to DNA through multiple modes that involve different physical configurations of the TF-DNA complex, resulting in multiple distinct motifs (340). Although the biological significance of these alternate motifs is unknown, many are bound by TFs *in vivo*, and therefore likely to play a functional role (338, 348). A key challenge now is to develop computational methods to leverage the raw data to accurately predict quantitative relative affinities and motif models that properly represent TF binding to all possible sequences.

TF binding *in vivo*

ChIP-seq has revolutionized the study of TF binding *in vivo* by enabling genome-wide mapping of TF occupancies. These maps immediately revealed that very few metazoan TFs occupy the majority of their motif sites in cellular contexts (349, 350). (The only clear exception is CTCF, which has a long 14-bp motif and occupies the majority of its ~14,000 sites in the human genome (351, 352).) Furthermore, ChIP-seq studies measuring TF binding across multiple cell types revealed extensive cell-specific DNA binding for nearly all TFs, indicating that the cellular context influences TF binding *in vivo* (353-357). Several cellular features that contribute to the additional specificity in genomic DNA binding are discussed below.

TF interactions with chromatin

Genomic DNA is wrapped around histone octamers *in vivo*, forming nucleosomes that facilitate DNA packaging in the nucleus. In order to bind to nucleosomal DNA, TFs must compete with nucleosomes or interact with nucleosomal DNA to access their binding sites. TF occupancy is highly correlated with nucleosome-depleted regions (NDRs) across multiple species and cell types (201, 358-360), indicating TF binding and nucleosome occupancy are generally mutually exclusive. Systematic *in vitro* characterization of the binding of 220 TFs to free and nucleosomal DNA showed the majority of TFs are unable bind nucleosomal DNA (361), indicating that DNA accessibility helps to govern which sites TFs can bind. Consistent with this model, several nuclear receptors have been shown to require remodeling by the chromatin-remodeling complex BAF to facilitate their binding (362, 363). DNA accessibility is

particularly important for the occupancy of low-affinity binding sites, which are important in development (346, 347) and tend to show more cell-specific binding (364-366).

Collaborative TF binding

TFs typically bind to regions that contain clusters of different TF binding sites (367, 368). Multiple studies across species and cell type have shown that depletion of a TF or mutations in a motif site affects not only the binding of the cognate TF, but also the binding of neighboring TFs (192, 193, 353, 369, 370). This combinatorial binding requirement confers greater specificity and allows for different subsets of motif sites to be bound in different cell types, depending on the set of TFs present in a particular cell. Cooperative binding can occur through several different mechanisms, including indirect mechanisms that impact the local chromatin state and direct TF-TF interactions that stabilize binding.

In order to overcome the barrier presented by nucleosomes, TFs can collaborate to aid each other in binding DNA. While the affinity of a nucleosome for DNA is greater than that of an individual TFs, multiple TFs binding to clustered sites can collaborate with each other to compete with and displace nucleosomes (referred to as "collaborative competition") (188, 371, 372). This kind of cooperativity does not require direct protein-protein interactions between TFs, and thus allows for a flexible arrangement of binding sites within enhancers.

A specialized class of TFs called pioneer factors are defined by their ability to bind within closed chromatin, promote DNA accessibly and enable the subsequent binding of additional TFs. For example, the canonical pioneer factor FOXA1 can bind its target sites on nucleosomal DNA and contact core histone proteins through a C-terminal

domain that facilitates chromatin opening (373, 374). Another pioneer factor, PU.1, can expand the linker regions between nucleosomes and promote nucleosome depletion and chromatin modifications (353, 370, 375). Consistent with a pioneering role in enhancer opening, FOXA1 and PU.1 have been shown to occupy binding sites in enhancers before other TFs (376, 377), and their expression is required for induction of liver (378) and hematopoietic (376) gene expression programs respectively.

In recent years, the ability of many TFs to bind nucleosomal DNA and affect DNA accessibility were systematically analyzed in high-throughput studies using a modified SELEX protocol for nucleosomal DNA (361) and unique TF DNaseI footprints in accessible regions during a developmental timecourse (379). These studies revealed a wide range of nucleosome-binding abilities and numerous distinct modes of TF-nucleosome interaction, suggesting it may be more accurate to view TF pioneering activity as a continuum. A more nuanced view is consistent with recent observations that even the binding of pioneer TFs can depend on partner TFs. For example, both FOXA1 and PU.1 depend on CEBPα to bind many genomic sites (380, 381). In general, sites where pioneer TFs can act autonomously tend to have higher-affinity motif matches and show binding across multiple cell types, whereas sites requiring collaborative binding tend to be cell-type specific and are associated with lower-affinity motif matches (366, 370, 382). Thus, the relative contributions of traditional pioneer factors and collaborative binding can vary in different contexts, generating distinct regulatory strategies at different target sites.

Cooperative binding through TF-TF interactions

Direct protein-protein interactions can also give rise to cooperative TF binding. The prototypic example of this type of cooperativity is the binding of the *E. coli* lambda repressor dimers to neighboring sites on DNA (383). Due to physical interactions between lambda dimers, binding of a lambda dimer to the first site increases the affinity of a second dimer. This cooperative binding leads to a nonlinear, switch-like relationship between TF concentration and binding site occupancy, and is often used in developmental contexts to drive cell fate decisions and generate sharp boundaries of gene expression (384, 385). Unlike collaborative binding interactions, which do not depend on precise spacing between TFs and allow for flexible binding site arrangements, direct TF-TF interactions rely on defined protein interaction interfaces and impose spatial constraints on the binding sites.

The TF-TF interactions underlying direct cooperativity span a range of affinities. Some TFs cannot bind DNA as monomers, and form homodimers or heterodimers in solution with increased DNA affinity and specificity. For example, the AP-1 factor Fos interacts with other AP-1 factors via a leucine zipper and can bind to DNA as either a homodimer or a Fos/Jun heterodimer (186). Similarly, nuclear receptors such as PPARγ, RAR and LXR form heterodimers with RXR that bind to composite motifs with direct or inverted half-sites separated by a specific number of nucleotides. Cooperativity can also arise, however, from TFs that interact only weakly in the absence of DNA, but form specific contacts when bound to neighboring sites on DNA. A systematic binding analysis of 9,400 human TF pairs using SELEX identified a large number of DNA-dependent TF-TF interactions, indicating such interactions are common and can occur

between TFs in different structural families (386). Interestingly, many interacting pairs of TFs bind composite motifs different than the two individual motifs; furthermore, interactions between TFs and cofactors can also change the DNA sequence specificity of a TF (387). These interaction-induced differences in sequence specificity may contribute to *in vivo* context-dependent TF binding.

Binding cooperativity can also be mediated by DNA shape, if binding of one TF influences the local shape in a manner that promotes the binding of a second TF. For example, the highly ordered IFNβ "enhanceosome" contains overlapping, spatially constrained binding sites for eight TFs, which form a platform to recruit cofactors. Structural analysis revealed a paucity of protein-protein interactions, however, and suggests that cooperativity occurs via binding-induced conformational changes in DNA structure and stabilization of the complex by multivalent interactions with cofactors (388). Similarly, the TFs AML1 and RUNX1 bind cooperatively to DNA and form a binary complex. They do not interact with each other directly, but induce a bend in the DNA between them that likely mediates the cooperativity (389).

Long-range TF binding interdependencies

The mechanisms discussed above all involve interactions between TFs bound to proximal sites. However, multiple studies of TF binding variation and allele-specific DNA binding have observed that some differential TF binding events cannot be explained by nucleotide changes in motif site of the studied TF or proximal TF motif sites (390, 391). Furthermore, a study in human LCLs found that PU.1 binding variation often correlated with variation in chromatin marks such as H3K4me1 or H3K27ac not only locally but also over extended distances, suggesting that TF binding can be affected not only by

the proximal region but also by long-range mechanisms. These regions with a high level of coordination between molecular phenotypes, recently termed "variable chromatin modules" (VCMs), tend to be contained within TADs, suggesting they may correspond to regions interacting in three-dimensional space (390). Similar long-range coordination has been observed in studies of allele-specific TF binding and chromatin structure, indicating the effect is largely driven by genetic variation instead of *trans* effects (392, 393). While the mechanisms underlying these long-range effects remain unclear, the existing evidence supports a model in which the alteration of the binding of one or a few TFs determines the VCM's activity state and affects all the molecular phenotypes in the VCM, including the binding of other TFs.

TF effector functions in transcriptional activation

In addition to a DBD, nearly all TFs have one or more "effector" domains. These domains are sufficient to modulate transcription when recruited to promoter or enhancer regions, even when separated from their native DBD (315, 394, 395). Unlike the well-characterized DBDs, comparatively little is understood about the structure of effector domains and how they interact with other transcriptional regulatory proteins. In general, these domains interact with chromatin remodeling enzymes and general transcription factors, helping to recruit them to enhancers and promoters. Here, I discuss common features of effector domains, and give examples of effector domains that influence transcriptional regulation by (1) interacting with the basal transcriptional machinery and general cofactors to promote multiple steps in the transcriptional cycle, (2) interacting with other TFs to facilitate cooperative binding and (3) recruiting histone remodelers and

chromatin modifying enzymes. Finally, I discuss the recently proposed phase separation model for transcriptional control.

Effector domains interact with the general transcriptional machinery

By and large, general transcription factors (GTFs) involved in transcription initiation and elongation cannot stably and specifically bind DNA on their own. Instead, activating TFs play a pivotal role in the transcriptional by enhancing recruitment and stabilization of the basal transcriptional machinery and GTFs at core promoters (Fig. 5A). TF effector domains that contact GTFs and cofactors and stimulate transcriptional activation are called activation domains (ADs). ADs from many different TFs have been shown to make direct contacts with GTFs, including TBP, numerous TAFs, TFIIA, TFIIB, TFIIF and TFIIH, and various components of Mediator (396-404).

In contrast to the structured and conserved DBD, the ADs of most TFs are low-complexity amino acid sequences with low conservation and little homology with other ADs. These intrinsically disordered regions (IDRs) are often classified by their amino acid composition as acidic (e.g. E2F1 and p53), glutamine-rich (e.g. Oct1, Oct2, and Sp1), proline-rich (e.g. AP-2 and CTF/NF1), or serine/threonine-rich (e.g. ATF2). Each of these classes of ADs has been observed to interact with components of the basal transcriptional machinery (405-407). For instance, the glutamine-rich AD of Sp1 interacts with $TAF_{II}130$, a subunit of TFIID, and mutations in the AD that inhibit the binding of $TAF_{II}130$ result in reduced expression (398, 399, 408). Notably, several acidic and proline-rich ADs were shown not to interact with $TAF_{II}130$, suggesting different ADs might recruit different elements of the transcriptional machinery (400, 408, 409).

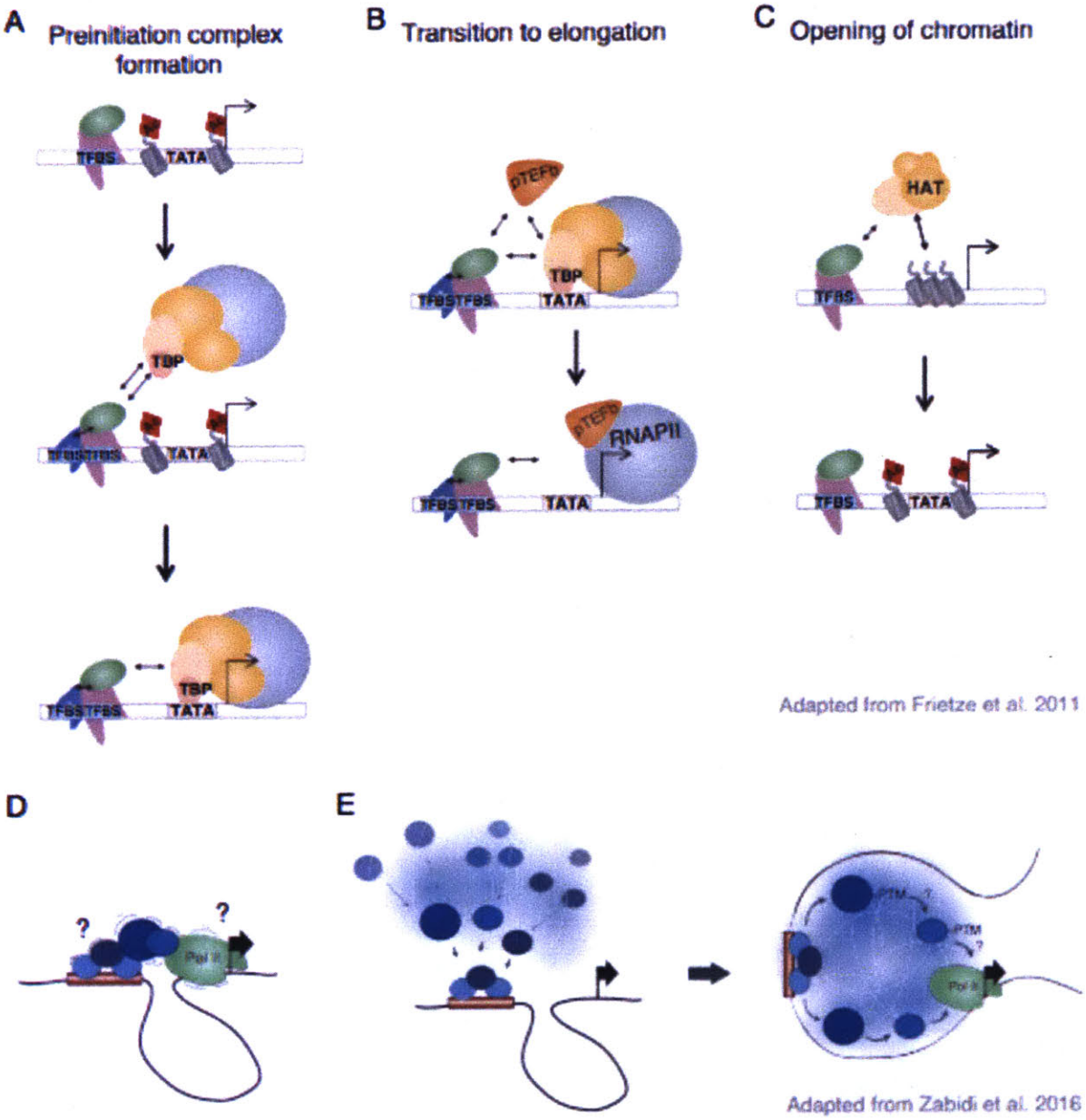


Figure 5. Functions and properties of TF effector domains. Functions of TF effector domains include **(A)** facilitating the assembly of the preinitiation complex, which can involve inter- actions between different TFs and general factors such as TBP; **(B)** promoting of the transition to productive elongation, which involves interaction between DNA-bound proteins and enzymes such as the pTEFb kinase; and **(C)** creating open chromatin, which involves interactions between DNA-bound proteins and histone modifying enzymes (e.g. a HAT which can acetylate histones). Although in this

schematic the TFs are shown binding to TFBS proximal to the TSS, TFs can also perform these functions when bound to enhancers looped to a target promoter. (**D**) Static model of transcription regulation in which defined protein complexes formed by static protein–protein interactions at enhancers exert their function on core promoters. This model is incompatible with some observations such as the simultaneous activation of two core promoters by a single enhancer. **(E)** Transcription regulation might alternatively occur via a phase-separated activating microenvironment in which enhancers and core promoters recruit trans factors to create a high concentration of regulatory proteins that dynamically interact with each other, and enable regulatory communication through post-transcriptional modifications (PTMs, top) or via dynamic protein–protein interactions and recruitment (bottom). Figure modified from (410) and (196).

The initial mechanistic models of AD function focused on the enriched residues (e.g. the "acidic blob" model). However, directed mutagenesis of acidic and glutamine-rich ADs showed that acidic or glycine residues often have only a small effect on activity. Instead, activity is most dependent on bulky hydrophobic residues interspersed in the acidic or glutamine residues (398, 411-413). Based on this observation, a model was proposed for AD function stemming from short linear motifs of hydrophobic residues embed in disordered regions, which are frequently presented to binding partners on one side of an α-helix (414). A recent high-throughput mutagenesis scan tested this model by investigating the function of sequence features of the Gcn4 acidic AD, including acidity, hydrophobicity, intrinsic disorder, and short linear motifs. Activity and inducibility was found to depend on key aromatic residues positioned in a disordered acidic peptide scaffold, supporting a unified model of AD function in which acidic residues and conformational disorder keep key hydrophobic exposed to solvent and binding partners (415).

In several cases, IDRs have been shown to fold into a more stable structure upon interacting with their binding partners. For example, the unstructured N-terminal

region of p53 containing its acidic AD folds into an α-helix when it binds to either the PIC or the p53 attenuator Mdm2 (416-418). A similar phenomenon was observed for IDR in nuclear receptors, which folds into an α-helix in response to protein-protein interactions and some ligands (419-421). This "induced fit" could explain how ADs interact with numerous different binding partners and avoid restricted relationships between certain general factors and ADs.

ADs can also be functionally grouped by the transcriptional step they stimulate. While ADs are most commonly thought to act in PIC assembly and transcription initiation, a number of studies have shown that ADs can affect multiple steps of the transcriptional cycle, including promoter clearance, elongation rate, and pause release (Fig. 5B; 144, 146, 422-425). For example, the c-Myc AD recruits P-TEFb, which phosphorylates the RNAPII CTD and releases paused polymerase. An elegant biochemical study showed that TFs with activation domains that worked on different steps (i.e. initiation vs. elongation) acted synergistically, while those that worked on the same step did not (423), supporting the notion of functionally distinct classes of ADs.

TF-TF interactions through effector domains

As discussed above, effector domains mediate cooperative physical interactions between TFs that increase their affinity and specificity for DNA sequences. Some TF-TF interactions occurs in solution, generating homo- or heterodimers that bind DNA as a complex. Some of the heterodimeric TF complexes, such as E2F and its obligate partner DP, dimerize using similar heterodimerization domains (426), while others use two dissimilar domains, like RXR with PPARγ and VDR (427). Other pairs of effector domains with weaker affinities mediate protein-protein interactions after the two TFs are

brought into proximity on DNA. Oct4 and Sox2, for instance, interact weakly in solution, but this low affinity interaction is crucial for the cooperative binding of Oct4 and Sox2 to neighboring sites on DNA in the regulation of ESC-specific gene expression (428-431). In some cases, binding of one TF to DNA can induce an allosteric change in its effector domain that increases its affinity for another TF. For example, different motif variants cause different allosteric changes in the TR receptor, modulating its DNA-dependent dimerization with RXR (432, 433).

Recruitment of chromatin-modifying enzymes by effector domains

The majority of eukaryotic TFs interact with cofactors (coactivators or corepressors), large multi-subunit complexes or multidomain proteins that regulate transcription through several mechanisms (Fig. 5C; 434). Most cofactors include domains involved in chromatin recognition, nucleosome remodeling and/or covalent modification of chromatin. Coactivators modify chromatin to displace or evict nucleosomes or modify histones to loosen their contacts with DNA, while corepressors effect a more closed chromatin confirmation and deposit repressive histone modifications. Some also covalently modify other proteins, such as TFs and RNAPII.

One of the most common coactivators recruited by TFs is p300/CBP, a histone acetyltransferase (HAT) found at a large number of enhancers that also a bromodomain to recognizes acetylated lysines on histones as well as domains that mediate interactions with RNAPII and a multitude of GTFs and sequence-specific TFs (435, 436). Because of its numerous interaction partners, p300/CBP can act as a hub to integrate signals from multiple TFs. The IFNβ enhanceosome exemplifies such signal integration through synergistic coactivator recruitment, with multiple bound TFs forming

a platform to recruit GCN5/KAT2A and p300/CBP HATs through multivalent interactions, leading to histone acetylation and subsequent nucleosome remodeling by SWI/SNF (388).

ATP-dependent chromatin remodeling complexes such as SWI/SNF, ISWI and WINAC are also directly recruited by many TF effector domains, including nuclear receptors, AP-1, C/EBP and c-Myc (437-440). Dynamic changes in chromatin remodeling complex subunit composition during development have been shown to result gene- and cell-specific targeting by TFs (441-443). As the IFNβ enhanceosome illustrates, histone acetylation has been shown to stabilize SWI/SNF binding to nucleosomes through several bromodomain-containing subunits (444), assisting in stabilizing the complex at activator-bound sites. Recruitment of these chromatin remodeling complexes leads to nucleosome rearrangement and additional chromatin modifications, allowing Mediator, GTFs and RNAPII to bind to the region.

TF effector domains can also recruit repressive chromatin modifying complexes. For example, the nuclear receptors TR, RAR and VDR can repress transcription in the absence of their ligands by recruiting corepressor complexes containing NCoR and SMRT (445). Corepressors establish repressive chromatin using histone methyltransferases (HMTs) and demethylases, histone deacetylases (HDACs), and nucleosome remodeling complexes such as NURD (446). The most common repressive effector domain is the KRAB domain, which is found in over ~350 different human C2H2-ZF TFs and interacts with the corepressor TRIM28/KAP1 (447). KAP1 functions as a scaffold to recruit HDACs, HP1 and SETDB1, resulting in the deposition of the repressive chromatin mark H3K9me3 and a closed chromatin state (448). Thus, KRAB

effector domain in zinc finger TFs link the KAP1 corepressor complex to specific genomic sites to silence gene expression.

A phase separation model of transcriptional control

One of the remarkable aspects of TF effector function is the ability of hundreds of different TFs to interact with the same small set of cofactors. For example, over 50 TFs have been reported to physically interact with components of Mediator (449). Furthermore, ADs that share little sequence homology can substitute for each other functionally across multiple enhancer contexts (450). This complementation argues against a classic lock-and-key protein interaction mechanism. Recently, structural and functional studies of the AD of the yeast TF GCN4 showed it binds to a Mediator subunit in multiple conformations and orientations using only hydrophobic interactions, forming a "fuzzy complex" (414, 451, 452). Such multi-conformational IDR-IDR interactions are a hallmark of phase-separated biomolecular condensates such as nucleoli, Cajal bodies, and nuclear speckles (sites of rRNA biogenesis, snRNP assembly, and mRNA splicing factor storage respectively) (453-456).

In view of this observation, Sharp and colleagues recently proposed that TF ADs and cofactors could participate in the IDR-mediated formation of phase-separated condensates that control transcription (Fig. 5D,E; 232). Consistent with this phase-separation model, RNAPII and many cofactors also possess low-complexity IDRs (457). This model would explain several key features of superenhancers (large enhancers occupied by a high density of TFs that often control genes with prominent roles in cell identity), such as their ability to drive unusually high levels of transcription and their vulnerability to perturbations. It would also explain the recent observation that a single

enhancer can drive transcriptional bursts at multiple promoters simultaneously (165). In the past few months, Sharp and colleagues showed that several TFs can indeed form phase-separated condensates with Mediator, providing functional support for the phase-separation model (458). Furthermore, Mediator and the coactivator BRD4 form phase-separated condensates at superenhancers *in vivo* that control the expression key cell-identity genes (459). Overall, these results suggest that the phase-separating properties of IDR in TF ADs and cofactors facilitate the formation of condensates that compartmentalize and concentrate the transcriptional machinery at specific genes.

SECTION V. Enhancer grammar and combinatorial TF function

Transcriptional regulation is mediated by distinct combinations of TFs at each enhancer, enabling a few hundred TFs to generate a diverse array of gene expression patterns. Enhancer sequences encode the regulatory instructions for the proper expression of 20,000 genes across thousands of cell types and conditions. The lexicon of these instructions is composed of recognition motifs encoding binding sites for specific TFs. The number, affinity, spacing, orientation, and order of motif sites can all influence the expression output of an enhancers. The rules governing the organization of binding sites in regulatory sequences, referred to as regulatory grammar, are complex and remain an area of active research. In this section, I review our current understanding of the grammatical rules of enhancers, and discuss some of the potential mechanisms underlying these rules.

Properties of enhancer grammar

The general rules of the organization of TF binding sites in regulatory sequences have been studied using various approaches, including comparative analyses of

orthologous or functionally similar enhancers, computational modeling and synthetic enhancer reporter assays. Analyses of native enhancers can uncover regulatory features associated with transcriptional outcomes, whereas synthetic enhancer assays enable systematic manipulations of different properties of regulatory sequences to elucidate the causal parameter. A number of loci have been laboriously characterized through extensive regulatory sequence perturbations, providing qualitative insights into *cis*-regulatory logic and enhancer grammar. In recent years, a surge of new methods linking reporter assays with high-throughput sequencing enabled quantitative measurements of the activity of tens of thousands of synthetically designed enhancer sequences in a single experiment (23, 460-463). These methods revolutionized the study of regulatory sequences, making it possible to evaluate the generality of grammatical rules and develop quantitative models.

Motif site number and affinity

Multiple copies of the same TF motif are often found in proximal promoters and enhancers (464). This pattern of enrichment for homotypic clusters is conserved between vertebrates and invertebrates (464), suggesting that motif copy number can serve to fine-tune gene expression. Consistent with this notion, studies in both yeast and human cells have shown that for many TFs expression output initially increases with the addition of TF binding sites and then saturates above a specific number of sites (465, 466). In yeast, the relationship between TF multiplicity and expression was quantitively examined for 2 TFs across all possible combinations of 1-7 sites across two promoter contexts. For both TFs, the relationship accurately followed a logistic function; however, the number of binding sites and expression level at which saturation occurred

varied by TF and promoter context (465). Such saturation was also observed in smaller studies of individual TFs (467, 468), but it is unclear at present whether it occurs at the level of binding or activation of transcription.

Both the number of binding sites for a specific TF and the affinities of the binding sites can serve as a sensor for the concentration or activity of the TF. For example, the yeast TF Pho4 is activated during phosphate starvation and induces many phosphate response genes. In intermediate phosphate conditions, Pho4 is partially activated and induces target genes that have accessible high-affinity binding sites in their promoters, while expression of target genes with low-affinity binding sites remains at low levels (469). In addition to modulating the gene expression for different levels of stimuli, suboptimal binding sites also play an important role in promoting cell-type specific enhancer activity (470-472). For example, the neural plate-specific Otx-a enhancer in *Ciona* contains suboptimal GATA and ETS binding sites that differ from the consensus motifs by a couple of nucleotides. A recent study showed that changing these binding sites to perfect consensus motifs results in robust but ectopic enhancer activity patterns (470).

Organization of motifs in regulatory sequences

TFs bound to enhancers mediate gene expression through interactions with other TFs, cofactors, chromatin remodelers, and the transcriptional machinery. Furthermore, some enhancers are first poised for activation by pioneer factors, which facilitate the binding of the rest of the TFs. The underlying organization of motifs in the sequence can affect these complex and hierarchical interactions, resulting in grammatical rules

relating the spacing and arrangement of the binding sites and the enhancer activity. The prevalence and rigidity of these rules remains an open question.

On one end of the spectrum, some enhancers require precise motif arrangement, orientation and spacing for function. The canonical example of such an "enhanceosome" is the viral-induced mammalian IFNβ enhancer, which requires the binding of eight TFs to overlapping binding sites to synergistically activate transcription (388). A recent mutational scan found that substitutions or insertions at nearly every position within the enhanceosome were deleterious to enhancer activity (460), consistent with its high evolutionary conservation and structural studies that showed contacts with virtually every nucleotide (473). This strong synergy results in switch-like behavior, integrating inputs from multiple signaling pathways.

On the other end of the spectrum are enhancers described by the "billboard mode." These enhancers show little to no constraint on the spacing or organization of their motif sites, which instead function as largely independent modules, resulting in additive behavior. Examples of this type of enhancer include several zebrafish notochord enhancers, which show no fixed position (474), distance or relative organization of key TF motif sites, and *even-skipped* enhancers, which appear to lack any conserved arrangement of binding sites between *Drosophila* and sepsids (475, 476). Similarly, a recent reporter screen tested the activity of synthetic enhancers containing various combinations of 12 liver-specific TFs, and found evidence for flexibility of binding site order (466). A large-scale comparison of enhancer function across species generally supported a flexible organization model, finding conservation of function of orthologous enhancers between species, despite extensive reorganization

of the TF binding sites (477). However, a SELEX analysis of pairwise TF binding affinities found that the majority of TF-TF interactions have novel consensus motifs different from the motifs bound by either TF in isolation (386), raising the possibility that studies looking for constrained spacing between the individual TF binding motifs are missing real cooperative binding events.

Between these two extremes, other enhancers require particular organization of a subset of the constituent motifs. For example, the extensively-characterized *sparkling* enhancer (*spa*) of the *Drosophila* dPax2 gene has some linked sites with apparently conserved arrangements and spacing, while other parts of the enhancer have undergone extensive reorganization (255). Consistent with this evolutionary observation, systematic mutation of *spa* showed that many reorganizations of the binding sites or deletions of spacer regions resulted in aberrant expression patterns, but some were permitted, suggesting a mix of the enhanceosome and billboard models in the enhancer's architecture (478). Similar partial organizational constraints were observed in *Drosophila* immune and neurogenic ectoderm enhancers (4, 479). Notably, constraints on spacing have been shown in some cases to facilitate cooperative interactions between TFs binding to adjacent sites, but in others to prevent such an interaction, which would result in ectopic expression (472, 480). Thus, many enhancers may not fit a straightforward billboard or enhanceosome model.

Another intermediate between the billboard and enhanceosome model was recently proposed, called a "TF collective." In this model, TFs function synergistically, as in the enhanceosome model, but do not require a strict TF arrangement or composition (i.e. a subset of activating TFs are sufficient) (481). The basis for this model was the

genome-wide binding patterns for TFs that regulate heart development in *Drosophila*, which bind in a collective all-or-nothing fashion but with flexible binding site grammar (481, 482). Such a model would fit would also fit, for example, a kinetic control situation, where TFs act synergistically by affecting different steps in the transcriptional cycle (16).

Overall, it appears that importance of motif grammar in enhancers falls on a continuum between the two extremes. Some situations are better suited additive, billboard-type interactions, such as regulating graded gene expression for homeostatic responses. Others require the more switch-like behavior of enhanceosomes, such as binary cell-fate decisions. Due to this variability in enhancer architectures, grammatical rules have been hard to characterize even with the ability to extensively perturb and characterize various enhancers in different contexts. Identifying such rules may require greater understanding of the mechanisms underlying observed organizational constraints, shedding light on their generality or context of action.

Mechanisms of combinatorial TF function

TF cooperativity after DNA binding

In addition to TF cooperativity during binding, a great deal of evidence suggests that different combinations of TFs are more effective in activating transcription, suggesting that different TFs might have distinct functional roles and contribute different regulatory cues. For example, the study of synthetic enhancers with liver-specific TF motifs mentioned above found that enhancers with heterotypic combinations of TF binding sites drove higher expression than those with homotypic clusters of binding sites (466). Furthermore, a recent complementation study of 474 *Drosophila* TFs showed that different TFs were active in different contexts, suggesting there are classes

of TFs with distinct regulatory functions. The clearest distinction was found between TFs that preferentially activated housekeeping gene core promoters versus developmental gene core promoters, although 15 sub-classes were identified (195).

Interestingly, both studies also found that some TFs can activate transcription on their own while others can only activate transcription in combination with specific other TFs. Consistent with this observation, some TFs such as P53 are sufficient to activate transcription at many sites on their own (483), while other enhancers require multiple TFs that execute complimentary functions. For example, the LDLR enhancer is regulated cooperatively SP1 and SREBP, which recruit Mediator and TFIID respectively (484). Overall, these results suggest a possible regulatory code at the level of function, with complimentary regulatory functions contributed by different TFs.

References

1. Bhagwat AS & Vakoc CR (2015) Targeting Transcription Factors in Cancer. *Trends Cancer* 1(1):53-65.

2. Farh KK, *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518(7539):337-343.

3. Herz HM, Hu D, & Shilatifard A (2014) Enhancer malfunction in cancer. *Mol Cell* 53(6):859-866.

4. Crocker J, Tamori Y, & Erives A (2008) Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* 6(11):e263.

5. Frankel N, *et al.* (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* 474(7353):598-603.

6. Peter IS & Davidson EH (2011) Evolution of Gene Regulatory Networks Controlling Body Plan Development. *Cell* 144(6):970-985.

7. Prud'homme B, Gompel N, & Carroll SB (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A* 104 Suppl 1:8605-8612.

8. Shlyueva D, Stampfel G, & Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nature Publishing Group* 15(4):272-286.

9. Spitz F & Furlong EEM (2012) Transcription factors: from enhancer binding to developmental control. *Nature Publishing Group* 13(9):613-626.

10. Levine M & Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424(6945):147-151.

11. Lemon B & Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 14(20):2551-2569.

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

13. Struhl K (1991) Mechanisms for diversity in gene expression patterns. *Neuron* 7(2):177-181.

14. Weingarten-Gabbay S & Segal E (2014) The grammar of transcriptional regulation. *Hum Genet* 133(6):701-711.

15. Levo M & Segal E (2014) In pursuit of design principles of regulatory sequences. *Nature Publishing Group* 15(7):453-468.

16. Scholes C, DePace AH, & Sanchez A (2017) Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle. *Cell Syst* 4(1):97-108 e109.

17. Wasserman WW & Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5(4):276-287.

18. Boyle AP, *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132(2):311-322.

19. Consortium EP, *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799-816.

20. Heintzman ND, *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* 39(3):311-318.

21. Johnson DS, Mortazavi A, Myers RM, & Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830):1497-1502.

22. Robertson G, *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4(8):651-657.

23. Arnold CD, *et al.* (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339(6123):1074-1077.

24. Melnikov A, *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*:1-9.

25. Patwardhan RP, *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology* 30(3):265-270.

26. Kadonaga JT (2012) Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* 1(1):40-51.

27. Smale ST & Baltimore D (1989) The "initiator" as a transcription control element. *Cell* 57(1):103-113.

28. Burley SK & Roeder RG (1996) Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem* 65:769-799.

29. Louder RK, *et al.* (2016) Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature* 531(7596):604-609.

30. Patikoglou GA, *et al.* (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev* 13(24):3217-3230.

31. Lenhard B, Sandelin A, & Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* 13(4):233-245.

32. Haberle V & Stark A (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* 19(10):621-637.

33. FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, & Vinson C (2006) Comparative genomics of Drosophila and human core promoters. *Genome Biol* 7(7):R53.

34. Ohler U, Liao GC, Niemann H, & Rubin GM (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol* 3(12):RESEARCH0087.

35. Carninci P, *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38(6):626-635.

36. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, & Myers RM (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res* 16(1):1-10.

37. Kim TH, *et al.* (2005) A high-resolution map of active promoters in the human genome. *Nature* 436(7052):876-880.

38. Burke TW & Kadonaga JT (1996) Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* 10(6):711-724.

39. Kutach AK & Kadonaga JT (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol Cell Biol* 20(13):4754-4764.

40. Deng W & Roberts SG (2005) A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev* 19(20):2418-2423.

41. Lagrange T, Kapanidis AN, Tang H, Reinberg D, & Ebright RH (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* 12(1):34-44.

42. Lim CY, *et al.* (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* 18(13):1606-1617.

43. Lewis BA, Kim TK, & Orkin SH (2000) A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc Natl Acad Sci U S A* 97(13):7172-7177.

44. Lee DH, *et al.* (2005) Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol Cell Biol* 25(21):9674-9686.

45. Consortium F, *et al.* (2014) A promoter-level mammalian expression atlas. *Nature* 507(7493):462-470.

46. Hoskins RA, *et al.* (2011) Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* 21(2):182-192.

47. Vo Ngoc L, Cassidy CJ, Huang CY, Duttke SH, & Kadonaga JT (2017) The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev* 31(1):6-11.

48. Rach EA, Yuan HY, Majoros WH, Tomancak P, & Ohler U (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol* 10(7):R73.

49. Jiang C & Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* 10(3):161-172.

50. Mavrich TN, *et al.* (2008) Nucleosome organization in the Drosophila genome. *Nature* 453(7193):358-362.

51. Yuan GC, *et al.* (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* 309(5734):626-630.

52. Bonisch C & Hake SB (2012) Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res* 40(21):10719-10741.

53. Weber CM, Ramachandran S, & Henikoff S (2014) Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell* 53(5):819-830.

54. Rach EA, *et al.* (2011) Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet* 7(1):e1001274.

55. Cairns BR (2009) The logic of chromatin architecture and remodelling at promoters. *Nature* 461(7261):193-198.

56. Tirosh I & Barkai N (2008) Two strategies for gene regulation by promoter nucleosomes. *Genome Res* 18(7):1084-1091.

57. Tropberger P, *et al.* (2013) Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell* 152(4):859-872.

58. Tessarz P & Kouzarides T (2014) Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol* 15(11):703-708.

59. Ruthenburg AJ, Li H, Patel DJ, & Allis CD (2007) Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol* 8(12):983-994.

60. Berndsen CE, *et al.* (2007) Nucleosome recognition by the Piccolo NuA4 histone acetyltransferase complex. *Biochemistry* 46(8):2091-2099.

61. Li H, *et al.* (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* 442(7098):91-95.

62. Jacobson RH, Ladurner AG, King DS, & Tjian R (2000) Structure and function of a human TAFII250 double bromodomain module. *Science* 288(5470):1422-1425.

63. Hodl M & Basler K (2012) Transcription in the absence of histone H3.2 and H3K4 methylation. *Curr Biol* 22(23):2253-2257.

64. Pengelly AR, Copur O, Jackle H, Herzig A, & Muller J (2013) A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science* 339(6120):698-699.

65. Sainsbury S, Bernecky C, & Cramer P (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* 16(3):129-143.

66. He Y, Fang J, Taatjes DJ, & Nogales E (2013) Structural visualization of key steps in human transcription initiation. *Nature* 495(7442):481-486.

67. Kostrewa D, *et al.* (2009) RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* 462(7271):323-330.

68. Miller G & Hahn S (2006) A DNA-tethered cleavage probe reveals the path for promoter DNA in the yeast preinitiation complex. *Nat Struct Mol Biol* 13(7):603-610.

69. Chen HT & Hahn S (2004) Mapping the location of TFIIB within the RNA polymerase II transcription preinitiation complex: a model for the structure of the PIC. *Cell* 119(2):169-180.

70. Juven-Gershon T, Hsu JY, & Kadonaga JT (2006) Perspectives on the RNA polymerase II core promoter. *Biochem Soc Trans* 34(Pt 6):1047-1050.

71. Juven-Gershon T & Kadonaga JT (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* 339(2):225-229.

72. Kokubo T, Yamashita S, Horikoshi M, Roeder RG, & Nakatani Y (1994) Interaction between the N-terminal domain of the 230-kDa subunit and the TATA box-binding subunit of TFIID negatively regulates TATA-box binding. *Proc Natl Acad Sci U S A* 91(9):3520-3524.

73. Kotani T, *et al.* (1998) Identification of highly conserved amino-terminal segments of dTAFII230 and yTAFII145 that are functionally interchangeable for inhibiting TBP-DNA interactions in vitro and in promoting yeast cell growth in vivo. *J Biol Chem* 273(48):32254-32264.

74. Liu WL, *et al.* (2009) Structures of three distinct activator-TFIID complexes. *Genes Dev* 23(13):1510-1521.

75. Papai G, *et al.* (2010) TFIIA and the transactivator Rap1 cooperate to commit TFIID for transcription initiation. *Nature* 465(7300):956-960.

76. Imbalzano AN, Zaret KS, & Kingston RE (1994) Transcription factor (TF) IIB and TFIIA can independently increase the affinity of the TATA-binding protein for DNA. *J Biol Chem* 269(11):8280-8286.

77. Tan S, Hunziker Y, Sargent DF, & Richmond TJ (1996) Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature* 381(6578):127-151.

78. Geiger JH, Hahn S, Lee S, & Sigler PB (1996) Crystal structure of the yeast TFIIA/TBP/DNA complex. *Science* 272(5263):830-836.

79. Bleichenbacher M, Tan S, & Richmond TJ (2003) Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. *J Mol Biol* 332(4):783-793.

80. Kokubo T, Swanson MJ, Nishikawa JI, Hinnebusch AG, & Nakatani Y (1998) The yeast TAF145 inhibitory domain and TFIIA competitively bind to TATA-binding protein. *Mol Cell Biol* 18(2):1003-1012.

81. Chi T, Lieberman P, Ellwood K, & Carey M (1995) A general mechanism for transcriptional synergy by eukaryotic activators. *Nature* 377(6546):254-257.

82. Zhao X & Herr W (2002) A regulated two-step mechanism of TBP binding to DNA: a solvent-exposed surface of TBP inhibits TATA box recognition. *Cell* 108(5):615-627.

83. Sawadogo M & Roeder RG (1985) Factors involved in specific transcription by human RNA polymerase II: analysis by a rapid and quantitative in vitro assay. *Proc Natl Acad Sci U S A* 82(13):4394-4398.

84. Pinto I, Ware DE, & Hampsey M (1992) The yeast SUA7 gene encodes a homolog of human transcription factor TFIIB and is required for normal start site selection in vivo. *Cell* 68(5):977-988.

85. Barberis A, Muller CW, Harrison SC, & Ptashne M (1993) Delineation of two functional regions of transcription factor TFIIB. *Proc Natl Acad Sci U S A* 90(12):5628-5632.

86. Buratowski S & Zhou H (1993) Functional domains of transcription factor TFIIB. *Proc Natl Acad Sci U S A* 90(12):5633-5637.

87. Tsai FT & Sigler PB (2000) Structural basis of preinitiation complex assembly on human pol II promoters. *EMBO J* 19(1):25-36.

88. Littlefield O, Korkhin Y, & Sigler PB (1999) The structural basis for the oriented assembly of a TBP/TFB/promoter complex. *Proc Natl Acad Sci U S A* 96(24):13668-13673.

89. Bushnell DA, Westover KD, Davis RE, & Kornberg RD (2004) Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. *Science* 303(5660):983-988.

90. Sainsbury S, Niesser J, & Cramer P (2013) Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature* 493(7432):437-440.

91. Muhlbacher W, *et al.* (2014) Conserved architecture of the core RNA polymerase II initiation complex. *Nat Commun* 5:4310.

92. Conaway RC, Garrett KP, Hanley JP, & Conaway JW (1991) Mechanism of promoter selection by RNA polymerase II: mammalian transcription factors alpha and beta gamma promote entry of polymerase into the preinitiation complex. *Proc Natl Acad Sci U S A* 88(14):6205-6209.

93. Fishburn J & Hahn S (2012) Architecture of the yeast RNA polymerase II open complex and regulation of activity by TFIIF. *Mol Cell Biol* 32(1):12-25.

94. Cabart P, Ujvari A, Pal M, & Luse DS (2011) Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes. *Proc Natl Acad Sci U S A* 108(38):15786-15791.

95. Ghazy MA, Brodie SA, Ammerman ML, Ziegler LM, & Ponticelli AS (2004) Amino acid substitutions in yeast TFIIF confer upstream shifts in transcription initiation and altered interaction with RNA polymerase II. *Mol Cell Biol* 24(24):10975-10985.

96. Pan G & Greenblatt J (1994) Initiation of transcription by RNA polymerase II is limited by melting of the promoter DNA in the region immediately upstream of the initiation site. *J Biol Chem* 269(48):30101-30104.

97. Chen HT, Warfield L, & Hahn S (2007) The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat Struct Mol Biol* 14(8):696-703.

98. Flores O, Maldonado E, & Reinberg D (1989) Factors involved in specific transcription by mammalian RNA polymerase II. Factors IIE and IIF independently interact with RNA polymerase II. *J Biol Chem* 264(15):8913-8921.

99. Ohkuma Y, Hashimoto S, Wang CK, Horikoshi M, & Roeder RG (1995) Analysis of the role of TFIIE in basal transcription and TFIIH-mediated carboxy-terminal domain phosphorylation through structure-function studies of TFIIE-alpha. *Mol Cell Biol* 15(9):4856-4866.

100. Ohkuma Y & Roeder RG (1994) Regulation of TFIIH ATPase and kinase activities by TFIIE during active initiation complex formation. *Nature* 368(6467):160-163.

101. Conaway RC & Conaway JW (1989) An RNA polymerase II transcription factor has an associated DNA-dependent ATPase (dATPase) activity strongly stimulated by the TATA region of promoters. *Proc Natl Acad Sci U S A* 86(19):7356-7360.

102. Conaway RC & Conaway JW (1993) General initiation factors for RNA polymerase II. *Annual review of biochemistry* 62(1):161-190.

103. Moreland RJ, *et al.* (1999) A role for the TFIIH XPB DNA helicase in promoter escape by RNA polymerase II. *J Biol Chem* 274(32):22127-22130.

104. Goodrich JA & Tjian R (1994) Transcription factors IIE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II. *Cell* 77(1):145-156.

105. Holstege FC, van der Vliet PC, & Timmers HT (1996) Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. *EMBO J* 15(7):1666-1677.

106. Schaeffer L, *et al.* (1993) DNA repair helicase: a component of BTF2 (TFIIH) basic transcription factor. *Science* 260(5104):58-63.

107. Svejstrup JQ, *et al.* (1995) Different forms of TFIIH for transcription and DNA repair: holo-TFIIH and a nucleotide excision repairosome. *Cell* 80(1):21-28.

108. Makela TP, *et al.* (1995) A kinase-deficient transcription factor TFIIH is functional in basal and activated transcription. *Proc Natl Acad Sci U S A* 92(11):5174-5178.

109. Tirode F, Busso D, Coin F, & Egly JM (1999) Reconstitution of the transcription factor TFIIH: assignment of functions for the three enzymatic subunits, XPB, XPD, and cdk7. *Mol Cell* 3(1):87-95.

110. Guzman E & Lis JT (1999) Transcription factor TFIIH is required for promoter melting in vivo. *Mol Cell Biol* 19(8):5652-5658.

111. Coin F, Oksenych V, & Egly JM (2007) Distinct roles for the XPB/p52 and XPD/p44 subcomplexes of TFIIH in damaged DNA opening during nucleotide excision repair. *Mol Cell* 26(2):245-256.

112. Feaver WJ, Gileadi O, Li Y, & Kornberg RD (1991) CTD kinase associated with yeast RNA polymerase II initiation factor b. *Cell* 67(6):1223-1230.

113. Serizawa H, *et al.* (1995) Association of Cdk-activating kinase subunits with transcription factor TFIIH. *Nature* 374(6519):280-282.

114. Grunberg S, Warfield L, & Hahn S (2012) Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nat Struct Mol Biol* 19(8):788-796.

115. Sainsbury S, Bernecky C, & Cramer P (2015) Structural basis of transcription initiation by RNA polymerase II. *Nature reviews. Molecular cell biology* 16(3):129-143.

116. Hahn S (2004) Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structural & Molecular Biology* 11(5):394-403.

117. Wang W, Carey M, & Gralla JD (1992) Polymerase II promoter activation: closed complex formation and ATP-driven start site opening. *Science* 255(5043):450-453.

118. Pal M, Ponticelli AS, & Luse DS (2005) The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. *Mol Cell* 19(1):101-110.

119. Liu X, Bushnell DA, Wang D, Calero G, & Kornberg RD (2010) Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science* 327(5962):206-209.

120. Kugel JF & Goodrich JA (2002) Translocation after synthesis of a four-nucleotide RNA commits RNA polymerase II to promoter escape. *Mol Cell Biol* 22(3):762-773.

121. Myers LC, *et al.* (1998) The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. *Genes Dev* 12(1):45-54.

122. Holstege FC, Fiedler U, & Timmers HT (1997) Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO J* 16(24):7468-7480.

123. Rasmussen EB & Lis JT (1993) In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A* 90(17):7923-7927.

124. Proudfoot NJ, Furger A, & Dye MJ (2002) Integrating mRNA processing with transcription. *Cell* 108(4):501-512.

125. Guenther MG, Levine SS, Boyer LA, Jaenisch R, & Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130(1):77-88.

126. Zeitlinger J, *et al.* (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* 39(12):1512-1516.

127. Rougvie AE & Lis JT (1988) The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* 54(6):795-804.

128. Shao W & Zeitlinger J (2017) Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet* 49(7):1045-1051.

129. Gressel S, *et al.* (2017) CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6.

130. Boettiger AN & Levine M (2009) Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science* 325(5939):471-473.

131. Lagha M, *et al.* (2013) Paused Pol II coordinates tissue morphogenesis in the Drosophila embryo. *Cell* 153(5):976-987.

132. Williams LH, *et al.* (2015) Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. *Mol Cell* 58(2):311-322.

133. Yamaguchi Y, *et al.* (1999) NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* 97(1):41-51.

134. Landick R (2006) The regulatory roles and mechanism of transcriptional pausing. *Biochem Soc Trans* 34(Pt 6):1062-1066.

135. Renner DB, Yamaguchi Y, Wada T, Handa H, & Price DH (2001) A highly purified RNA polymerase II elongation control system. *J Biol Chem* 276(45):42601-42609.

136. Zhong H, *et al.* (2004) COBRA1 inhibits AP-1 transcriptional activity in transfected cells. *Biochem Biophys Res Commun* 325(2):568-573.

137. Aiyar SE, *et al.* (2004) Attenuation of estrogen receptor alpha-mediated transcription through estrogen-stimulated recruitment of a negative elongation factor. *Genes Dev* 18(17):2134-2146.

138. Ye Q, *et al.* (2001) BRCA1-induced large-scale chromatin unfolding and allele-specific effects of cancer-predisposing mutations. *J Cell Biol* 155(6):911-921.

139. Kim JB & Sharp PA (2001) Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain independently of cyclin-dependent kinase-activating kinase. *J Biol Chem* 276(15):12317-12323.

140. Ivanov D, Kwak YT, Guo J, & Gaynor RB (2000) Domains in the SPT5 protein that modulate its transcriptional regulatory properties. *Mol Cell Biol* 20(9):2970-2983.

141. Cheng B & Price DH (2007) Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *J Biol Chem* 282(30):21901-21912.

142. Peterlin BM & Price DH (2006) Controlling the elongation phase of transcription with P-TEFb. *Mol Cell* 23(3):297-305.

143. Li B, Carey M, & Workman JL (2007) The role of chromatin during transcription. *Cell* 128(4):707-719.

144. Rahl PB, *et al.* (2010) c-Myc regulates transcriptional pause release. *Cell* 141(3):432-445.

145. Barboric M, Nissen RM, Kanazawa S, Jabrane-Ferrat N, & Peterlin BM (2001) NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. *Mol Cell* 8(2):327-337.

146. Eberhardy SR & Farnham PJ (2002) Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter. *J Biol Chem* 277(42):40156-40162.

147. Yang Z, *et al.* (2005) Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell* 19(4):535-545.

148. Jang MK, *et al.* (2005) The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell* 19(4):523-534.

149. Lis JT, Mason P, Peng J, Price DH, & Werner J (2000) P-TEFb kinase recruitment and function at heat shock loci. *Genes Dev* 14(7):792-803.

150. Henriques T, *et al.* (2013) Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol Cell* 52(4):517-528.

151. Jonkers I, Kwak H, & Lis JT (2014) Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* 3:e02407.

152. Chen FX, *et al.* (2017) PAF1 regulation of promoter-proximal pause release via enhancer activation. *Science* 357(6357):1294-1298.

153. Belotserkovskaya R, *et al.* (2003) FACT facilitates transcription-dependent nucleosome alteration. *Science* 301(5636):1090-1093.

154. Xin H, *et al.* (2009) yFACT induces global accessibility of nucleosomal DNA without H2A-H2B displacement. *Mol Cell* 35(3):365-376.

155. Bortvin A & Winston F (1996) Evidence that Spt6p controls chromatin structure by a direct interaction with histones. *Science* 272(5267):1473-1476.

156. Ardehali MB, *et al.* (2009) Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J* 28(8):1067-1077.

157. Danko CG, *et al.* (2013) Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* 50(2):212-222.

158. Yudkovsky N, Ranish JA, & Hahn S (2000) A transcription reinitiation intermediate that is stabilized by activator. *Nature* 408(6809):225-229.

159. Dieci G & Sentenac A (2003) Detours and shortcuts to transcription reinitiation. *Trends Biochem Sci* 28(4):202-209.

160. Struhl K (1996) Chromatin structure and RNA polymerase II connection: implications for transcription. *Cell* 84(2):179-182.

161. Iyer V & Struhl K (1995) Mechanism of differential utilization of the his3 TR and TC TATA elements. *Mol Cell Biol* 15(12):7059-7066.

162. Yean D & Gralla J (1997) Transcription reinitiation rate: a special role for the TATA box. *Mol Cell Biol* 17(7):3809-3816.

163. Darzacq X & Singer RH (2008) The dynamic range of transcription. *Mol Cell* 30(5):545-546.

164. Tantale K, *et al.* (2016) A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nat Commun* 7:12248.

165. Fukaya T, Lim B, & Levine M (2016) Enhancer Control of Transcriptional Bursting. *Cell* 166(2):358-368.

166. Chubb JR, Trcek T, Shenoy SM, & Singer RH (2006) Transcriptional pulsing of a developmental gene. *Curr Biol* 16(10):1018-1025.

167. Raj A, Peskin CS, Tranchina D, Vargas DY, & Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4(10):e309.

168. Hornung G, *et al.* (2012) Noise-mean relationship in mutated promoters. *Genome Res* 22(12):2409-2417.

169. Blake WJ, *et al.* (2006) Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell* 24(6):853-865.

170. Tirosh I, Weinberger A, Carmi M, & Barkai N (2006) A genetic signature of interspecies variations in gene expression. *Nat Genet* 38(7):830-834.

171. Bartman CR, Hsu SC, Hsiung CC, Raj A, & Blobel GA (2016) Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Mol Cell* 62(2):237-247.

172. Schaffner W (2015) Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. *Biol Chem* 396(4):311-327.

173. Banerji J, Rusconi S, & Schaffner W (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27(2 Pt 1):299-308.

174. Banerji J, Olson L, & Schaffner W (1983) A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* 33(3):729-740.

175. Gillies SD, Morrison SL, Oi VT, & Tonegawa S (1983) A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* 33(3):717-728.

176. Struhl K, Kadosh D, Keaveney M, Kuras L, & Moqtaderi Z (1998) Activation and repression mechanisms in yeast. *Cold Spring Harb Symp Quant Biol* 63:413-421.

177. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.

178. Levine M (2010) Transcriptional enhancers in animal development and evolution. *Curr Biol* 20(17):R754-763.

179. de Laat W & Duboule D (2013) Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* 502(7472):499-506.

180. Levine M & Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424(6945):147-151.

181. Lettice LA, *et al.* (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12(14):1725-1735.

182. Fulco CP, *et al.* (2016) Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* 354(6313):769-773.

183. Martin DI, Fiering S, & Groudine M (1996) Regulation of beta-globin gene expression: straightening out the locus. *Curr Opin Genet Dev* 6(4):488-495.

184. Calhoun VC & Levine M (2003) Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. *Proc Natl Acad Sci U S A* 100(17):9878-9883.

185. Su W, Jackson S, Tjian R, & Echols H (1991) DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes Dev* 5(5):820-826.

186. Kerppola TK & Curran T (1991) Fos-Jun heterodimers and Jun homodimers bend DNA in opposite orientations: implications for transcription factor cooperativity. *Cell* 66(2):317-326.

187. Miller JA & Widom J (2003) Collaborative competition mechanism for gene activation in vivo. *Mol Cell Biol* 23(5):1623-1632.

188. Mirny LA (2010) Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A* 107(52):22534-22539.

189. Barozzi I, *et al.* (2014) Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Molecular Cell* 54(5):844-857.

190. Chaya D, Hayamizu T, Bustin M, & Zaret KS (2001) Transcription factor FoxA (HNF3) on a nucleosome at an enhancer complex in liver chromatin. *Journal of Biological Chemistry* 276(48):44385-44389.

191. Iwafuchi-Doi M, *et al.* (2016) The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Molecular Cell* 62(1):79-91.

192. Sun Y, *et al.* (2015) Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation. *Genome Res* 25(11):1703-1714.

193. Schulz KN, *et al.* (2015) Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo. *Genome Res* 25(11):1715-1726.

194. Kakidani H & Ptashne M (1988) GAL4 activates gene expression in mammalian cells. *Cell* 52(2):161-167.

195. Stampfel G, *et al.* (2015) Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature*.

196. Zabidi MA & Stark A (2016) Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors. *Trends in genetics : TIG* 32(12):801-814.

197. Lidor Nili E, *et al.* (2010) p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome Res* 20(10):1361-1368.

198. Charoensawan V, Janga SC, Bulyk ML, Babu MM, & Teichmann SA (2012) DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes. *Mol Cell* 47(2):183-192.

199. Svaren J, Klebanow E, Sealy L, & Chalkley R (1994) Analysis of the competition between nucleosome formation and transcription factor binding. *J Biol Chem* 269(12):9335-9344.

200. Walter PP, Owen-Hughes TA, Cote J, & Workman JL (1995) Stimulation of transcription factor binding and histone displacement by nucleosome assembly protein 1 and nucleoplasmin requires disruption of the histone octamer. *Mol Cell Biol* 15(11):6178-6187.

201. Liu X, Lee CK, Granek JA, Clarke ND, & Lieb JD (2006) Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res* 16(12):1517-1528.

202. Hargreaves DC & Crabtree GR (2011) ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell Res* 21(3):396-420.

203. Buenrostro JD, Wu B, Chang HY, & Greenleaf WJ (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* 109:21 29 21-29.

204. Garcia-Ramirez M, Rocchini C, & Ausio J (1995) Modulation of chromatin folding by histone acetylation. *J Biol Chem* 270(30):17923-17928.

205. Tse C, Sera T, Wolffe AP, & Hansen JC (1998) Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. *Mol Cell Biol* 18(8):4629-4638.

206. Vermeulen M, *et al.* (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* 131(1):58-69.

207. Jenuwein T & Allis CD (2001) Translating the histone code. *Science* 293(5532):1074-1080.

208. Dorighi KM, *et al.* (2017) Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* 66(4):568-576 e564.

209. Rickels R, *et al.* (2017) Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet* 49(11):1647-1653.

210. Hilton IB, *et al.* (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* 33(5):510-517.

211. Boija A, *et al.* (2017) CBP Regulates Recruitment and Release of Promoter-Proximal RNA Polymerase II. *Mol Cell* 68(3):491-503 e495.

212. Pradeepa MM, *et al.* (2016) Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat Genet* 48(6):681-686.

213. Edmunds JW & Mahadevan LC (2004) MAP kinases as structural adaptors and enzymatic activators in transcription complexes. *J Cell Sci* 117(Pt 17):3715-3723.

214. Kim MY, Woo EM, Chong YT, Homenko DR, & Kraus WL (2006) Acetylation of estrogen receptor alpha by p300 at lysines 266 and 268 enhances the deoxyribonucleic acid binding and transactivation activities of the receptor. *Mol Endocrinol* 20(7):1479-1493.

215. Roe JS, Mercan F, Rivera K, Pappin DJ, & Vakoc CR (2015) BET Bromodomain Inhibition Suppresses the Function of Hematopoietic Transcription Factors in Acute Myeloid Leukemia. *Mol Cell* 58(6):1028-1039.

216. Pollex T & Furlong EEM (2017) Correlation Does Not Imply Causation: Histone Methyltransferases, but Not Histone Methylation, SET the Stage for Enhancer Activation. *Mol Cell* 66(4):439-441.

217. Kim TK, *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465(7295):182-187.

218. Andersson R, *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493):455-461.

219. De Santa F, *et al.* (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* 8(5):e1000384.

220. Li W, *et al.* (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 498(7455):516-520.

221. Schaukowitch K, *et al.* (2014) Enhancer RNA facilitates NELF release from immediate early genes. *Mol Cell* 56(1):29-42.

222. Arner E, *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* 347(6225):1010-1014.

223. Kaikkonen MU, *et al.* (2013) Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* 51(3):310-325.

224. Michel M, *et al.* (2017) TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Mol Syst Biol* 13(3):920.

225. Mousavi K, *et al.* (2013) eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell* 51(5):606-617.

226. Gilchrist DA, *et al.* (2010) Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* 143(4):540-551.

227. Sigova AA, *et al.* (2015) Transcription factor trapping by RNA in gene regulatory elements. *Science* 350(6263):978-981.

228. Gardini A, *et al.* (2014) Integrator regulates transcriptional initiation and pause release following activation. *Mol Cell* 56(1):128-139.

229. Bose DA, *et al.* (2017) RNA Binding to CBP Stimulates Histone Acetylation and Transcription. *Cell* 168(1-2):135-149 e122.

230. Hsieh TH, *et al.* (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* 162(1):108-119.

231. Isoda T, *et al.* (2017) Non-coding Transcription Instructs Chromatin Folding and Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. *Cell* 171(1):103-119 e118.

232. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, & Sharp PA (2017) A Phase Separation Model for Transcriptional Control. *Cell* 169(1):13-23.

233. Muerdter F & Stark A (2016) Gene Regulation: Activation through Space. *Curr Biol* 26(19):R895-R898.

234. Jain A & Vale RD (2017) RNA phase transitions in repeat expansion disorders. *Nature* 546(7657):243-247.

235. Berry J, Weber SC, Vaidya N, Haataja M, & Brangwynne CP (2015) RNA transcription modulates phase transition-driven nuclear body assembly. *Proc Natl Acad Sci U S A* 112(38):E5237-5245.

236. Young RS, Kumar Y, Bickmore WA, & Taylor MS (2017) Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol* 18(1):242.

237. Natoli G & Andrau JC (2012) Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet* 46:1-19.

238. Lieberman-Aiden E, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289-293.

239. Dostie J, *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16(10):1299-1309.

240. de Wit E & de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev* 26(1):11-24.

241. Hou C, Li L, Qin ZS, & Corces VG (2012) Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* 48(3):471-484.

242. Sexton T, *et al.* (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 148(3):458-472.

243. Nora EP, *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485(7398):381-385.

244. Dixon JR, *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376-380.

245. Gibcus JH & Dekker J (2013) The hierarchy of the 3D genome. *Mol Cell* 49(5):773-782.

246. Rao SSP, *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*:1-30.

247. Vietri Rudan M, *et al.* (2015) Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* 10(8):1297-1309.

248. Harmston N, *et al.* (2016) Topologically associated domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *bioRxiv*.

249. Sofueva S, *et al.* (2013) Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J* 32(24):3119-3129.

250. Dowen JM, *et al.* (2014) Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159(2):374-387.

251. Guo Y, *et al.* (2015) CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* 162(4):900-910.

252. Lupianez DG, *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161(5):1012-1025.

253. Flavahan WA, *et al.* (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* 529(7584):110-114.

254. Zhou J & Levine M (1999) A novel cis-regulatory element, the PTS, mediates an anti-insulator activity in the Drosophila embryo. *Cell* 99(6):567-575.

255. Swanson CI, Schwimmer DB, & Barolo S (2011) Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr Biol* 21(14):1186-1196.

256. Deng W, *et al.* (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149(6):1233-1244.

257. Song SH, Hou C, & Dean A (2007) A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. *Mol Cell* 28(5):810-822.

258. Dynan WS & Tjian R (1983) The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* 35(1):79-87.

259. Mahmoudi T, Katsani KR, & Verrijzer CP (2002) GAGA can mediate enhancer function in trans by linking two separate DNA molecules. *EMBO J* 21(7):1775-1781.

260. de Belle I, Cai S, & Kohwi-Shigematsu T (1998) The genomic sequences bound to special AT-rich sequence-binding protein 1 (SATB1) in vivo in Jurkat T cells are tightly associated with the nuclear matrix at the bases of the chromatin loops. *J Cell Biol* 141(2):335-348.

261. Cai S, Han HJ, & Kohwi-Shigematsu T (2003) Tissue-specific nuclear architecture and gene expression regulated by SATB1. *Nat Genet* 34(1):42-51.

262. Cai S, Lee CC, & Kohwi-Shigematsu T (2006) SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet* 38(11):1278-1288.

263. Alvarez JD, *et al.* (2000) The MAR-binding protein SATB1 orchestrates temporal and spatial expression of multiple genes during T-cell development. *Genes Dev* 14(5):521-535.

264. Dickinson LA, Joh T, Kohwi Y, & Kohwi-Shigematsu T (1992) A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. *Cell* 70(4):631-645.

265. Wen J, *et al.* (2005) SATB1 family protein expressed during early erythroid differentiation modifies globin gene expression. *Blood* 105(8):3330-3339.

266. Britanova O, Akopov S, Lukyanov S, Gruss P, & Tarabykin V (2005) Novel transcription factor Satb2 interacts with matrix attachment region DNA elements in a tissue-specific manner and demonstrates cell-type-dependent expression in the developing mouse CNS. *Eur J Neurosci* 21(3):658-668.

267. Dobreva G, *et al.* (2006) SATB2 is a multifunctional determinant of craniofacial patterning and osteoblast differentiation. *Cell* 125(5):971-986.

268. Gomez-Marin C, *et al.* (2015) Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc Natl Acad Sci U S A* 112(24):7542-7547.

269. Mitchell JA & Fraser P (2008) Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev* 22(1):20-25.

270. Palstra RJ, *et al.* (2008) Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS One* 3(2):e1661.

271. Butler JE & Kadonaga JT (2001) Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* 15(19):2515-2519.

272. Zabidi MA, *et al.* (2015) Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518(7540):556-559.

273. Hay D, *et al.* (2016) Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet* 48(8):895-903.

274. Maeda RK & Karch F (2011) Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Curr Opin Genet Dev* 21(2):187-193.

275. Visel A, *et al.* (2009) Functional autonomy of distant-acting human enhancers. *Genomics* 93(6):509-513.

276. Yuh CH & Davidson EH (1996) Modular cis-regulatory organization of Endo16, a gut-specific gene of the sea urchin embryo. *Development* 122(4):1069-1082.

277. Yuh CH, Bolouri H, & Davidson EH (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279(5358):1896-1902.

278. Stine ZE, McGaughey DM, Bessling SL, Li S, & McCallion AS (2011) Steroid hormone modulation of RET through two estrogen responsive enhancers in breast cancer. *Hum Mol Genet* 20(19):3746-3756.

279. Maekawa T, Imamoto F, Merlino GT, Pastan I, & Ishii S (1989) Cooperative function of two separate enhancers of the human epidermal growth factor receptor proto-oncogene. *J Biol Chem* 264(10):5488-5494.

280. Perry MW, Boettiger AN, & Levine M (2011) Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc Natl Acad Sci U S A* 108(33):13570-13575.

281. Marinic M, Aktas T, Ruf S, & Spitz F (2013) An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. *Dev Cell* 24(5):530-542.

282. Leddin M, *et al.* (2011) Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood* 117(10):2827-2838.

283. Iampietro C, Gummalla M, Mutero A, Karch F, & Maeda RK (2010) Initiator elements function to determine the activity state of BX-C enhancers. *PLoS Genet* 6(12):e1001260.

284. Mihaly J, *et al.* (2006) Dissecting the regulatory landscape of the Abd-B gene of the bithorax complex. *Development* 133(15):2983-2993.

285. Barolo S (2012) Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays* 34(2):135-141.

286. Frankel N, *et al.* (2010) Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* 466(7305):490-493.

287. Perry MW, Boettiger AN, Bothma JP, & Levine M (2010) Shadow enhancers foster robustness of Drosophila gastrulation. *Curr Biol* 20(17):1562-1567.

288. Bothma JP, *et al.* (2015) Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. *Elife* 4.

289. Bejerano G, *et al.* (2004) Ultraconserved elements in the human genome. *Science* 304(5675):1321-1325.

290. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134(1):25-36.

291. Tak YG & Farnham PJ (2015) Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* 8:57.

292. Arnold CD, *et al.* (2014) Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* 46(7):685-692.

293. Villar D, *et al.* (2015) Enhancer evolution across 20 mammalian species. *Cell* 160(3):554-566.

294. Lambert SA, *et al.* (2018) The Human Transcription Factors. *Cell* 175(2):598-599.

295. Payvar F, *et al.* (1983) Sequence-specific binding of glucocorticoid receptor to MTV DNA at sites within and upstream of the transcribed region. *Cell* 35(2 Pt 1):381-392.

296. Singh H, Sen R, Baltimore D, & Sharp PA (1986) A nuclear factor that binds to a conserved sequence motif in transcriptional control elements of immunoglobulin genes. *Nature* 319(6049):154-158.

297. Payvar F, *et al.* (1981) Purified glucocorticoid receptors bind selectively in vitro to a cloned DNA fragment whose transcription is regulated by glucocorticoids in vivo. *Proc Natl Acad Sci U S A* 78(11):6628-6632.

298. Scheidereit C, Geisse S, Westphal HM, & Beato M (1983) The glucocorticoid receptor binds to defined nucleotide sequences near the promoter of mouse mammary tumour virus. *Nature* 304(5928):749-752.

299. Dynan WS & Tjian R (1983) Isolation of transcription factors that discriminate between different promoters recognized by RNA polymerase II. *Cell* 32(3):669-680.

300. Rosenfeld PJ & Kelly TJ (1986) Purification of nuclear factor I by DNA recognition site affinity chromatography. *J Biol Chem* 261(3):1398-1408.

301. Kadonaga JT & Tjian R (1986) Affinity purification of sequence-specific DNA binding proteins. *Proc Natl Acad Sci U S A* 83(16):5889-5893.

302. Wu C, *et al.* (1987) Purification and properties of Drosophila heat shock activator protein. *Science* 238(4831):1247-1253.

303. Briggs MR, Kadonaga JT, Bell SP, & Tjian R (1986) Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. *Science* 234(4772):47-52.

304. Lee W, Mitchell P, & Tjian R (1987) Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements. *Cell* 49(6):741-752.

305. Mitchell PJ, Wang C, & Tjian R (1987) Positive and negative regulation of transcription in vitro: enhancer-binding protein AP-2 is inhibited by SV40 T antigen. *Cell* 50(6):847-861.

306. Imagawa M, Chiu R, & Karin M (1987) Transcription factor AP-2 mediates induction by two different signal-transduction pathways: protein kinase C and cAMP. *Cell* 51(2):251-260.

307. Kadonaga JT, Carner KR, Masiarz FR, & Tjian R (1987) Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell* 51(6):1079-1090.

308. Singh H, LeBowitz JH, Baldwin AS, Jr., & Sharp PA (1988) Molecular cloning of an enhancer binding protein: isolation by screening of an expression library with a recognition site DNA. *Cell* 52(3):415-423.

309. Miesfeld R, *et al.* (1984) Characterization of a steroid hormone receptor gene and mRNA in wild-type and mutant cells. *Nature* 312(5996):779-781.

310. Weinberger C, *et al.* (1985) Identification of human glucocorticoid receptor complementary DNA clones by epitope selection. *Science* 228(4700):740-742.

311. Govindan MV, Devic M, Green S, Gronemeyer H, & Chambon P (1985) Cloning of the human glucocorticoid receptor cDNA. *Nucleic Acids Res* 13(23):8293-8304.

312. Walter P, *et al.* (1985) Cloning of the human estrogen receptor cDNA. *Proc Natl Acad Sci U S A* 82(23):7889-7893.

313. Tacheny A, Dieu M, Arnould T, & Renard P (2013) Mass spectrometry-based identification of proteins interacting with nucleic acids. *J Proteomics* 94:89-109.

314. Hu S, *et al.* (2009) Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell* 139(3):610-622.

315. Brent R & Ptashne M (1985) A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* 43(3 Pt 2):729-736.

316. Finn RD, *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279-285.

317. Letunic I, Doerks T, & Bork P (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* 43(Database issue):D257-260.

318. Finn RD, *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* 45(D1):D190-D199.

319. Weirauch MT & Hughes TR (2011) A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. *Subcell Biochem* 52:25-73.

320. Fulton DL, *et al.* (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol* 10(3):R29.

321. Berman HM, *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235-242.

322. Rohs R, *et al.* (2010) Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* 79:233-269.

323. Slattery M, *et al.* (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* 39(9):381-399.

324. Stella S, Cascio D, & Johnson RC (2010) The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes Dev* 24(8):814-826.

325. Hancock SP, *et al.* (2013) Control of DNA minor groove width and Fis protein binding by the purine 2-amino group. *Nucleic Acids Res* 41(13):6750-6760.

326. Chen Y, *et al.* (2013) Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. *Nucleic Acids Res* 41(17):8368-8376.

327. Kitayner M, *et al.* (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat Struct Mol Biol* 17(4):423-429.

328. Chen Y, *et al.* (2012) DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep* 2(5):1197-1206.

329. Roth FP, Hughes JD, Estep PW, & Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 16(10):939-945.

330. Bailey TL & Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.

331. Pevzner PA & Sze SH (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol* 8:269-278.

332. Berger MF, *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24(11):1429-1435.

333. Berger MF & Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* 4(3):393-411.

334. Zhao Y, Granas D, & Stormo GD (2009) Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5(12):e1000590.

335. Jolma A, *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* 20(6):861-873.

336. Stormo GD & Zhao Y (2010) Determining the specificity of protein-DNA interactions. *Nat Rev Genet* 11(11):751-760.

337. Bulyk ML, Johnson PL, & Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* 30(5):1255-1261.

338. Jolma A, *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1-2):327-339.

339. Rohs R, *et al.* (2009) The role of DNA shape in protein-DNA recognition. *Nature* 461(7268):1248-1253.

340. Badis G, *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935):1720-1723.

341. Zhou Q & Liu JS (2008) Extracting sequence features to predict protein-DNA interactions: a comparative study. *Nucleic Acids Res* 36(12):4137-4148.

342. Zhao Y, Ruan S, Pandey M, & Stormo GD (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* 191(3):781-790.

343. Slattery M, *et al.* (2014) Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences* 39(9):381-399.

344. Wong D, *et al.* (2011) Extensive characterization of NF-kappaB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biol* 12(7):R70.

345. Siggers T, *et al.* (2011) Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-kappaB family DNA binding. *Nat Immunol* 13(1):95-102.

346. Rowan S, *et al.* (2010) Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes Dev* 24(10):980-985.

347. White MA, Parker DS, Barolo S, & Cohen BA (2012) A model of spatially restricted transcription in opposing gradients of activators and repressors. *Mol Syst Biol* 8:614.

348. Gordan R, *et al.* (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol* 12(12):R125.

349. Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10(9):605-616.

350. Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. *Dev Cell* 21(4):611-626.

351. Kim TH, *et al.* (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6):1231-1245.

352. Fu Y, Sinha M, Peterson CL, & Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* 4(7):e1000138.

353. Heinz S, *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576-589.

354. Slattery M, *et al.* (2013) Divergent transcriptional regulatory logic at the intersection of tissue growth and developmental patterning. *PLoS Genet* 9(9):e1003753.

355. Zhang JA, Mortazavi A, Williams BA, Wold BJ, & Rothenberg EV (2012) Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell* 149(2):467-482.

356. Kudron M, *et al.* (2013) Tissue-specific direct targets of Caenorhabditis elegans Rb/E2F dictate distinct somatic and germline programs. *Genome Biol* 14(1):R5.

357. Frietze S, *et al.* (2012) Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol* 13(9):R52.

358. Li XY, *et al.* (2011) The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol* 12(4):R34.

359. John S, *et al.* (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* 43(3):264-268.

360. Degner JF, *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482(7385):390-394.

361. Zhu F, *et al.* (2018) The interaction landscape between transcription factors and the nucleosome. *Nature* 562(7725):76-81.

362. John S, *et al.* (2008) Interaction of the glucocorticoid receptor with the chromatin landscape. *Mol Cell* 29(5):611-624.

363. Vicent GP, *et al.* (2009) Two chromatin remodeling activities cooperate during activation of hormone responsive promoters. *PLoS Genet* 5(7):e1000567.

364. Buck MJ & Lieb JD (2006) A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nat Genet* 38(12):1446-1451.

365. Carr A & Biggin MD (2000) Accessibility of transcriptionally inactive genes is specifically reduced at homeoprotein-DNA binding sites in Drosophila. *Nucleic Acids Res* 28(14):2839-2846.

366. Gertz J, *et al.* (2013) Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* 52(1):25-36.

367. Neph S, *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489(7414):83-90.

368. Gerstein MB, *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91-100.

369. Yanez-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(10):2018-2030.

370. Barozzi I, *et al.* (2014) Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell* 54(5):844-857.

371. Adams CC & Workman JL (1995) Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol Cell Biol* 15(3):1405-1421.

372. Polach KJ & Widom J (1996) A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J Mol Biol* 258(5):800-812.

373. Cirillo LA, *et al.* (2002) Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* 9(2):279-289.

374. Clark KL, Halay ED, Lai E, & Burley SK (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* 364(6436):412-420.

375. Ghisletti S, *et al.* (2010) Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* 32(3):317-328.

376. Decker T, *et al.* (2009) Stepwise activation of enhancer and promoter regions of the B cell commitment gene Pax5 in early lymphopoiesis. *Immunity* 30(4):508-520.

377. Gualdi R, *et al.* (1996) Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. *Genes Dev* 10(13):1670-1682.

378. Lee CS, Friedman JR, Fulmer JT, & Kaestner KH (2005) The initiation of liver development is dependent on Foxa transcription factors. *Nature* 435(7044):944-947.

379. Sherwood RI, *et al.* (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology* 32(2):171-178.

380. Stefflova K, *et al.* (2013) Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* 154(3):530-540.

381. Heinz S, *et al.* (2013) Effect of natural genetic variation on enhancer selection and function. *Nature* 503(7477):487-492.

382. Tuteja G, Jensen ST, White P, & Kaestner KH (2008) Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site. *Nucleic Acids Res* 36(12):4149-4157.

383. Ptashne M, *et al.* (1980) How the lambda repressor and cro work. *Cell* 19(1):1-11.

384. Lebrecht D, *et al.* (2005) Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. *Proc Natl Acad Sci U S A* 102(37):13176-13181.

385. Szymanski P & Levine M (1995) Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo. *EMBO J* 14(10):2229-2238.

386. Jolma A, *et al.* (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527(7578):384-388.

387. Slattery M, *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147(6):1270-1282.

388. Panne D (2008) The enhanceosome. *Curr Opin Struct Biol* 18(2):236-242.

389. Tahirov TH, *et al.* (2001) Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. *Cell* 104(5):755-767.

390. Waszak SM, *et al.* (2015) Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* 162(5):1039-1050.

391. Ding Z, *et al.* (2014) Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet* 10(11):e1004798.

392. Kilpinen H, *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 342(6159):744-747.

393. McVicker G, *et al.* (2013) Identification of genetic variants that affect histone modifications in human cells. *Science* 342(6159):747-749.

394. Keegan L, Gill G, & Ptashne M (1986) Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* 231(4739):699-704.

395. Lin YS, Carey MF, Ptashne M, & Green MR (1988) GAL4 derivatives function alone and synergistically with mammalian activators in vitro. *Cell* 54(5):659-664.

396. Cujec TP, *et al.* (1997) The human immunodeficiency virus transactivator Tat interacts with the RNA polymerase II holoenzyme. *Mol Cell Biol* 17(4):1817-1823.

397. Fry CJ, Slansky JE, & Farnham PJ (1997) Position-dependent transcriptional regulation of the murine dihydrofolate reductase promoter by the E2F transactivation domain. *Mol Cell Biol* 17(4):1966-1976.

398. Gill G, Pascal E, Tseng ZH, & Tjian R (1994) A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc Natl Acad Sci U S A* 91(1):192-196.

399. Goodrich JA, Hoey T, Thut CJ, Admon A, & Tjian R (1993) Drosophila TAFII40 interacts with both a VP16 activation domain and the basal transcription factor TFIIB. *Cell* 75(3):519-530.

400. Horikoshi M, Hai T, Lin YS, Green MR, & Roeder RG (1988) Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex. *Cell* 54(7):1033-1042.

401. Lin YS, Ha I, Maldonado E, Reinberg D, & Green MR (1991) Binding of general transcription factor TFIIB to an acidic activating region. *Nature* 353(6344):569-571.

402. Roberts SG, Choy B, Walker SS, Lin YS, & Green MR (1995) A role for activator-mediated TFIIB recruitment in diverse aspects of transcriptional regulation. *Curr Biol* 5(5):508-516.

403. Stringer KF, Ingles CJ, & Greenblatt J (1990) Direct and selective binding of an acidic transcriptional activation domain to the TATA-box factor TFIID. *Nature* 345(6278):783-786.

404. Zhu H, Joliot V, & Prywes R (1994) Role of transcription factor TFIIF in serum response factor-activated transcription. *J Biol Chem* 269(5):3489-3497.

405. Kim TK & Roeder RG (1994) Proline-rich activator CTF1 targets the TFIIB assembly step during transcriptional activation. *Proc Natl Acad Sci U S A* 91(10):4170-4174.

406. Chiang CM & Roeder RG (1995) Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. *Science* 267(5197):531-536.

407. Tanese N, Pugh BF, & Tjian R (1991) Coactivators for a proline-rich activator purified from the multisubunit human TFIID complex. *Genes Dev* 5(12A):2212-2224.

408. Hoey T, *et al.* (1993) Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. *Cell* 72(2):247-260.

409. Chen JL, Attardi LD, Verrijzer CP, Yokomori K, & Tjian R (1994) Assembly of recombinant TFIID reveals differential coactivator requirements for distinct transcriptional activators. *Cell* 79(1):93-105.

410. Frietze S & Farnham PJ (2011) Transcription factor effector domains. *Subcell Biochem* 52:261-277.

411. Cress WD & Triezenberg SJ (1991) Critical structural elements of the VP16 transcriptional activation domain. *Science* 251(4989):87-90.

412. Drysdale CM, *et al.* (1995) The transcriptional activator GCN4 contains multiple activation domains that are critically dependent on hydrophobic amino acids. *Mol Cell Biol* 15(3):1220-1233.

413. Regier JL, Shen F, & Triezenberg SJ (1993) Pattern of aromatic and hydrophobic amino acids critical for one of two subdomains of the VP16 transcriptional activator. *Proc Natl Acad Sci U S A* 90(3):883-887.

414. Warfield L, Tuttle LM, Pacheco D, Klevit RE, & Hahn S (2014) A sequence-specific transcription activator motif and powerful synthetic variants that bind Mediator using a fuzzy protein interface. *Proc Natl Acad Sci U S A* 111(34):E3506-3513.

415. Staller MV, *et al.* (2018) A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst* 6(4):444-455 e446.

416. Chi SW, *et al.* (2005) Structural details on mdm2-p53 interaction. *J Biol Chem* 280(46):38795-38802.

417. Uesugi M & Verdine GL (1999) The alpha-helical FXXPhiPhi motif in p53: TAF interaction and discrimination by MDM2. *Proc Natl Acad Sci U S A* 96(26):14801-14806.

418. Garza AS, Ahmad N, & Kumar R (2009) Role of intrinsically disordered protein regions/domains in transcriptional regulation. *Life Sci* 84(7-8):189-193.

419. Brzozowski AM, *et al.* (1997) Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* 389(6652):753-758.

420. Kumar R, Betney R, Li J, Thompson EB, & McEwan IJ (2004) Induced alpha-helix structure in AF1 of the androgen receptor upon binding transcription factor TFIIF. *Biochemistry* 43(11):3008-3013.

421. Pike AC, *et al.* (1999) Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist. *EMBO J* 18(17):4608-4618.

422. Blair WS, Bogerd HP, Madore SJ, & Cullen BR (1994) Mutational analysis of the transcription activation domain of RelA: identification of a highly synergistic minimal acidic activation module. *Mol Cell Biol* 14(11):7226-7234.

423. Blau J, *et al.* (1996) Three functional classes of transcriptional activation domain. *Mol Cell Biol* 16(5):2044-2055.

424. Krumm A, Hickey LB, & Groudine M (1995) Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev* 9(5):559-572.

425. Yankulov K, Blau J, Purton T, Roberts S, & Bentley DL (1994) Transcriptional elongation by RNA polymerase II is stimulated by transactivators. *Cell* 77(5):749-759.

426. Huber HE, *et al.* (1993) Transcription factor E2F binds DNA as a heterodimer. *Proc Natl Acad Sci U S A* 90(8):3525-3529.

427. Mangelsdorf DJ & Evans RM (1995) The RXR heterodimers and orphan receptors. *Cell* 83(6):841-850.

428. Remenyi A, *et al.* (2003) Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev* 17(16):2048-2059.

429. Remenyi A, Tomilin A, Scholer HR, & Wilmanns M (2002) Differential activity by DNA-induced quarternary structures of POU transcription factors. *Biochem Pharmacol* 64(5-6):979-984.

430. Chen X, *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133(6):1106-1117.

431. Mathur D, *et al.* (2008) Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET. *Genome Biol* 9(8):R126.

432. Reginato MJ, Zhang J, & Lazar MA (1996) DNA-independent and DNA-dependent mechanisms regulate the differential heterodimerization of the isoforms of the thyroid hormone receptor with retinoid X receptor. *J Biol Chem* 271(45):28199-28205.

433. Leng X, *et al.* (1994) Mechanisms for synergistic activation of thyroid hormone receptor and retinoid X receptor on different response elements. *J Biol Chem* 269(50):31436-31442.

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

435. Chan HM & La Thangue NB (2001) p300/CBP proteins: HATs for transcriptional bridges and scaffolds. *J Cell Sci* 114(Pt 13):2363-2373.

436. Kalkhoven E (2004) CBP and p300: HATs for different occasions. *Biochem Pharmacol* 68(6):1145-1155.

437. Belandia B & Parker MG (2003) Nuclear receptors: a rendezvous for chromatin remodeling factors. *Cell* 114(3):277-280.

438. Vierbuchen T, *et al.* (2017) AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Mol Cell* 68(6):1067-1082 e1012.

439. Pedersen TA, Kowenz-Leutz E, Leutz A, & Nerlov C (2001) Cooperation between C/EBPalpha TBP/TFIIB and SWI/SNF recruiting domains is required for adipocyte differentiation. *Genes Dev* 15(23):3208-3216.

440. Cheng SW, *et al.* (1999) c-MYC interacts with INI1/hSNF5 and requires the SWI/SNF complex for transactivation function. *Nat Genet* 22(1):102-105.

441. Lessard J, *et al.* (2007) An essential switch in subunit composition of a chromatin remodeling complex during neural development. *Neuron* 55(2):201-215.

442. Wu JI, *et al.* (2007) Regulation of dendritic development by neuron-specific chromatin remodeling complexes. *Neuron* 56(1):94-108.

443. Takeuchi JK & Bruneau BG (2009) Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors. *Nature* 459(7247):708-711.

444. Chandy M, Gutierrez JL, Prochasson P, & Workman JL (2006) SWI/SNF displaces SAGA-acetylated nucleosomes. *Eukaryot Cell* 5(10):1738-1747.

445. Privalsky ML (2004) The role of corepressors in transcriptional regulation by nuclear hormone receptors. *Annu Rev Physiol* 66:315-360.

446. Rosenfeld MG, Lunyak VV, & Glass CK (2006) Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes Dev* 20(11):1405-1428.

447. Urrutia R (2003) KRAB-containing zinc-finger repressor proteins. *Genome Biol* 4(10):231.

448. Groner AC, *et al.* (2010) KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. *PLoS Genet* 6(3):e1000869.

449. Borggrefe T & Yue X (2011) Interactions between subunits of the Mediator complex with gene-specific transcription factors. *Semin Cell Dev Biol* 22(7):759-768.

450. Stampfel G, *et al.* (2015) Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* 528(7580):147-151.

451. Brzovic PS, *et al.* (2011) The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. *Mol Cell* 44(6):942-953.

452. Tuttle LM, *et al.* (2018) Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. *Cell Rep* 22(12):3251-3264.

453. Mao YS, Zhang B, & Spector DL (2011) Biogenesis and function of nuclear bodies. *Trends Genet* 27(8):295-306.

454. Zhu L & Brangwynne CP (2015) Nuclear bodies: the emerging biophysics of nucleoplasmic phases. *Curr Opin Cell Biol* 34:23-30.

455. Banani SF, Lee HO, Hyman AA, & Rosen MK (2017) Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* 18(5):285-298.

456. Bergeron-Sandoval LP, Safaee N, & Michnick SW (2016) Mechanisms and Consequences of Macromolecular Phase Separation. *Cell* 165(5):1067-1079.

457. Kwon I, *et al.* (2013) Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell* 155(5):1049-1060.

458. Boija A, *et al.* (2018) Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* 175(7):1842-1855 e1816.

459. Sabari BR, *et al.* (2018) Coactivator condensation at super-enhancers links phase separation and gene control. *Science* 361(6400).

460. Melnikov A, *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30(3):271-277.

461. Patwardhan RP, *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30(3):265-270.

462. Murtha M, *et al.* (2014) FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat Methods* 11(5):559-565.

463. Dickel DE, *et al.* (2014) Function-based identification of mammalian enhancers using site-specific integration. *Nat Methods* 11(5):566-571.

464. Gotea V, *et al.* (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* 20(5):565-577.

465. Sharon E, *et al.* (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* 30(6):521-530.

466. Smith RP, *et al.* (2013) Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* 45(9):1021-1028.

467. Yu M, *et al.* (1997) GA-binding protein-dependent transcription initiator elements. Effect of helical spacing between polyomavirus enhancer a factor 3(PEA3)/Ets-binding sites on initiator activity. *J Biol Chem* 272(46):29060-29067.

468. Burz DS, Rivera-Pomar R, Jackle H, & Hanes SD (1998) Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. *EMBO J* 17(20):5998-6009.

469. Lam FH, Steger DJ, & O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453(7192):246-250.

470. Farley EK, *et al.* (2015) Suboptimization of developmental enhancers. *Science* 350(6258):325-328.

471. Crocker J, *et al.* (2015) Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* 160(1-2):191-203.

472. Evans NC, Swanson CI, & Barolo S (2012) Sparkling insights into enhancer structure, function, and evolution. *Curr Top Dev Biol* 98:97-120.

473. Panne D, Maniatis T, & Harrison SC (2007) An atomic model of the interferon-beta enhanceosome. *Cell* 129(6):1111-1123.

474. Rastegar S, *et al.* (2008) The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol* 318(2):366-377.

475. Hare EE, Peterson BK, & Eisen MB (2008) A careful look at binding site reorganization in the even-skipped enhancers of Drosophila and sepsids. *PLoS Genet* 4(11):e1000268.

476. Hare EE, Peterson BK, Iyer VN, Meier R, & Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet* 4(6):e1000106.

477. Taher L, *et al.* (2011) Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* 21(7):1139-1149.

478. Swanson CI, Evans NC, & Barolo S (2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell* 18(3):359-370.

479. Senger K, *et al.* (2004) Immunity regulatory DNAs share common organizational features in Drosophila. *Mol Cell* 13(1):19-32.

480. Liu F & Posakony JW (2012) Role of architecture in the function and specificity of two Notch-regulated transcriptional enhancer modules. *PLoS Genet* 8(7):e1002796.

481. Junion G, *et al.* (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* 148(3):473-486.

482. Erceg J, *et al.* (2014) Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLoS Genet* 10(1):e1004060.

483. Verfaillie A, *et al.* (2016) Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. *Genome Res* 26(7):882-895.

484. Naar AM, *et al.* (1998) Chromatin, TAFs, and a novel multiprotein coactivator are required for synergistic activation by Sp1 and SREBP-1a in vitro. *Genes Dev* 12(19):3020-3031.

Chapter 2

Systematic dissection of genomic features determining transcription factor binding and enhancer function

Parts of this chapter were first published as:

Grossman SR, *et al.* (2017) Systematic dissection of genomic features determining transcription factor binding and enhancer function. ***Proc Natl Acad Sci U S A*** 114(7):E1291-E1300.

ABSTRACT

Enhancers regulate gene expression through the binding of sequence-specific transcription factors (TFs) to cognate motifs. Various features influence TF binding and enhancer function—including the chromatin state of the genomic locus, the affinities of the binding site, the activity of the bound TFs, and interactions among TFs. Yet, the precise nature and relative contributions of these features remain unclear. Here, we used massively parallel reporter assays (MPRA) involving 32,115 natural and synthetic enhancers, together with high-throughput *in vivo* binding assays, to systematically dissect the contribution of each of these features to the binding and activity of genomic regulatory elements that contain motifs for PPARγ, a TF that serves as a key regulator of adipogenesis. We show that distinct sets of features govern PPARγ binding versus enhancer activity. PPARγ binding is largely governed by the affinity of the specific motif site and higher-order features of the larger genomic locus, such as

chromatin accessibility. In contrast, the enhancer activity of PPARγ binding sites depends on varying contributions from dozens of TFs in the immediate vicinity, including interactions between combinations of these TFs. Different pairs of motifs follow different interaction rules, including sub-additive, additive, and super-additive interactions among specific classes of TFs, with both spatially constrained and flexible grammars. Our results provide a paradigm for the systematic characterization of the genomic features underlying regulatory elements, applicable to the design of synthetic regulatory elements or the interpretation of human genetic variation.

INTRODUCTION

Regulatory sequences in DNA encode the information necessary to establish precise patterns of gene expression across cell types and conditions. While thousands of megabases of potential regulatory sequences have been identified (1, 2), deciphering the regulatory code – that is, being able to recognize and design regulatory sequences corresponding to particular expression patterns based on the underlying sequence – remains a major challenge in biology.

Gene expression is orchestrated by transcription factors (TFs), which bind to cognate binding sites (with characteristic sequence motifs) within regulatory sequences and recruit or modify components of the transcriptional machinery (3, 4). In the past decade, experimental advances have enabled characterization of the binding motifs for hundreds of TFs *in vitro* (5-9), mapping of the genome-wide binding sites of TFs *in vivo* (10-14), and functional characterization of the enhancer activity of thousands of genomic sequences (15-19). Comparisons between these experiments, however, have revealed that only a small fraction of the potential TF-binding sites (TFBS) in eukaryotic genomes are actually occupied by TFs in any given cell type, and that these sites vary substantially across cell types and conditions (3, 19-21). Moreover, only a subset (~25-50%) of bound TFBS can drive transcription in reporter assays (17-19, 22). Understanding the regulatory code involves being able to explain the sequence features and mechanisms underlying the ability of enhancers to bind specific TFs and to drive transcription in a given cellular context.

Several features could influence the TF binding and enhancer activity of specific motif sites *in vivo*. First, variation in the binding and activity of motif sites may reflect differences in the affinity of binding sites, due to the motif sequence (23-25), latent motif

preferences induced by cofactors (26, 27), or additional specificity determinants outside the core motifs, such as A/T-rich stretches (28-30). Second, TF access to motif sites may be governed by nucleosomes or the larger chromatin landscape (31-37). Third, additional TF motifs in the surrounding sequence could influence binding directly through protein-protein interactions, or indirectly through cooperative nucleosome displacement (38) or changes to the chromosome structure (39-41). Fourth, TFs at nearby sites could also contribute to transcriptional activation, either independently or through particular combinations of bound TFs acting in concert (e.g., by promoting better contacts with cofactors and general transcription factors) or in opposition (e.g., by inhibiting each other by disrupting these contacts) (42-44).

The extent to which TF binding and transcriptional activity of an enhancer are controlled by the same or separate factors is generally unclear. In particular, systematically identifying and characterizing TF interactions has proven difficult. Key questions include whether TFs fall into distinct functional groups and what constraints on motif positioning and orientation exist for various interactions.

To address these questions, we focused on PPARγ-response elements (PPREs) as a model set of regulatory sequences. PPARγ is a nuclear receptor that binds in cooperative fashion as a heterodimer with RXR to the canonical nuclear receptor direct repeat 1 (DR1) motif (45). It functions as a core regulator in adipocytes, localizing to PPREs during differentiation and primarily acting as a transcriptional activator (46, 47). In mouse adipocytes, PPARγ is bound to only ~1 in 200 genomic instances of this motif—and, even in regions of open chromatin, to only ~1 in 16 motif instances.

Furthermore, only ~15% of the genes closest to PPARγ binding sites are upregulated during adipogenesis (48).

We used massively parallel reporter assays (MPRA), together with high-throughput *in vitro* and *in vivo* binding assays, to systematically manipulate motif affinity, cooperative interactions, and chromatin accessibility across thousands of PPREs and measure the effect on PPARγ binding and expression. MPRA involves testing huge collections of short regulatory sequences (≤ 150 bp) in parallel by coupling each to a transcription unit containing a matched DNA barcode (49) (**Fig. 1A,B**). In total, we collected data on 32,115 regulatory sequences.

We show that PPARγ binding depends on the affinity of the PPARγ motif and on the larger chromatin landscape, but not significantly on the sequence in the immediate vicinity of the PPARγ binding site, indicating cooperative protein-protein binding interactions are relatively scarce. In contrast, enhancer activity strongly depends on the motifs in the immediate vicinity, particularly a core set of 20-30 additional TF motifs. Notably, we show that in addition to the individual contributions of these motifs, particular combinations are also an important determinant of expression. We systematically identify and functionally test these interactions, and find diverse interaction rules for different pairs of TFs, including additive, inhibitory, and synergistic interactions with varying constraints on motif positioning. Notably, we found consistent interactions between families of TFs, suggesting they may influence transcriptional activation through distinct mechanisms. Together, these experiments present a comprehensive approach to dissect the sequence grammar that determines TF binding and enhancer function. Applying this approach to a broad range of enhancers and cell

types will help to determine the prevalence and generality of these rules, and potentially yield universal models to predict expression from regulatory sequences across diverse cellular contexts.

RESULTS

In vivo PPARγ binding on plasmids is governed by the core PPARγ motif

The mouse genome contains ~1.5 million PPARγ-motif sites, defined based on the canonical 16-base PPARγ/RXR DR1 motif (see Methods). Of these, only a small minority—between 5,000-10,000 sites—are actually occupied by PPARγ in adipocytes (46-48). In principle, the binding at specific sites in vivo might depend on: (i) latent properties of the motif instance, not captured by the consensus sequence (30), (ii) cooperative binding in the immediate vicinity by other TFs expressed in the cell type (either directly through protein-protein interactions or indirectly through cooperative competition with nucleosomes) (50), and (iii) differences in the accessibility of the sites due to the chromatin landscape (36, 51, 52).

To investigate the first two possibilities (latent motif affinity and cooperative binding), we used a pooled plasmid-based reporter system to explore PPARγ binding *in vivo* to genomic motif sites outside of their native chromatin context (**Fig. 1A,B**). We randomly chose 750 of the 6835 sites we previously identified in ChIP-seq experiments ("bound genomic sites") (48) and, for each, we chose a matched site that contained an identical 16-bp PPARγ motif but was not bound by PPARγ in adipocytes ("unbound genomic sites"). We synthesized an oligonucleotide pool (Pool 1) containing 145 bp centered on the PPARγ motif from each of these 1500 site, as well as 1500 control sequences, one for each site in which the core PPARγ motif was disrupted by swapping

A↔T and G↔C (**Fig. 1C**). We cloned these 3,000 sequences ("candidate enhancers") into plasmids containing a minimal promoter and an open reading frame (ORF) containing a unique barcode that identifies its specific upstream enhancer (49).

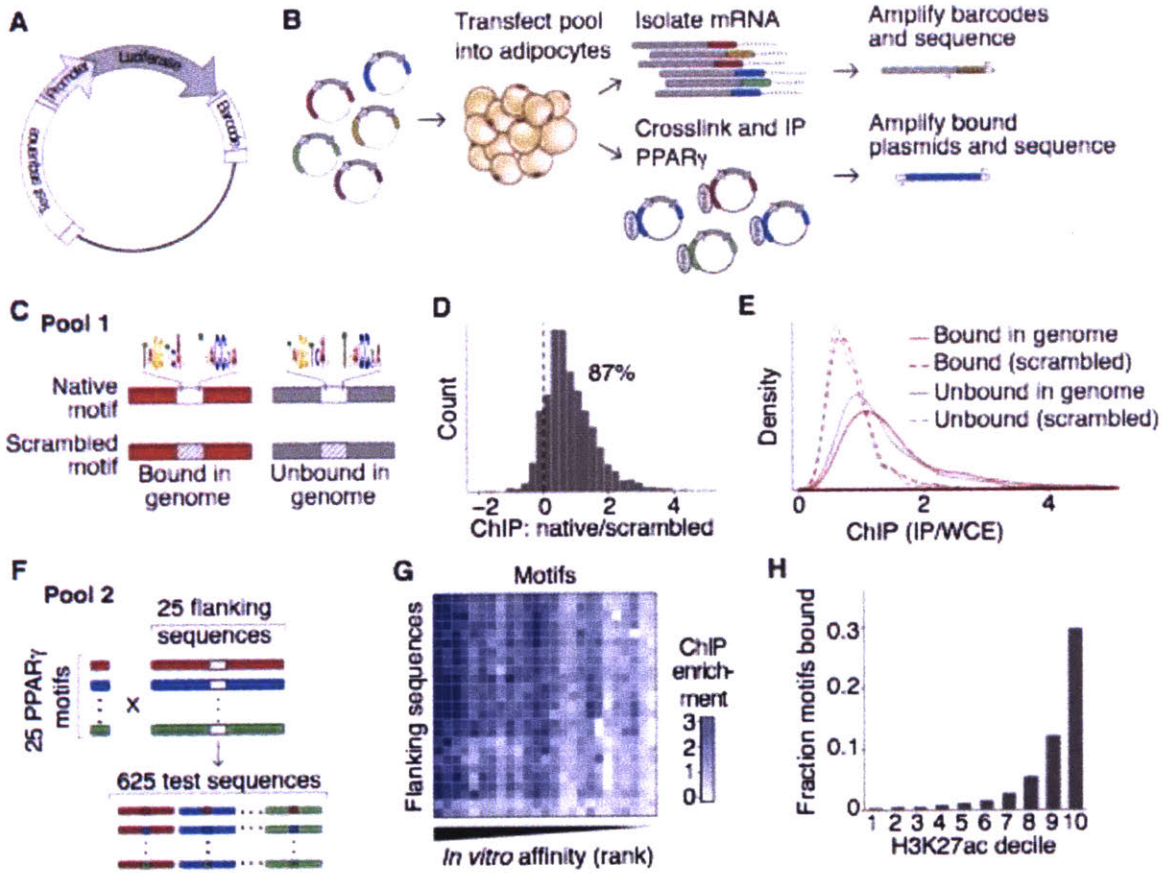


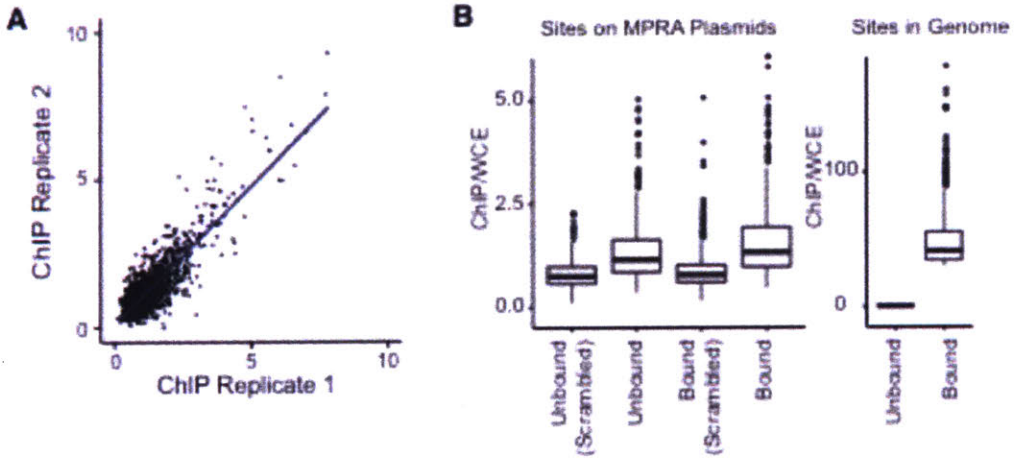
Figure 1. In vitro PPARγ binding is determined by core motif affinity. (A,B) Overview of pooled reporter system. **(A)** Candidate sequences were cloned into plasmids upstream of a minimal promoter and barcoded *luc2* ORF. **(B)** Plasmid pools were transfected into adipocytes, and assayed for PPARγ binding by ChIP-seq (bottom) and for enhancer activity by RNA-seq (top). **(C)** Schematic of Pool 1, containing 750 bound genomic PPARγ motif sites, 750 unbound genomic PPARγ motif sites, and these 1500 sites with the core PPARγ motif disrupted. (**D**) Log_2-ratio of ChIP enrichment for each genomic sequence with an intact and disrupted central PPARγ motif **(E)** PPARγ ChIP enrichments for bound and unbound genomic sequences with intact and disrupted core PPARγ motifs. **(F)** Schematic of Pool 2. The core PPARγ motif from 25 bound genomic sites was swapped into each of the other 24 flanking sequences, yielding a

matrix of 625 enhancer constructs. **(G)** ChIP enrichment for each core motif (columns) and flanking sequence (rows) in Pool 2. Core motifs were arranged by affinity measured by MITOMI (see Methods). (**H**) Fraction of genomic PPARγ motif sites bound by PPARγ, conditional on the H3K27ac ChIP enrichment score (48).

We transfected this plasmid pool into mouse 3T3-L1 adipocytes 7 days post-differentiation; grew the cells for 16 hours; and measured PPARγ binding by performing ChIP-Seq and calculating the relative enrichment of reads corresponding to each sequence. Measurements of relative binding activity were highly reproducible between two biological replicates (r=0.93, **Fig. S1A**). As a control, we also examined PPARγ binding across the genome from the same experiment and confirmed that the observed binding sites were consistent with those identified in our previous ChIP-Seq experiments with PPARγ (**Fig. S1B-D**).

For candidate enhancers corresponding to both bound and unbound genomic sites, disrupting the PPARγ motif significantly reduced binding ($p_{Wilcox}<2.2x10^{-16}$, **Fig. 1D**). The native sequence showed stronger binding than the disrupted control in 87% of cases. For the other 13%, the difference in binding to the native and disrupted motif sites was negligible (less than the technical variance between replicates (SI Appendix, **Fig. S2A**)). In half of these cases, these sequences contained a second PPARγ motif site. In the remaining cases, the native sequence exhibited only weak binding and tended to contain less robust matches of the PPARγ motif, suggesting they have lower affinity for PPARγ (SI Appendix, **Fig. S2B,C**).

To our surprise, we found little difference in PPARγ-binding between candidate enhancers corresponding to bound genomic sites versus those corresponding to unbound genomic sites (**Fig. 1E**). More precisely, the bound genomic sites showed



C

Quartile	Peaks in Mikkelsen et al.	ChIP Enrichment
	1st	1.97
	2nd	2.23
	3rd	3.22
	4th	13.79

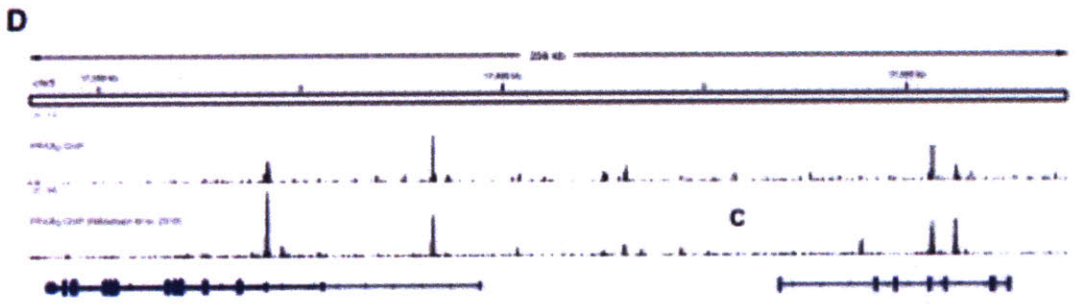


Figure S1. Validation of PPARγ ChIP. (A) Reproducibility of PPARγ ChIP enrichment ([reads in ChIP]/[reads in WCE]) of candidate enhancers in two biological replicates. (**B**) Distribution of PPARγ ChIP enrichment at bound and unbound sequences on plasmids and in the genome. (**C**) Genomic PPARγ ChIP enrichment at PPARγ binding sites in the genome detected in Mikkelsen et al., conditioned on PPARγ ChIP quartile in Mikkelsen et al. (**D**) Histogram of PPARγ ChIP enrichment at the *cd36* locus in this study (top) and Mikkelsen et al. (bottom).

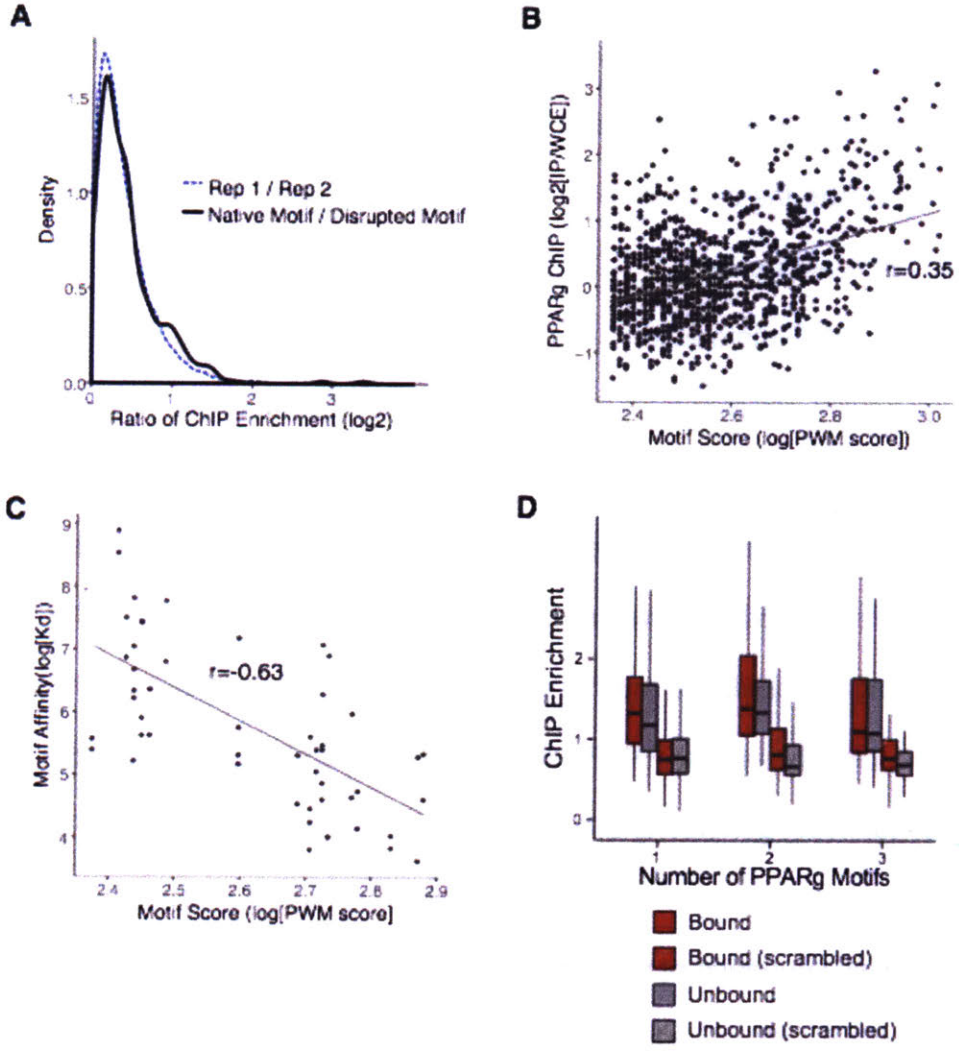


Figure S2. Sequences with stronger ChIP enrichment at the disrupted motif site than the native motif site contain multiple PPARγ motif sites or weak motif sites. **(A)** For sequences with stronger ChIP enrichment at the disrupted motif site than the native motif site, the difference between the two is negligible (within the technical noise of the system). Black distribution represents differences between native site and disrupted site for all sites with higher binding to the disrupted site. Blue dashed distribution represents the differences between the ChIP scores in two technical replicates. (**B**) Relationship between PWM score and ChIP signal on plasmids. (C) Relationship between PWM and K_d measured *in vitro*. (**D**) PPARγ binding to candidate enhancers, conditioned on the total number of PPARγ motif sites in the sequence.

slightly higher average binding, but this difference is largely explained by the fact that the sequences flanking the bound sites have a slightly higher average number of PPARγ-motif sites (average of 0.5 in bound vs. 0.2 in unbound). For bound and unbound genomic sites with the *same* number of PPARγ-motif sites, the difference is no longer statistically significant ($p_{F\text{-}test}$=0.67; SI Appendix, **Fig. S2D**).

The fact that the DNA immediately surrounding bound and unbound sites appears to have equivalent ability to bind PPARγ when reintroduced on plasmids into adipocytes suggests that PPARγ-binding depends primarily on the core PPARγ motif site and is not significantly influenced by elements or interactions in the immediate surroundings. To test this hypothesis, we selected 25 bound genomic sites from the original pool with a range of predicted motif strengths, and created a second plasmid pool (Pool 2) by swapping the central PPARγ motif site from each sequence into the other 24 flanking sequences, generating a "matrix" of 625 candidate enhancers (**Fig. 1F**). Consistent with our hypothesis, in vivo binding to the plasmid pool strongly depended on the precise sequence of the core PPARγ motif site rather than on the flanking sequence (40% vs. 6% of variance explained; **Fig. 1G**).

For each central PPARγ motif sequence, we directly measured the *in vitro* binding affinity, using a microfluidic device that assays association and dissociation of fluorescently labeled DNA oligonucleotides with PPARγ protein immobilized on the surface of the device (see Methods) (53). The binding observed in our cellular ChIP assay was was well predicted by the *in vitro* affinity measurements of the motifs (**Fig. 1G;** SI Appendix, **S3A, Table S1**). Specifically, binding (enrichment in the ChIP assay) fell linearly with affinity ($\log(K_d)$) down to affinity of $\log(K_d)$ = 6.5 and remained

thereafter. (The predictions were between the 25th and 75th percentiles for all but 4 of the 25 motifs and between the 5th and 95th percentiles for all of the 25 motifs.) Thus, the ability of genomic sequences containing PPARγ motifs to bind PPARγ on episomes is mainly determined—in a quantitative manner—by the core motif site, and therefore not by cooperative binding interactions with elements in the flanking sequence.

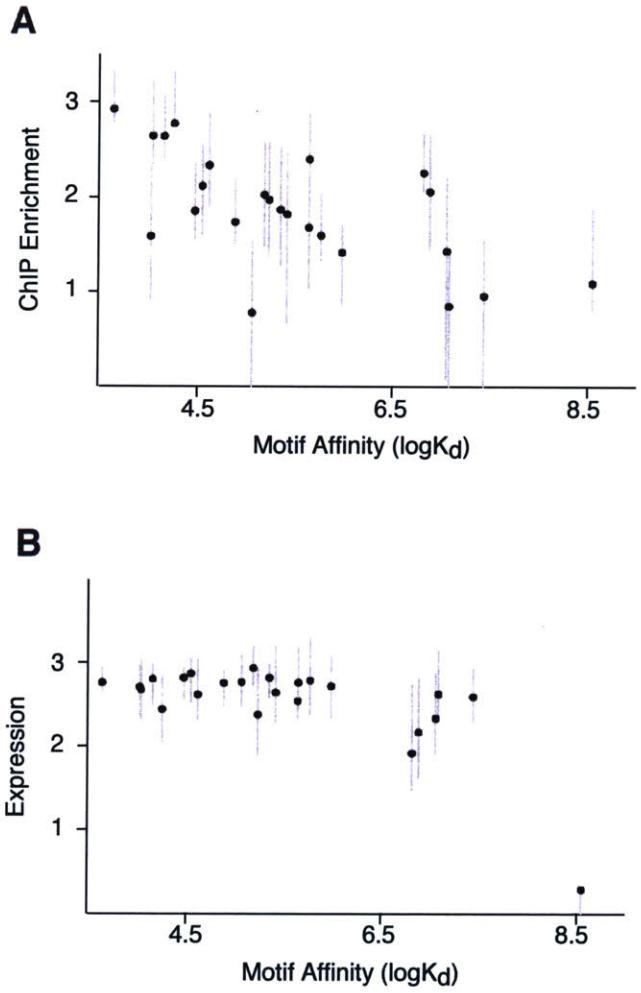


Figure S3. In vitro Kd of motif site predicts PPARγ binding but not enhancer activity. (**A**) Correlation between *in vitro* affinity of core PPARγ motif instance (x-axis) and median PPARγ binding to candidate enhancers in Pool 2 containing each motif instance (y-axis). (**B**) Correlation between in vitro affinity of core PPARγ motif instance (x-axis) and median expression (log_2[RNA/DNA]) of candidate enhancers in Pool 2 containing each motif instance (y-axis). Error bars represent first and third quartiles.

In vivo PPARγ binding to genomic PPARγ motifs is closely related to the chromatin landscape

Our results suggest that the explanation for why certain PPARγ motif sites are differentially bound *in vivo* lies neither in differences in affinities of the core motif site, nor in cooperative protein-protein interactions with TFs that bind in the immediate vicinity. Instead, our data indicate that PPARγ binding at a motif site is strongly correlated with the epigenomic context of the larger genomic locus.

The vast majority (85%) of the PPARγ-bound motif sites lie in regions of open chromatin, defined in terms of DNAse hypersensitivity (as assayed by FAIRE-seq (54)) and marked by H3K4me1/H3K27ac identified in adipocytes using ChIP-Seq (48). Moreover, genomic PPARγ binding *within* open regions was strongly dependent on the quantitative DNA accessibility and the enrichment of chromatin marks such as H3K27ac (**Fig. 1H;** SI Appendix, **S4A,B**). While only ~1 in 10 PPARγ motif sites in open regions enriched for active chromatin marks (H3K4me1/2/3 or H3K27ac) are bound, we correctly identify the majority (82%) of bound motif sites with a precision of ~1 in 4 by using a logistic classifier based on five chromatin modifications (H3K4me1/2/3, H3K27ac, and H3K27me)(SI Appendix, **Fig. S4C**). Among sites predicted to be bound, those that are actually bound tend to have motif sites with a better motif match (as measured by PWM score; SI Appendix, **Fig. S4D**). This suggests some remaining specificity may be due to differences in motif affinity, consistent with our findings for motif sites on plasmids.

In fact, our observation that *within open chromatin* TF binding is strongly correlated with the quantitative level of active chromatin marks appears to apply to many TFs. We analyzed 61 sequence-specific TFs profiled in ENCODE in 7 cell types

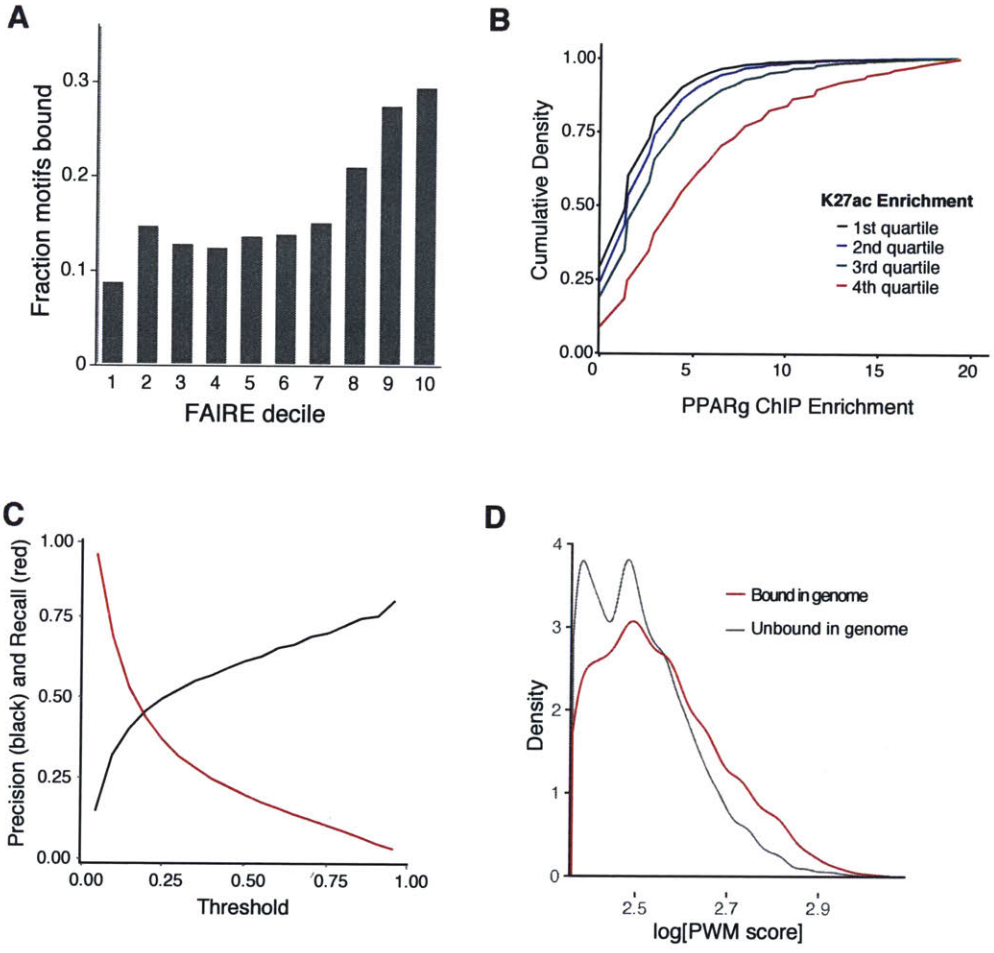


Figure S4. DNA accessibility and chromatin marks predict PPARγ binding. (A) Fraction of bound PPARγ motif sites in within FAIRE-seq peak, conditional on quantitative FAIRE-seq signal. (**B**) Distribution of genomic PPARγ ChIP enrichment at motif sites in each quartile of K27ac enrichment. (**C**) Performance of classifier predicting genomic PPARγ binding using chromatin marks. Sensitivity and specificity of logistic classifier predicting PPARγ occupancy at genomic motif sites using ChIP enrichment of H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K27me. (**D**) Distribution of PWM scores at motif sites predicted to be bound by classifier, conditional on observed PPARγ binding.

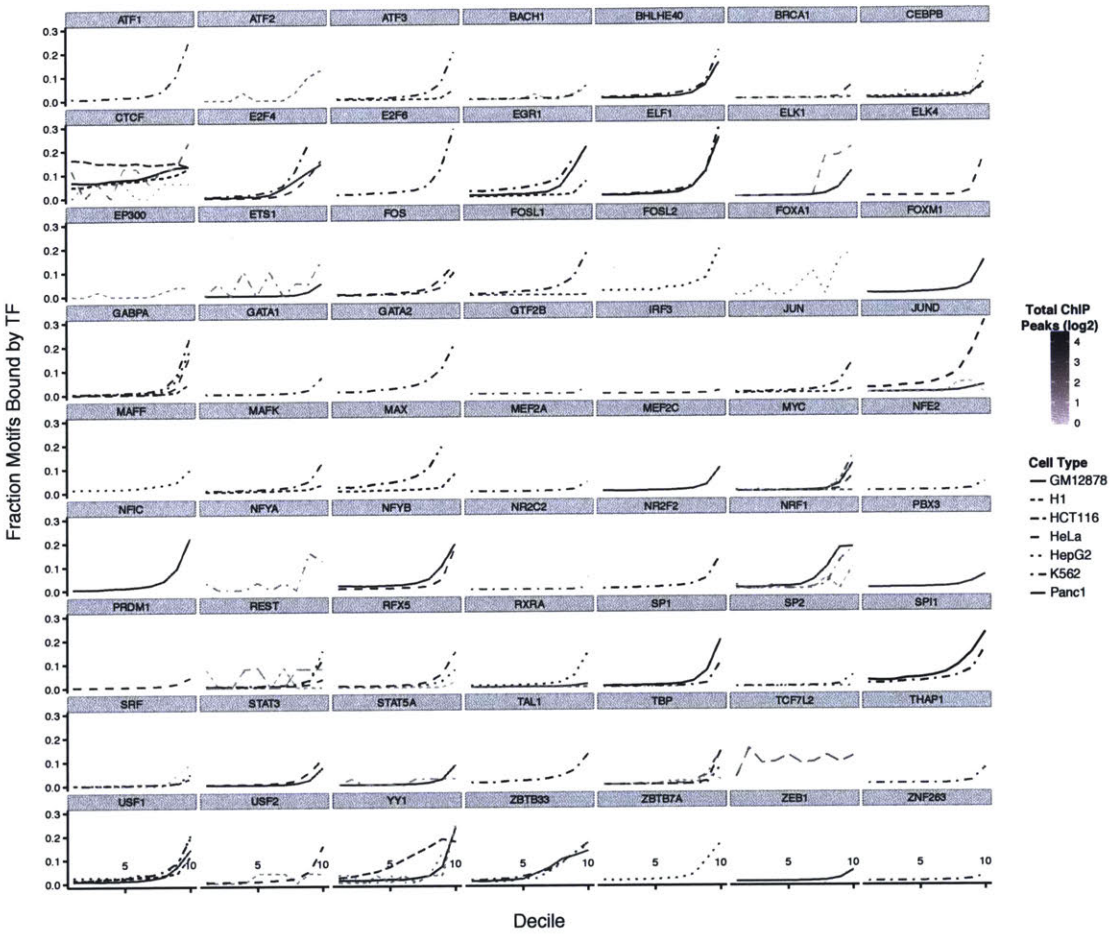


Figure S5. Quantitative relationship between TF binding and H3K27ac in ENCODE datasets. Fraction of TF motif sites in human genome with significant ChIP enrichment of the corresponding TF, conditioned on quartile of H3K27ac.

(121 total TF-cell type pairs) and found that the binding of 35 of these TFs were significantly correlated with quantitative DNA accessibility (measured by DNase-seq), and 45 were significantly correlated with enrichment of H3K27ac (Bonferroni-corrected $p_{Spearman}$ <0.01) (SI Appendix, **Fig S5**). TFs whose binding was not correlated with DNA accessibility include several pioneer factors, such as FOXA1, C/EBP, and NF-YA (46, 55, 56), the silencing factor REST, which maintains a repressive chromatin state (57), and CTCF, which binds insulator elements without active chromatin marks (58).

We note that the fact that TF binding is *correlated* with activating chromatin marks does not prove the direction of causality: it is possible that PPARγ binding not only depends on, but also contributes to chromatin state (59, 60). With respect to DNAse hypersensitivity, however, it is known that many (33%) of the genomic sites bound by PPARγ in terminally differentiated adipocytes show DNase hypersensitivity in the first 4 hours of adipogenesis, *before* PPARγ is expressed (61).

Elements in the sequence flanking PPARγ motifs strongly affect gene expression

We next sought to understand the determinants of enhancer activity for PPARγ-motif sites in adipocytes. To explore this question, we measured the transcriptional activity of the 3000 sequences in Pool 1, consisting of 750 bound genomic sites, 750 unbound genomic sites, and their corresponding controls with disrupted core motif sites. We transfected the plasmid pool into 3T3-L1 cells 7 days post-differentiation; grew the cells for 16 hours; and extracted both RNA and DNA. We calculated a "relative enhancer activity" for each candidate enhancer, defined as the ratio of the proportion of total RNA to the proportion of total DNA corresponding to the enhancer (using the median ratio across the unique barcodes for each) (**Fig. 2A**). Measurements of relative enhancer activity were highly robust across three biological replicates (r=0.96-0.97) (SI Appendix, **Fig. S6A,B**).

As expected, disrupting the PPARγ motif site substantially decreased the relative enhancer activity in candidate enhancers, consistent with enhancer activity depending on PPARγ binding. Transcription was lower in 71% of all cases and 94% of cases where the expression from the native sequence was above the mean (**Fig. 2A**).

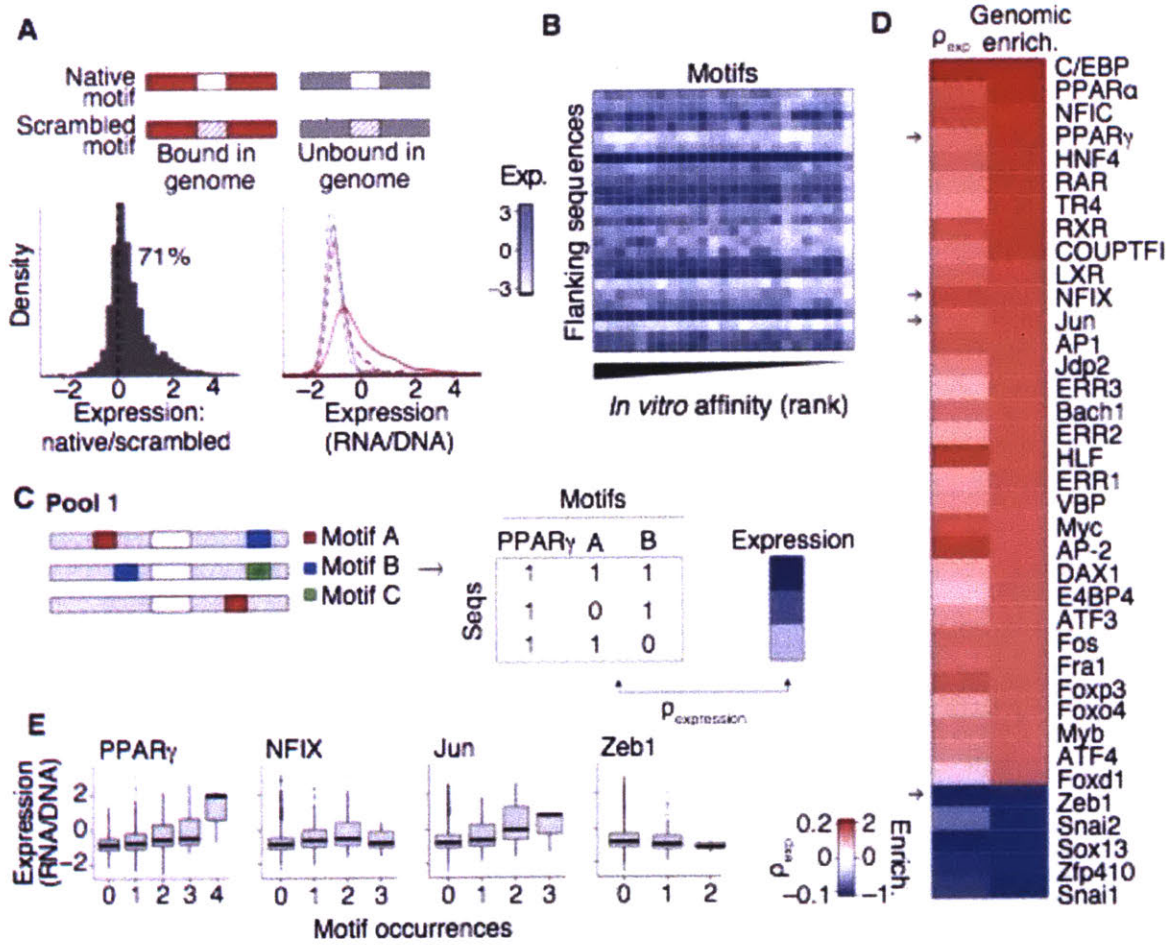


Figure 2. Elements in flanking sequence govern enhancer activity. (A) Left, ratio of expression (log_2[RNA/DNA]) for each genomic sequence with an intact vs. disrupted central PPARγ motif. Right, expression corresponding to bound and unbound genomic sites with intact and disrupted core PPARγ motifs. **(B)** Expression driven by sequence constructs in Pool 2, comprising 25 core PPARγ motifs (columns) swapped into 25 flanking sequences (rows). **(C)** Schematic of identification of TF motifs correlated with enhancer activity. For each TF motif, we calculated the correlation between motif counts and expression in Pool 1. **(D)** Counts of 38 motifs were significantly correlated with expression (FDR<0.01, top). These motifs are enriched or depleted around all 6835 bound motif sites in the genome (bottom). Arrows indicate motifs depicted in (E). **(E)** Expression of candidate enhancers in Pool 1, conditional on the number of occurrences of each motif.

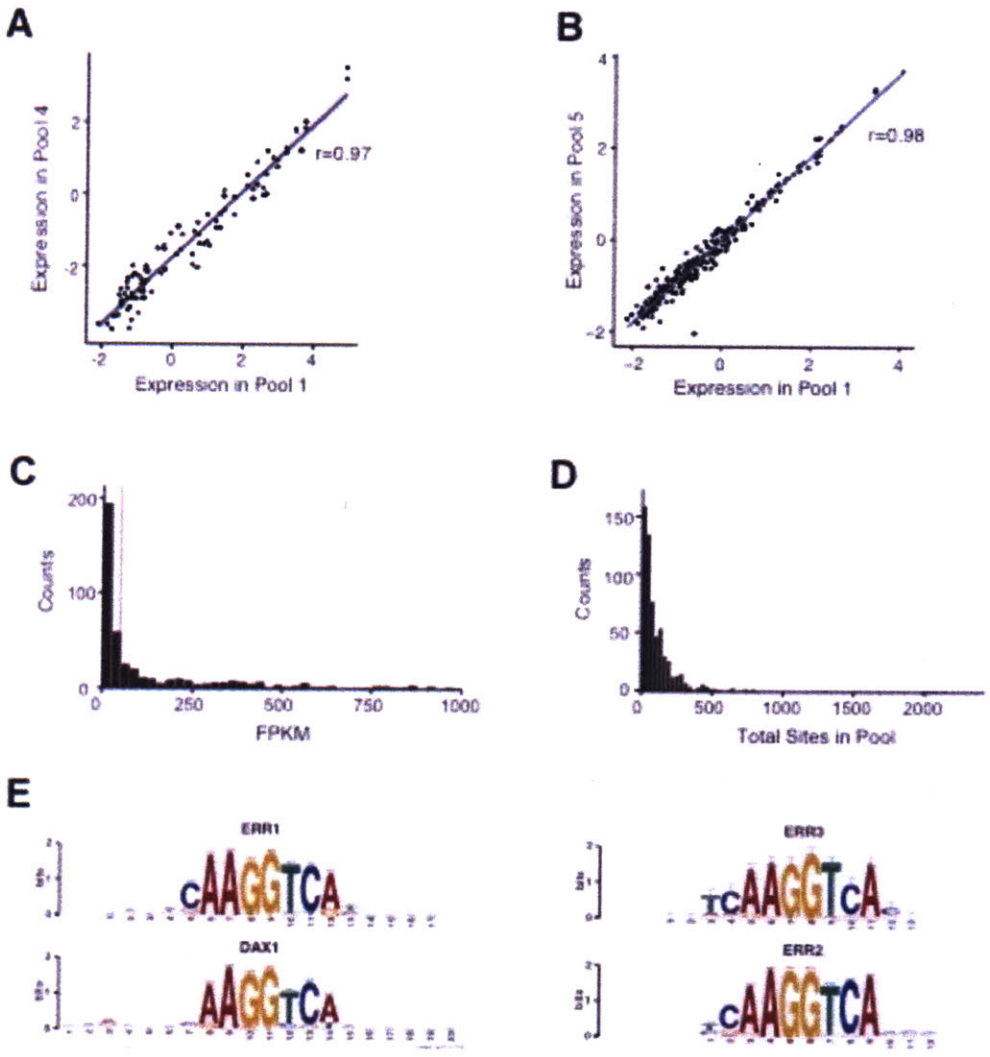


Figure S6. Validation of enhancer activity measurements and TF motif sites of candidate enhancers. (A, B) Reproducibility across 3 biological replicates of expression measurements (log_2[RNA/DNA]) of candidate enhancers included in multiple pools. (**C**) Histogram of expression of 614 TFs in 3T3-L1 adipocytes. (**D**) Histogram of total motif counts for 614 TFs in the 1500 genomic sequences included in Pool 1 (**E**) The motifs of correlated TFs not expressed in adipocytes (ERR2, ERR3, and DAX-1) are similar to the motif of ERR1, which is expressed.

Surprisingly, as described above, while the bound and unbound genomic PPARγ sites showed no significant difference in PPARγ binding affinity (**Fig. 1E**), these sites exhibited sharply different enhancer activity (**Fig. 2A**). Half of the sequences from

bound sites drove expression levels above the 95th percentile for sequences from unbound genomic sites. Moreover, the transcriptional activity from the unbound sites was only weakly affected by disrupting the core PPARγ motif site (median=1.05-fold decrease), indicating the expression from the unbound genomic sites is close to the background levels. Notably, the sequences from *bound* sites with *disrupted* PPARγ motif sites showed higher expression than sequences from the *unbound* genomic sites with the *native* PPARγ motifs (p_{Wilcox} =4.7x10^{-13}), suggesting the bound genomic sites are enriched for critical enhancer elements outside the core PPARγ motif site itself. (This observation holds true even when excluding sequences with addition PPARγ motif sites (p_{Wilcox} =4.3x10^{-6}).) Moreover, enhancer activity among bound genomic sites was only weakly correlated with PPARγ binding on plasmids ($\rho_{Spearman}$=0.24) or in the genome ($\rho_{Spearman}$=0.12). Together, these observations show that, although bound and unbound genomic sites do not differ in their inherent ability to bind PPARγ, they differ sharply in their enhancer activity as a result of additional elements in the surrounding sequence.

These results indicate that both the core motif sequence and the flanking sequence contribute to enhancer activity. To assess the relative contributions of each, we measured the enhancer activity of plasmids in Pool 2, comprising 25 core motifs inserted into 25 different flanking sequences (described above). Unlike binding activity, enhancer activity was largely explained by the flanking sequence rather than by the core motif sequence (84% vs 6% of variance explained; **Fig. 2B**). The median transcriptional activity was not substantially affected by changing the affinity of the core motif, although it did drop off for the motif with the lowest affinity as measured in the *in*

vitro assay (SI Appendix, **Fig. S1G**), consistent with the effect of disrupting the PPAR motif site. Thus, unlike PPAR binding, enhancer activity depends largely on the flanking sequence, provided that PPAR binding exceeds a threshold level.

Specific TF motifs correlate with transcriptional activity

To identify the elements in the flanking sequence that determine enhancer activity, we searched the sequences for known TF-binding motifs (**Fig. 2C**). We scanned the sequences using 1490 vertebrate motifs corresponding to 612 TFs (of which 400 are expressed in adipocytes; SI Appendix, **Fig. S6C,D**), and counted the number of (non-overlapping) occurrences of each motif. Across the 1500 candidate enhancers, the number of motif occurrences per TF ranged from 5 (Tcf7l2) to 807 (Sp1) (median=63; SI Appendix, **Fig. S6D**). Enhancer activity showed significant correlations with occurrences of 38 TF motifs, comprised of 33 positively correlated motifs and 5 negatively correlated motifs (FDR<0.001 in permuted datasets) (**Fig. 2D,E** and SI Appendix, **Table S2,S3**).

Several lines of evidence suggest the TFs corresponding to the correlated motifs may functionally contribute to gene regulation in adipocytes.

(i) 35 of the 38 TFs are expressed in adipocytes (vs. an expectation of only 26 by chance; hypergeometric test $p=5.0x10^{-6}$). The three "non-expressed" TFs (ERR2, ERR3, and DAX-1) are nuclear receptors with motifs highly similar to the motif of ERR1, another nuclear receptor expressed in adipocytes (SI Appendix, **Fig. S6E**).

(ii) Consistent with previous observations that functional enhancers often contain homotypic clusters of motif sites (62, 63), the presence of additional PPARγ/RXR motif sites correlated strongly with enhancer activity. Candidate enhancers containing

additional PPARγ/RXR motif sites showed nearly 2-fold higher enhancer activity than those with only a single motif site.

(iii) The TFs that recognize several of the positively correlated motifs are known to promote adipocyte differentiation (64, 65) or regulate gene expression in various stages of adipogenesis (54, 66-69). Conversely, the TFs that recognize several of the negatively correlated motifs are transcriptional repressors involved in inhibiting adipocyte-specific genes (68, 70) or promoting an alternate cell fate (71) (see Supplemental Note in SI Appendix).

(iv) Occurrences of the correlated motifs are enriched in the immediate vicinity of the bound PPARγ sites in the genome. Of the 33 positively correlated TF motifs (detected in the 750 bound sites included in Pool 1), 31 were significantly enriched. Conversely, all 5 negatively correlated TF-motifs were significantly depleted across the full set of 6,835 PPARγ-bound genomic sites compared to unbound sites in adipocytes (hypergeometic test $p=10^{-6}$ to 10^{-300}) (**Fig. 2D**). Moreover, the 31 positively correlated TFs were the most significantly enriched and the 5 negatively correlated TFs were the most significantly depleted TFs among all 612 TFs tested. (The two TFs that were not significantly enriched, Foxo4 and Foxp3, have fairly degenerate 4-nucleotide motifs.)

Together, these observations suggest that most of the correlated TF motifs identified in our assay indeed correspond to key regulators of gene expression in the adipocyte lineage.

TF motifs directly influence transcription

We reasoned that correlation between (i) the presence of specific TF-binding motifs in sequences from bound genomic sites and (ii) transcriptional activity in our

enhancer assay does not necessarily imply that TF-binding at these sites plays a causal role in determining transcriptional activity. An alternative possibility, for example, is that some TF-binding sites might be present at active enhancers because they were used for opening the chromatin before or during adipogenesis but do not contribute to driving transcriptional activity in adipocytes. We therefore next sought to identify motifs directly involved in transcription by deleting and inserting them in controlled contexts.

Disrupting TF motifs causes changes in expression

We first used an unbiased approach to identify elements that directly affect transcription, either motif sites and otherwise. We created an MPRA pool (Pool 4; "block-mutated enhancers") that systematically introduced mutations in sliding windows in 25 of the candidate enhancer sequences from bound genomic sites. First, we disrupted 10-bp blocks, tiled every 5 bp across the sequence (excluding the core PPARγ/RXR motif site). Next, we swapped 20-bp blocks of sequence between the bound genomic site and a matched unbound site (**Fig. 3A**, see Methods). For each mutant, we measured the enhancer activity and calculated the change in activity from the wild-type enhancer (**Fig. 3B**).

On average, mutations in the bound genomic sites that disrupted positively correlated TF-motif sites reduced expression more than those that did not disrupt such sites (median of 1.7-fold vs. 1.2-fold), and were 7 times more likely to cause a major decrease (>2-fold). Inserting blocks containing negatively correlated motifs (such as Zeb1 in the example shown in Fig. 3B) into the bound sites also substantially reduced expression (median of 2.0-fold). Finally, inserting blocks containing positively correlated motifs into unbound sites were 7 times more likely to cause a major increase in

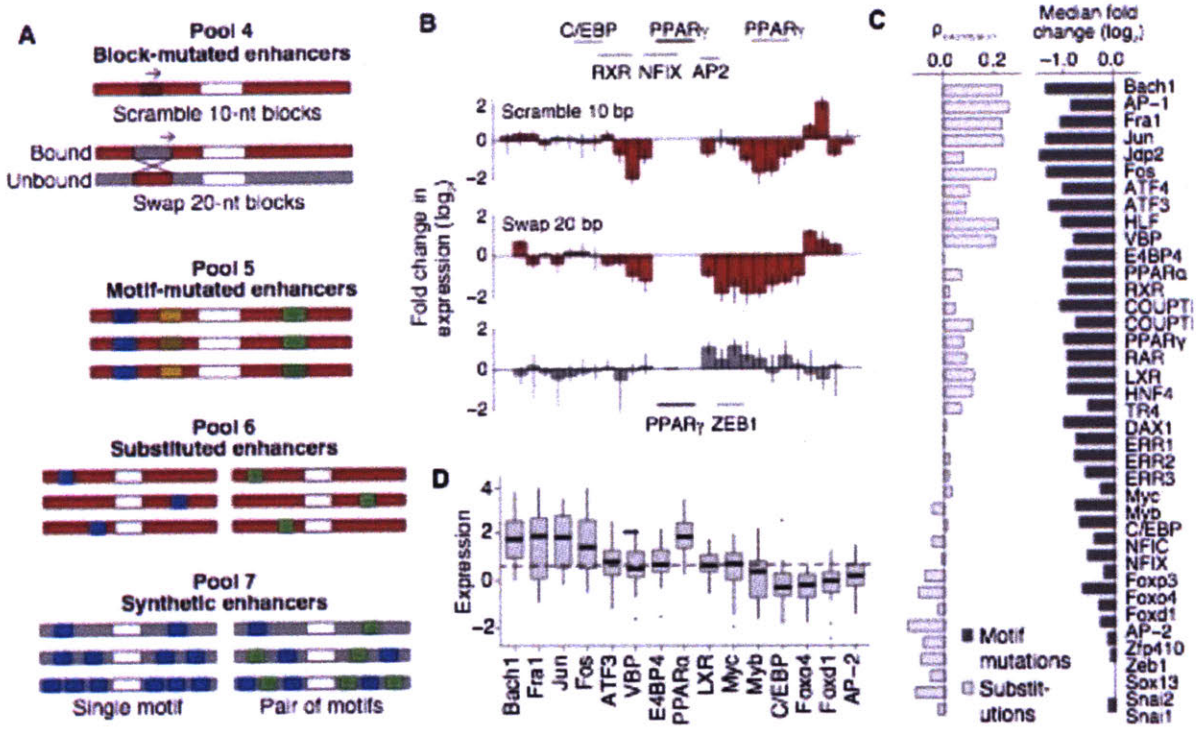


Figure 3. Disrupting TF motifs affects enhancer activity. (A) Schematic of motif deletion pools. Pool 4 (block-mutated enhancers, left): For 25 bound genomic sites, we disrupted 10 bp blocks tiled every 5 bp across the sequence (top), and swapped 20 bp blocks tiled every 5 bp across the sequence between bound and unbound genomic sites, matched by the sequence of the central PPARγ motif (bottom). In each case, the central PPARγ motif was left intact. Pool 5 (motif-mutated enhancers, right): each occurrence of the 38 significantly correlated motifs were disrupted across 375 bound genomic sites. Pool 6 (substituted enhancers, top): motif sites for each of the 38 correlated motifs were substituted into 90 existing motif sites in bound genomic sites. Pool 7 (synthetic enhancers, bottom): motif sites for 15 of the positively correlated motifs were added individually (left) and in pairs (right) to 3 neutral templates in various configurations (see text). **(B)** Example of changes in expression caused by tiled mutations in a bound sequences (red, chr8:90491327-90491472, top) and unbound sequence (gray, chr14:57223369-57223514, bottom). Bars represent the $\log_2$-ratio of the mutant and wild-type expression for the block centered at that position. **(C)** Right, median change in expression due to mutations in each motif across 375 bound genomic sites (see Fig. 3C). Left, correlation between change in counts and change in expression for each motif in the substituted enhancers. **(D)** Expression of synthetic enhancers containing multiple copies of one motif.

expression. Overall, these data suggests that the correlated motifs account for the majority of elements that strongly contribute to expression in these enhancers.

We next sought to distinguish the contribution of the individual correlated TF motifs to transcription levels. We created an MPRA pool (Pool 5; "motif-mutated enhancers") in which we systematically mutated each of the 38 significantly correlated motifs in 375 bound genomic sequences (**Fig. 3A,** see Methods). For each of the mutated motif sites, we also created a control by mutating an equally sized block that did not overlap any motif.

For the majority (27 of 33) of the positively correlated TF motifs, the mutations in the motif sites reduced expression significantly more than the mutations in control regions (p_{Wilcox} <0.05; SI Appendix, **Table S2**). Mutating ATF and AP-1 factors had the largest effect (87% of mutations decreasing expression by >2-fold; **Fig. 3C).** Mutations overlapping additional PPARγ-motifs sites surrounding the core PPARγ-motif site also substantially reduced expression (64% reducing expression by >2-fold).

Most interesting were the remaining 6 TFs (NFIX, NFIC, Foxo4, Foxp3, Foxd1, Myb, and AP-2), which did not appear to affect transcription in our assay – that is, these motifs are (i) enriched in genomic sequences that drive reporter expression but (ii) are not required for expression in our assay. A likely explanation is that these TFs act in different cellular contexts or have roles *in vivo* at these enhancers, such as remodeling chromatin, that are not required for activating transcription in our plasmid-based assay. Consistent with the latter notion, the list includes 3 Fox family TFs (Foxo4, Foxp3, and Foxd1), which act as pioneer factors that open chromatin during genomic enhancer activation (72-74), and 2 NFI TFs (NFIX and NFIC) that interact with histones (75, 76)

and contribute to remodeling of nucleosome architecture (77, 78). The remaining 2 TFs (Myb and AP-2) have plausible roles in early adipocyte differentiation, regulating the final cell division (79, 80) and repressing an alternate cell fate (81) respectively.

Together, the mutagenesis data show that (i) for the majority of correlated motifs, disruption of a single TF motif site has a strong effect on transcription, while (ii) the remaining correlated motifs, although not necessary for reporter enhancer activity in adipocytes, may contribute to transcriptional activation in the genome through chromatin remodeling or during different stages of development.

TF motifs drive expression when inserted into new contexts

The deletion analysis above revealed which motifs are required for expression. We next investigated the sufficiency of these correlated motifs for driving transcription when inserted into a new sequence context.

First, we substituted binding sites for each motif into existing motif sites in bound genomic sequences with strong activity in our assay. The resulting MPRA pool (Pool 6, "motif-substituted enhancers") contained each of the 38 significantly correlated TF consensus motifs substituted into 95 distinct locations, yielding a "matrix" of 3,160 enhancer constructs (**Fig. 3A,** see Methods).

Second, we added binding sites for 15 of the positively correlated motifs into sequences with low baseline activity, chosen from genomic regions that are bound by PPARγ in macrophages but not in adipocytes (46), and that contain a central PPARγ motif site but none of the other positive or negative TF motifs identified above. The motif sites were added individually and in pairs to these templates in 9 different

configurations, with 2, 4, or 6 total sites (**Fig. 3A,** see Methods). This pool (Pool 7; "synthetic enhancers") contained 4,324 sequences.

For each construct in Pools 6 and 7, we calculated the "incremental enhancer activity" of the motif-substituted enhancer relative to its background sequence.

The relative strengths of the motif sites in driving transcription in the substituted and synthetic enhancers were highly consistent with the relative effects of their disruption measured previously. Of the 27 motifs whose sites caused significantly reduced expression when mutated, 25 led to increased expression when substituted into the bound genomic sequences, whereas the 6 TF motifs that did not significantly reduce expression when disrupted also did not affect expression when substituted (**Fig. 3C**; SI Appendix, **Fig. S7A, Table S2**). The 5 negatively correlated motifs detected in the native enhancers strongly reduced transcription in the substituted enhancers. Similarly, synthetic enhancers containing motif sites associated with reduced expression in the mutated enhancers had activity significantly above the background level in most cases (7 of 9), whereas synthetic enhancers containing motif sites not associated with reduced expression did not (6 of 6 cases)(**Fig. 3D**). Moreover, the average quantitative effects of the motifs on enhancer activity were highly concordant in the substituted and synthetic enhancers (Spearman ρ=0.85, **Fig. S7C,D**, SI Appendix, **Table S2**).

Overall, 70% of these synthetic enhancers had higher transcriptional output than the template. Moreover, the proportion increased with the number of copies of the motif: 54% with 2 motifs, 73% with 4 motifs, and 81% with 6 motifs (**Fig. S7E**). Thus, these motifs in combination with a central PPARγ motif are sufficient to drive expression independent of positioning and sequence background.

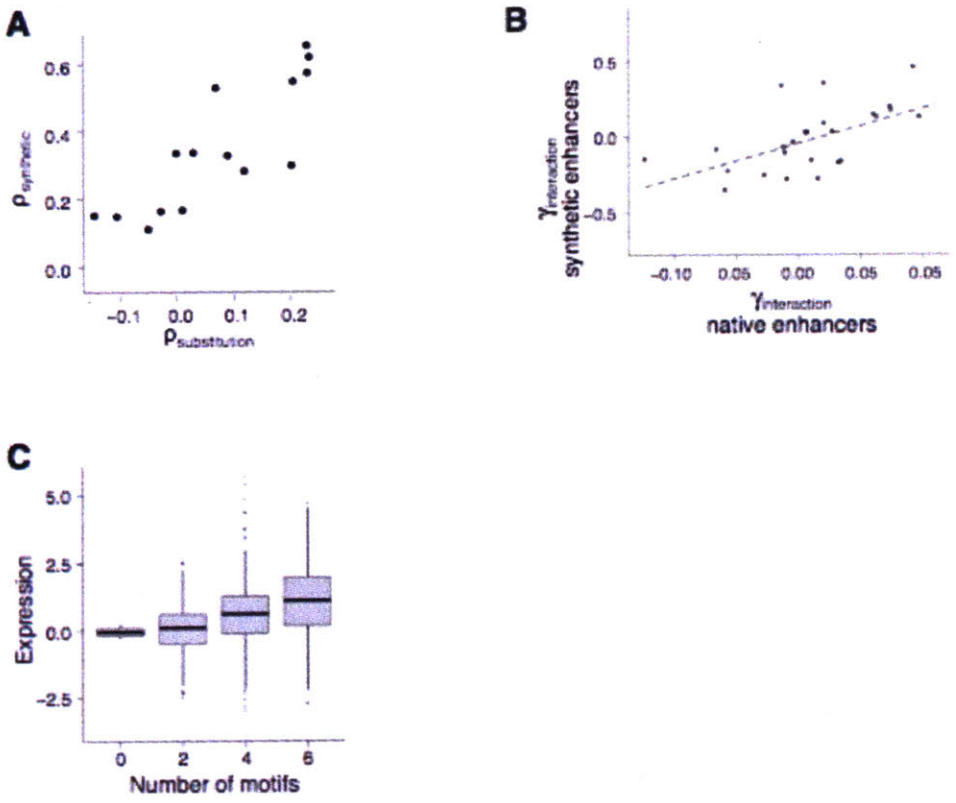


Figure S7. Effects of TF motif disruptions and insertions is consistent. (A) Comparison of quantitative effect ($\rho_{expression}$) of TF motifs on activity in the substituted (x-axis) and synthetic (y-axis) enhancers (Spearman correlation=0.85). **(B)** Comparison of non-additive interaction effects (γ) in native and synthetic enhancers. Only significant interactions ($p_{F\text{-}test}<0.01$ in native enhancers) are shown. **(C)** Expression of synthetic enhancers (normalized to template) conditional on number of sites.

The composition of TF motif sites predicts quantitative expression levels

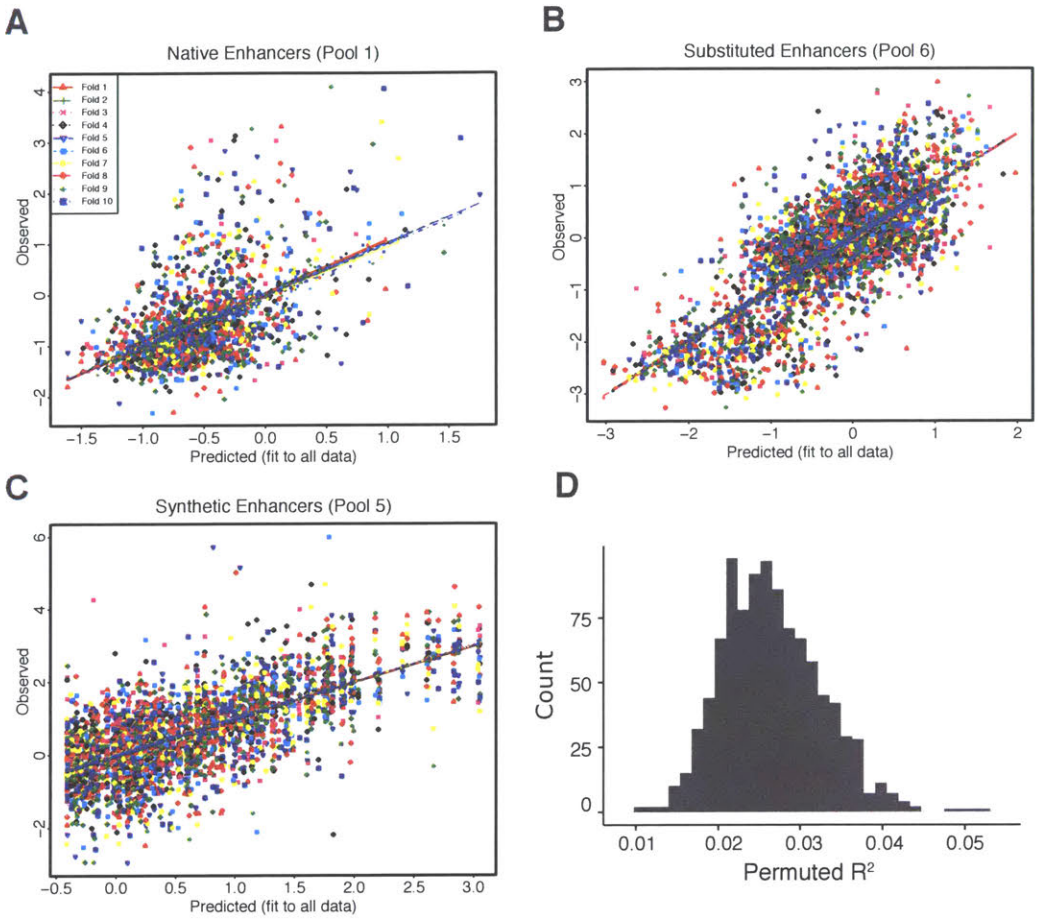
We next sought to explore how much of the quantitative enhancer activity could be explained by the composition of TF motif sites (i.e. number and identity) in the enhancers. To explore how well this model captures enhancer activities in the naturally occurring enhancers, we fit a linear regression (in which each motif contributes additively to the total expression level) to predict the log-transformed transcript levels associated with the original candidate enhancer from 750 bound and 750 unbound

genomic sites (Pool 1) based on the number of motif sites for each of the 38 TFs identified above, as well as overall GC content (which is often elevated at TF-binding sites (18, 82, 83)). Because some TFs have similar binding motifs, we removed redundant variables using stepwise variable selection to minimize the Akaike information criterion (AIC), and evaluated the performance of the selected model using 10-fold cross-validation.

The selected linear model included 23 of the 38 TFs and explained one-fifth of the variance in enhancer activity in the training dataset (cross-validation r^2=0.20, SI Appendix, Fig. **S8A,D-E**). The model explains 15% of the variance among the bound genomic sites alone, indicating it is not simply differentiating between the bound and unbound class. To validate the model, we created an MPRA pool (Pool 3) containing a new set of 750 bound motif sites and 750 unbound motif sites and measured their enhancer activity (SI Appendix, **Fig. S9A,B**). The model explained an equal amount of variance in enhancer activity (20%) among the test set (**Fig. 4A**). (We note that ~5% of the variance is due to inherent noise in the assay, as determined from comparison of biological replicates.)

Potential sources of the remaining variance include differences in motif-site affinities, interactions between motif sites, effects of motif positioning, and additional features in the background sequence below our detection limit.

We thus next sought to determine how much additional variance could be explained if we held the affinities of the motif sites constant and controlled for the activity of the background sequence. We fit linear models to predict the *incremental* enhancer activity (described above) of the enhancers with inserted motif sites (Pools 5



E

	First-order model				Model with interactions		
Pool	R2 in training data	R2 in 10-fold CV	R2 in test data	Number of coefficients	R2 in 10-fold CV	R2 in test data	Number of coefficients
Native	0.24	0.21	0.20	26	0.40	0.38	169
Synthetic	0.46	0.45		14	0.52		80
Substitution	0.57	0.56		33	0.74		136

Figure S8. Cross validation and performance of linear models of quantitative expression. (A-C) Predicted expression in 10-fold cross validation of native enhancers (A), substituted enhancers (B), and synthetic enhancers (C). **(D)** Histogram of R^2 values linear models fit on 1000 permuted datasets. **(E)** Performance of linear and Lasso models in each pool.

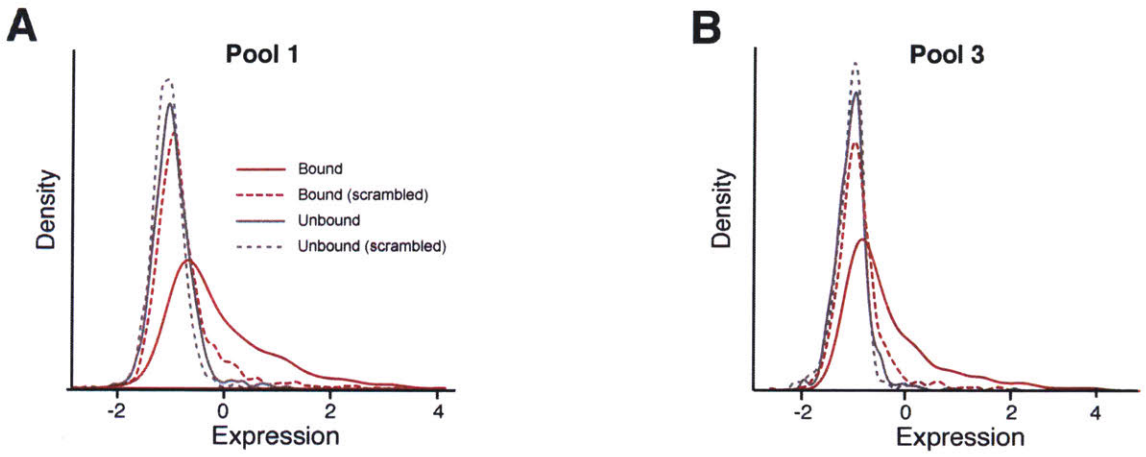


Figure S9. Enhancer activity of bound and unbound sequences is consistent in Pool 1 and Pool 3. Expression corresponding to bound and unbound genomic sites with intact and scrambled core PPARγ motifs in Pool 1 (**A**) and Pool 3 (**B**).

and 6) based on the changes in the number of motif sites relative to its background sequence. (As before, we removed redundant variables using stepwise variable selection, and evaluated the performance of the selected models using 10-fold cross-validation.) The selected models explained 56% of the incremental activity of the substituted enhancers with a single inserted motif site (Pool 5), and 45% of the incremental activity of the synthetic enhancers with pairs of sites inserted (Pool 6)(SI Appendix, **Fig. S8B-C**). These results show that motif affinity and background sequence contribute to variation in enhancer activity, but also suggest a substantial role for such features as non-additive interactions and motif positioning.

Interactions between motif sites explain additional quantitative variance in expression

Under our simple linear model, each motif site contributes a fixed amount to enhancer function, independent of the other sites in the immediate region and their arrangement. We next wondered if pairwise combinations of motifs could account for a

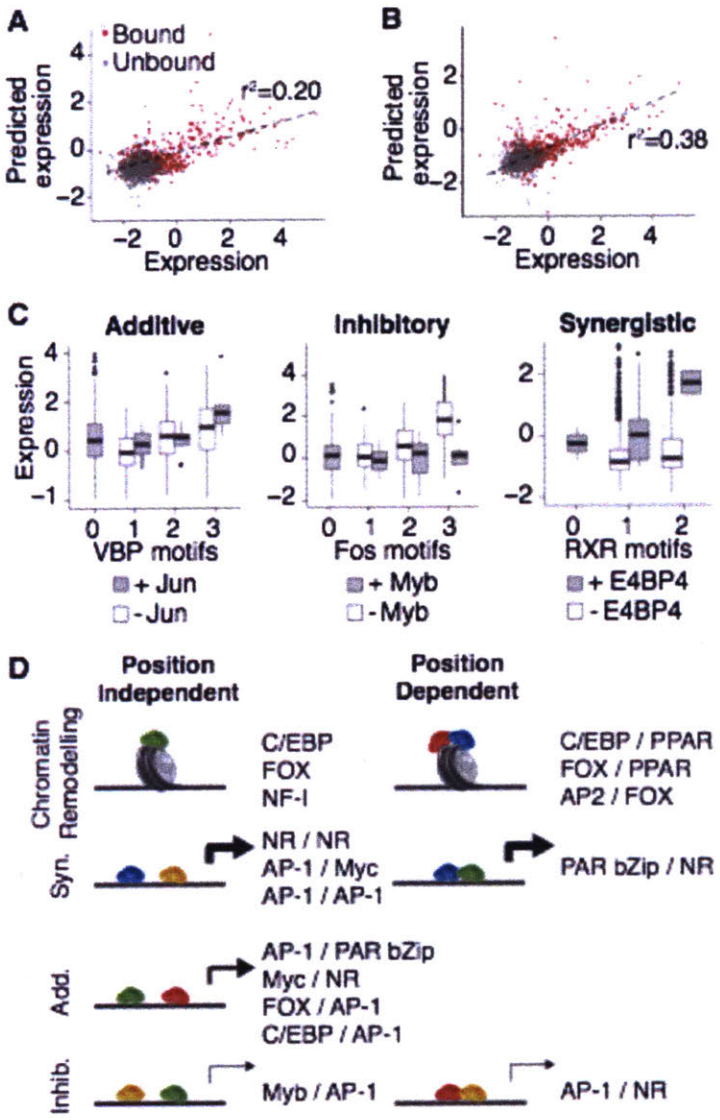


Figure 4. Interactions between motifs contribute substantially to enhancer activity. (A,B) Performance of linear model (A) and Lasso model (B) predicting expression levels based on motif counts in independent test dataset (Pool 3). (**C**) Boxplots represent expression of sequences in Pool 1 (synergistic plot, left) or Pool 5 (additive and inhibitory plots, left and center), conditioned on counts of the two motifs. **(D)** Modes of interaction between TFs. Pioneer factors (top) are required to open chromatin at enhancers in the genome, but do not contribute strongly to transcriptional activation. Some pairs of TF enhance each other's activity, resulting in super-additive transcriptional output (center-top). Other pairs of TFs function independently of each other, contributing additively to the transcriptional output (center-bottom). Finally, some TFs mutually inhibit each other's activity, resulting in sub-additive transcriptional output (bottom).

substantial fraction of the variance not explained by the simple additive model. To the linear models above, we added second-order interaction terms for motif pairs that co-occur in at least 10 sequences. Since only a minority of the potential interaction terms are likely to be relevant, we used a Lasso regression model, which selects sparse models, and optimized the tuning parameter by 10-fold cross-validation.

The models with interactions explained substantially more variance than the linear models for both the natural enhancers and the enhancers with inserted motifs. For the natural enhancers, the selected model included 73 out of 384 possible interaction terms, and explained 40% of the variance in the training data in 10-fold cross-validation (Pool 1) and 38% in the test data (Pool 3) (**Fig. 4B;** SI Appendix, **S10A, Table S4**). For the substituted and synthetic enhancers, the selected models respectively explained 74% and 52% of the incremental expression in 10-fold cross-validation (SI Appendix, **Fig. S10B,C**). The improvement in the performance of this model compared to the additive model suggests that TF interactions play an important role in the function of these enhancers to generate combinatorial enhancer activity. (Because the Lasso process involves some arbitrary choices among correlated variables, the specific terms in the model should not be regarded as a comprehensive list of biologically meaningful interactions. Below we consider interactions between specific TFs.)

Synergistic and inhibitory interactions occur in synthetic enhancers

To explore combinatorial interactions between specific TFs, we first focused on interactions present in synthetic enhancers discussed above, containing all pairwise combinations of motif sites for 15 TFs inserted into inactive template sequences with a

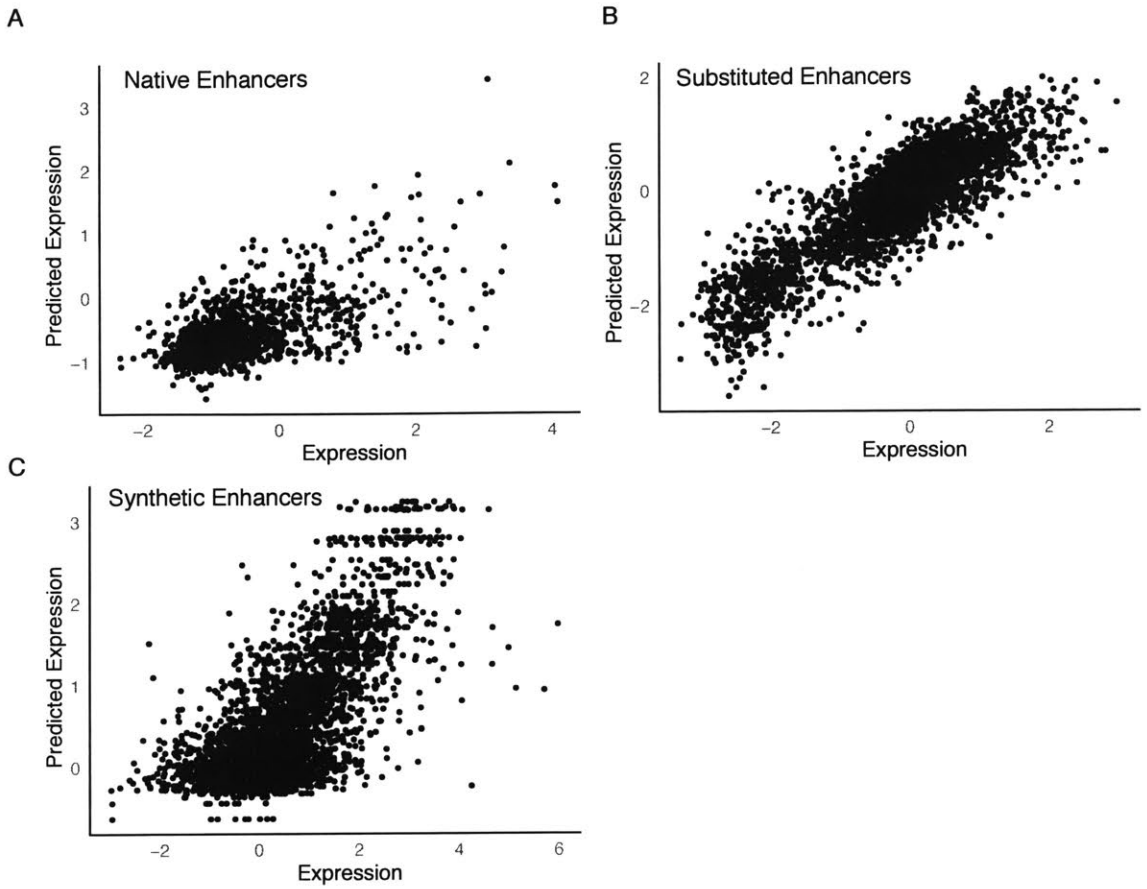


Figure S10. Cross validation and performance of Lasso models of quantitative expression. (A-C) Expression predicted by Lasso models in 10-fold cross validation of native enhancers (A), substituted enhancers (B), and synthetic enhancers (C).

central PPARγ motif. While motifs may co-occur in active native enhancers for a variety of reasons, only those pairs of TFs that functionally interact to drive enhancer function do not require specific positioning will be detected in our synthetic enhancers.

Using analysis of variance (ANOVA) to study interactions between pairs of TF motif sites, we identified 21 significant positive and negative interactions among the 15 TF motifs tested in the synthetic enhancers (Bonferroni-corrected $p_{F\text{-test}} < 0.01$; **Fig. 4C,D;** SI Appendix, **Fig. S11, Table S5,S6**). (For simplicity below, we refer to these as

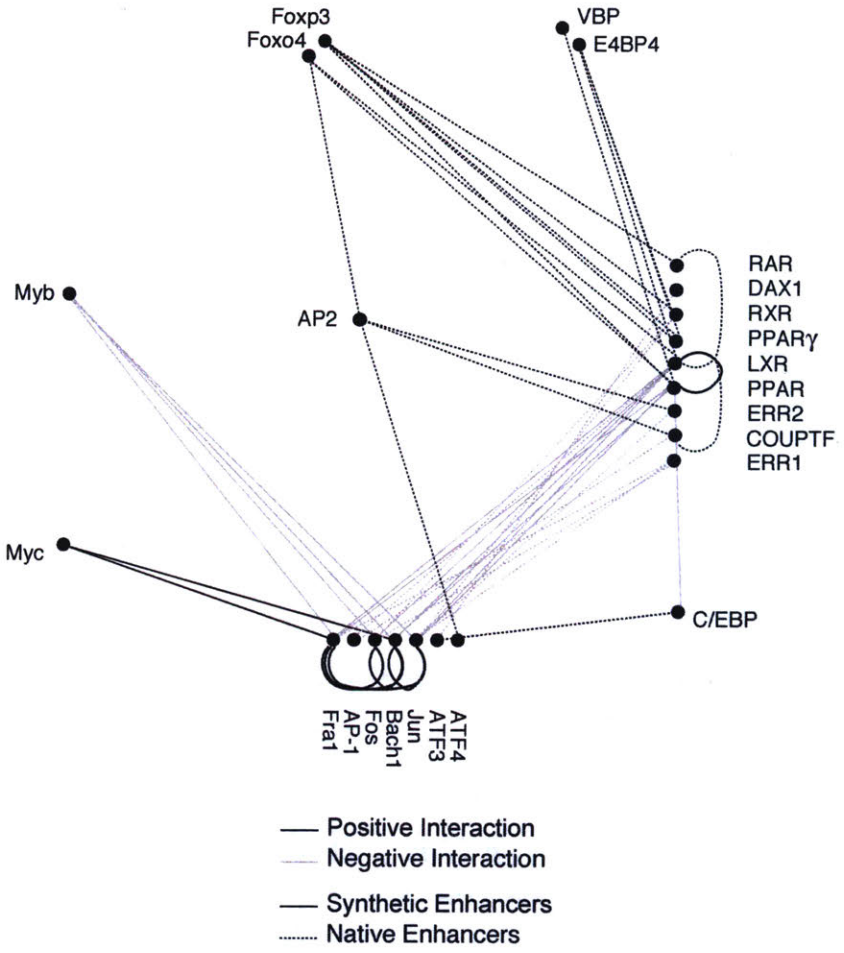


Figure S11. Graph of interactions between TFs. Graph showing all interactions identified in the synthetic enhancers (solid lines) and native enhancers (dashed lines). Red lines represent positive interactions, and blue lines represent negative interactions. TFs are group into structural families.

interactions between the TFs expected to being to the motif sites in adipocytes; however, we note that the presence of the motif site does not necessarily imply that the corresponding TF is bound in all cases.) These interactions fall into four main classes (**Fig. 4E**). The first two classes comprised synergistic interactions between various AP-1 family factors (6 pairs) and between AP-1 factors and Myc (2 pairs). AP-1 family members are strong activators that are able to interact with a wide range of other TFs

(84), and pairs of AP-1 factors are known to cooperatively induce DNA bending (85). The third class consisted of repressive interactions between AP-1 factors and nuclear receptors (7 pairs). Consistent with this result, AP-1 factors and the nuclear receptors PR, GR and ER have been shown to mutually inhibit each other's ability to activate transcription (86, 87). Finally, the fourth class contained repressive interactions involving Myb (4 pairs; **Fig. 4C**). Myb contains a repressive domain, and can function as a transcriptional repressor in some contexts (88). In the synthetic enhancers in our assay, Myb appears to act as a dominant repressor, returning transcription close to baseline levels. We also detected two interactions that did not fall into any of these four classes: a synergistic interaction between two nuclear receptors, PPAR and LXR, and a repressive interaction between C/EBP and PPAR.

To confirm the interactions, we examined the results from the mutated and substituted enhancer pools (Pools 5 and 6) to assess the effect of adding or deleting one of a pair of interacting motifs at several hundred sites in active naturally occurring sequences. For the majority of identified pairs, the effect of disrupting or inserting an interacting motif site differed significantly in sequences that contained the partner motif compared with sequences that did not (82% for motif disruptions and 86% for motif insertions; $p_{Wilcox}<0.0001$), and the direction of the effect was consistent with the interaction detected in the synthetic enhancers.

Finally, we tested whether the 21 interacting pairs showed significant interaction with respect to transcriptional activity of the naturally occurring enhancers (Pool 1). Eleven of the pairs (including most AP-1/AP-1, AP-1/Myc, and Myb/TF pairs) co-occurred too rarely in the native enhancers (frequency<0.01) to allow meaningful

analysis. For the remaining 10 pairs, we expected to have power to see 3-5 significant interactions (Bonferroni-corrected $p_{F\text{-}test} < 0.05$) based on the effect sizes and counts in the natural enhancers (see Methods). Consistent with this, we found 5 pairs with significant interaction terms, all of which had the same sign as the interaction in the synthetic enhancers (SI Appendix, **Table S5**).

Naturally occurring enhancers contain additional classes of interactions

We next examined native enhancer sequences (Pool 1) to identify additional TF interactions not detected in the synthetic enhancers. Such interactions might fall into three classes: (i) pairs of motifs that do not involve the 15 TFs used in creating synthetic enhancers on a neutral template, (ii) pairs involving TFs such as pioneer factors that are correlated in the genome with an effective enhancer but not necessary for expression in adipocytes, or (iii) pairs that require specific spacing and orientation, which were not imposed in the synthetic enhancers.

Among the 38 TFs correlated with expression, we detected 25 significant positive and negative interactions beyond those seen in the synthetic enhancers (**Fig. 4C,D**; SI Appendix, **Fig. S11**, **Table S5,S7**). They fell into each of the three classes. The first class included 11 pairs of inhibitory AP-1/NR pairs that were not tested in the synthetic enhancers. The second class (TFs not required for reporter expression) consisted of two groups: the first group (6 pairs) involves synergistic interactions between nuclear receptors and FOX TFs, which have pioneering ability (72-74); the second group (4 pairs) involves interactions between various TFs and AP-2, which may play a role in early adipocyte differentiation by repressing an alternate cell fate (81), but is down-regulated in terminally differentiated adipocytes.

To study the third class (spatially constrained interactions), we evaluated whether the two motifs occurred adjacently (<10 bp apart) more often than would be expected by chance in naturally occurring enhancers. Of the 25 interacting pairs, 8 showed significant enrichment of *adjacent* co-occurrences in bound PPARγ enhancers in the genome (Bonferroni-corrected $p_{Fisher}<0.01$; SI Appendix, **Fig. S11**). The 8 pairs are CEBP/ATF-3, AP-2/Foxo4, 3 AP-1/NR pairs, and 3 FOX/NR pairs. The first pair, for example, is known to bind as a heterodimer to composite motifs in the genome (89, 90), suggesting that the enriched configurations reflect functional physical interactions. AP-1 and NRs both directly interact with CBP/p300 to activate transcription (91), and could interfere with each other's interaction when in close proximity. FOXO family TFs have also been shown to physically interact with a number of nuclear receptors, often in a ligand-dependent manner, resulting in changes in the activity of the two TFs (92).

Of the 5 adjacent pairs that include an asymmetric motif, 4 were enriched for a specific orientation of the two motifs relative to each other. Interestingly, 8 of the interactions detected in the synthetic enhancers were also biased towards a specific configuration in the natural enhancers, suggesting that these pairs may interact more efficiently in one orientation.

DISCUSSION

Deciphering the regulatory code of enhancers requires understanding how the combinatorial input of different TFBS lead to precise TF-binding patterns and gene expression outputs. Here, we use a series of MPRA experiments, involving 32,115 distinct enhancer constructs, to systematically evaluate the factors that govern PPARγ binding and regulation in adipocytes. We demonstrate that (i) the PPARγ motif affinity

(and not cooperative elements in the immediately flanking sequence) largely determines PPARγ binding to genomic sequences when removed from their chromatin context; (ii) enhancer activity depends not only on PPARγ binding, but also on a network of 20-30 TF motifs in the flanking sequence that have distinct quantitative contributions to expression; and (iii) various pairs of motifs interact in additive, inhibitory, and synergistic ways with varying constraints on motif positioning. Although in this study we measured enhancer activity in an episomal context, a recent study found that enhancer activity was highly concordant between episomal and genomic contexts (r=0.86 across 2,236 candidate enhancers vs. 0.90-0.98 for replicates within each context) (93). Importantly, our results show that PPARγ binding and enhancer activity are independently regulated.

Studies of several TFs, including PPARγ, have observed strong correlations between DNA accessibility and TF binding, leading to the hypothesis that TF binding for non-pioneer factors is largely governed by nucleosomes or the larger chromatin landscape (31-33, 36, 51, 52, 94). In this model, pioneer factors bind to sites in closed chromatin and displace surrounding nucleosomes, allowing other TFs to bind to neighboring sites, which may then reinforce nucleosome exclusion. Our results support this model, demonstrating that, in an episomal context, both bound and unbound genomic motif sites bind PPARγ equally well (excluding the possibility of latent features controlling motif affinity) and that binding is largely independent of sequences immediately surrounding the PPARγ motif (excluding a major role for direct cooperative binding). While the presence of H3K27ac at bound PPARγ sites has been widely appreciated, our results suggest a graded effect even among open sites, whereby stronger quantitative chromatin accessibility is associated with more frequent TF

occupancy. This relationship appears to be general to other TFs: we see a similar quantitative correlation between quantitative H3K27ac signal and TF-binding for nearly all of the TFs and cell lines profiled in the ENCODE Project.

Our study identifies a collection of ~20 TF motifs that are correlated with higher enhancer activity in naturally occurring enhancers and, with the exception of pioneer factors, play direct roles in enhancer activity (as assayed by mutational perturbation). Intriguingly, the TF motifs that affect enhancer activity correspond closely to those that are most enriched in the genomic sequences of PPARγ binding sites in adipocytes. If this observation can be confirmed for some additional TFs and cell types, it may allow the use of motif co-occurrences in the genome to be used to predict the functional activities of TFs.

While cooperative binding of TFs to composite motif sites has been studied in depth, much less is known about how sets of TFs, once bound, influence gene expression. Characterizing such interactions is difficult due to the large number of possible combinations and uncertainty about spatial constraints. By choosing a related set of enhancers and focusing on TFs that correlate individually with enhancer activity, our approach yields a tractable number of potential interactions for functional characterization, facilitating the identification of a basic set of grammatical rules governing the activity of these enhancers.

Using this approach, we detect examples sub-additive, additive, and super-additive interactions between different pairs of TFs with varying degrees of spatial constraint. The identified interactions fall into several classes, comprising TFs from specific structural families that interact similarly with TFs from other families. We

reproduce several types of interactions supported by previous studies, such as mutual inhibition of nuclear receptors and AP-1 factors and synergistic interactions between pairs of AP-1 factors (84, 86), and identify intriguing new interactions, such as quenching of enhancer activity by Myb. Our results highlight the need for better understanding of the molecular and biochemical basis of TF activity in order to understand the mechanisms underlying TF cooperativity and combinatorial transcriptional activation.

The approach described here provides a framework to dissect the regulatory grammar underlying enhancer function by (i) systematically identifying TFBS correlated with activity, (ii) isolating the independent quantitative contributions of each TF to enhancer output by disrupting and inserting binding sites in controlled contexts, (iii) identifying interactions between TFBS in synthetic and natural enhancers that influence enhancer activity, and (iv) characterizing spatial constraints on the identified interactions. Our approach is readily applicable to many TFs and cell types to understand the regulatory grammar of diverse sets of regulatory elements, revealing the prevalence and generality of these rules. Understanding this regulatory code is critical in understanding how gene expression drives fundamental biological processes such as differentiation and development, as well as interpreting the increasing number of variants in regulatory regions implicated in cancer and other diseases.

ACKNOWLEDGEMENTS

We thank Aviv Regev for valuable advice regarding the design and interpretation of the experiments, and Brian Cleary, Tim Wang, Mitch Guttman, Chris Burge, members of the Lander Lab and numerous colleagues for helpful comments and discussion. This

work was supported by the National Human Genome Research Institute (2U54HG003067-10) (E.S.L.), the National Institute of General Medical Sciences (T32GM007753) (S.R.G.), and the National Institute of Health (R01 HG006785) (T.S.M.), as well as funding by SystemsX.ch (B.D.) and internal support by the EPFL (B.D.)

MATERIALS AND METHODS

Motif Analysis

All vertebrate motifs from the TRANSFAC (Matys et al., 2006) and JASPAR (Mathelier et al., 2014) databases were included in analysis. Motif sites were identified by running FIMO (Grant et al., 2011) with a p-value threshold of 10^{-4}. In cases where the combined databases contained multiple PWMs corresponding to a single transcription factor, we merged overlapping motif sites before calculating motif counts. The motif instance with the highest PWM score for each TF was used as the consensus motif in Pools 6 and 7.

Pool Design

Pools 1 and 3

PPARγ motif sites in the mouse genome were identified using the 16-bp PPARγ motif reported in (Mikkelsen et al., 2010) with a p-value threshold of 10^{-4}. To create the oligonucleotide pool, we randomly selected 750 PPARγ motif sites that had significant (FDR<0.01) ChIP enrichment in 3T3-L1 adipoctyes (Mikkelsen et al., 2010). For each bound site, we selected a matching unbound PPARγ motif site in the genome with the same 16-bp sequence that showed no PPARγ ChIP enrichment. 92% of the bound sites were in regions with open chromatin marks (significant ChIP enrichment for H3K4me1/2/3 or H4K27ac), whereas only 5% of the unbound sites were in open

regions. 145-bp regions centered on each of the 1500 motif sites were included in the pool. To create the scrambled motif sequences, the core 16-bp motif was disrupted by swapping A↔T and G↔C.

Pool 2

To create Pool 2, we selected 25 of the bound motif sequences included in Pool 1. The central 16-bp PPARγ motif sequence was replaced with the central 16-bp motif sequence from each of the other 24 sequences, yielding a pool of 625 sequences. The *in vitro* affinity for PPARγ of the 25 central 16-bp motif site flanked by 5 bp on each side from the original genomic sequence was determined using MITOMI with recombinant PPARγ (as in (Isakova et al., 2016)).

Pool 4

To create Pool 4, we performed scanning mutagenesis across 25 bound sequences (the same as were included in Pool 2). We used 3 mutagenesis strategies. The first consisted of bound sequences with 10-bp blocks scrambled (by swapping A↔T and G↔C), tiled every 5 bp across the sequences. The second consisted of bound sequences with 20-bp blocks (tiled every 5 bp) replaced with 20-bp blocks from corresponding position in the matched unbound sequence (see Pool 1). The third consisted of the unbound sequences with 20-bp blocks (tiled every 5 bp) replaced with 20-bp blocks from corresponding position in the matched bound sequence.

Pool 5

To generate Pool 5 sequences, we mutated each motif instance of the 38 correlated motifs (1803 total sites) by swapping A↔T and G↔C in the 375 most active bound sequences from Pool 1, yielding 1803 total variants. For each motif site, we also created

a control by mutating an equally-sized block in the sequence that did not overlap any of the correlated motif sites.

Pool 6

To create Pool 6, we randomly selected 15 bound sequences with activities in Pool 1 above the median. We identified 95 total instances of the 38 correlated motifs in these sequences. At each of the 95 sites, we replaced the original motif instance with a consensus instance of the 38 correlated motifs, yielding 3610 variants. The variant sequences were trimmed or padded with random bases to a length of 150 bp.

Pool 7

We chose 4 template sequences from the mouse genome that (i) contained a PPARγ motif site in the center, (ii) were reported to have significant PPARγ ChIP enrichment in cultured macrophages (Lefterova et al., 2010) but not in 3T3-L1 adipocytes (Mikkelsen et al., 2010), and (iii) did not contain any instances of the 38 correlated motifs. We created synthetic enhancers by adding 2, 4, or 6 motif sites to each of the 4 templates in 9 total configurations. First, we created synthetic enhancers containing multiple copies of a single motif, using consensus motif sequences for 15 of the positively correlated motifs (PPAR, LXR, ATF3, Jun, Fra1, Bach1, VBP, E4BP4, Myc, Myb, AP-2, Fos, C/EBP, Foxd1, and Foxo4), yielding 540 total enhancers. Next, we created synthetic enhancers containing each pairwise combination of the 15 motifs (105 total pairs), yielding 3780 total enhancers.

Oligonucleotide library design and synthesis

Oligonucleotide libraries for MPRA Pools 1-4 were designed as in (Melnikov et al., 2012) to contain (in order) the universal primer site ACTGGCCGCTTCACTG, the 145-

bp test sequence, KpnI/XbaI restriction sites (GGTACCTCTAGA), a variable 10-bp tag sequence, and the universal primer site ATCGGAAGAGCGTCG. Sequences from Pools 1, 2, 3, and 6 linked with 9, 20, 13, and 15 unique tags (respectively) in order to reduce variance due to stochastic rates of amplification and transfection of specific plasmids.

Oligonucleotide libraries for Pools 5-7 were designed to contain (in order) the universal primer site ACTGGCCGCTTGACG, the 145-bp test sequence, the universal primer site CACTGCGGCTCCTGC. After synthesis, a variable 20-bp tag sequence and additional adapter sequences were added by performing emulsion PCR reactions with Q5 DNA Polymerase (NEB) and the primers MPRA_v4_amp_F and MPRA_v4_amp_R. Amplified emulsion reaction mixture was broken and purified using AMPure XP SPRI beads (Beckman Coulter).

Oligonucleotide pools were synthesized by Agilent, Inc.

MPRA plasmid construction

Full-length oligonucleotides were isolated by running the libraries on 10% TBE-Urea denaturing polyacrylamide gel. The purified oligonucleotides were amplified by PCR using Herculase II Fusion DNA Polymerase and the primers GCTAAGGGCCTAACTGGCCGCTTCACTG and GTTTAAGGCCTCCGTGGCCGACGCTCTTC. Purified PCR products were digested with *SfiI* and ligated into the *SfiI*-digested MPRA vector pGL4.10M. The ligation reaction was transformed into One Shot TOP10 Electrocomp *E. coli* cells (Invitrogen) with transformation efficiency >3 x 10^8 cfu/ug. The isolated plasmid library was then digested with *KpnI* and *XbaI* to cut between the test sequence and tag, and ligated to

the luc2 ORF fragment isolated from pGL4.10 by *KpnI-XbaI* digestion. The ligation reaction was transformed into *E. coli* as described above and the plasmid library isolated. To remove intermediate constructs, the libraries were digested with KpnI, size-selected on 1% agarose gel, self-ligated, and transformed into *E. coli* as described above.

Cell culture and transfection

3T3-L1 preadipocytes (ATCC CL-173) were grown in DMEM supplemented with 10% calf serum. Two days after confluence, the medium was replaced with DMEM supplemented with 10% FBS, 1 uM dexamethasone, 1.7 uM insulin, and 0.5 mM 3-isobutyl-1-methylxanthine (IBMX) to induce differentiation. 48 hours later, the medium was replaced by DMEM supplemented with 10% FBS, and changed every 2 days. For the transfections, 100 ug of pool DNA was introduced into 2.2 x 10^7 cells using a Nucleofector 96-well Shuttle System device with SE Cell Line Nucleofector Kits and program DS-137.

ChIP-seq

16 h post-transfection, the cells were treated with 1% formaldehyde for 10 min at 37C and stored at -80C. Crosslinked cells were resuspended in Cell Lysis Buffer (0.25 M sucrose, 20 mM Tris-HCl pH 8.0, 85 mM KCl, 0.5% NP40, 0.5 mM DTT) and homogenized with a dounce homogenizer. To extract the nuclei, the homogenate was centrifuged at 1000g for 10 min at 4C and the nuclear pellet was washed twice with Cell Lysis Buffer. Nuclei were lysed by incubation for 10 m in Nuclear Lysis Buffer (10 mM Tris-HCl pH 7.5, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS) at 4C. To remove

contaminants, lysate was sonicated for 2 m using a Branson instrument and spun for 10 m at maximum speed. ChIP for PPARγ was performed as described in (Mikkelsen et al., 2010). Plasmid library DNA was amplified from ChIP fragments by PCR (25 cycles) with Herculase II DNA polymerase and the primers CTAACTGGCCGCTTCACTG and AGCTAGCGAGCTCAGGTACC, which hybridize to constant regions on the MPRA plasmid backbone. The PCR fragments were size-selected on 2% agarose gel. Illumina sequencing libraries were constructed from purified PCR fragments as described in (Mikkelsen et al., 2010). Libraries were sequenced using 150-bp paired-end reads on an Illumina MiSeq instrument.

Reads were mapped to plasmids using BWA 0.7.12 (Li and Durbin, 2009) with default parameters. Reads with MAPQ<20 were removed.

Analysis of genomic TF binding

PPARγ: We defined "open chromatin regions" as regions called as FAIRE-seq peaks by Waki et al and significantly enriched in ChIP for at least one activating chromatin mark (H3K27ac, H3K4me1/2/3). For quantitative analysis, we used the maximum quantitative FAIRE-seq signal and the total ChIP enrichment within 200 bp centered around the PPARγ motif. To predict binding, we fit a logistic regression using the 4 activating chromatin marks as well as H3K27me.

ENCODE: We defined "open chromatin regions" as regions called as DNaseI hotspots ("Uniform DNase Hotspots"). For all sequence-specific TF profiled in ENCODE that had a cognate motif in Transfac or JASPAR, we identified genomic motif instances using FIMO ($p<10^{-4}$). At each motif instance, we used the maximum DNase score reported by ENCODE within 200 bp of the motif site, and calculated ChIP enrichment for H3K27ac,

H3K4me1, and K3K4me3 over the same window. To define TF-bound sites, we used the ChIP peaks reported by ENCODE (narrowPeak).

Tag-Seq

The cells were lysed 16 h post-transfection with RLT buffer (Qiagen) and frozen at -80. Total RNA was isolated using RNeasy Kit (Qiagen), and mRNA was isolated from total RNA using MicroPoly(A)Purist Kit (Ambion). To remove any plasmid contamination, mRNA was treated with DNase I using the Turbo DNA-free kit (Ambion). First-strand cDNA was synthesized using the SuperScript II First-Strand Synthesis kit (Life Technologies). Tag-seq libraries were generated from cDNA and plasmid libraries by PCR (26 cycles) with Pfu Ultra II HS DNA polymerase and the primers AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT and CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGAGGTGCCTAAAGG, where XXXXXXXX represents a library-specific barcode. The resulting ~250 bp PCR product was isolated on 2% agarose gel. The libraries were sequenced using 36-nt single-end reads on an Illumina HiSeq 2000 instrument.

Barcode/Oligo mapping

To determine the oligo/barcode combinations for Pools 4 and 5, sequencing libraries were prepared from the intermediate plasmid library (prior to SfiI digest) by PCR using the primers TruSeq_Universal_Adapter and MPRA_v3_TruSeq_Amp2Sa_F. Sequencing indexes were added by PCR with the same primers used in the Tag-seq

library preparation. Samples were sequenced using paired-end 150 bp reads on an Illumina MiSeq instrument.

Paired-end reads were aligned to the oligonucleotide pool using Bowtie2 2.1.0 (Langmead and Salzberg, 2012), trimming the tag sequence and adapters. Read pairs that were not aligned in proper pairs, had mapping scores less than 25, or edit distance greater than 5 were discarded. Remaining oligo/barcode pairs were merged. Barcodes that were observed in only 1 read or that matched multiple test sequences were removed from further analysis.

Data processing and normalization

ChIP-Seq sequence reads were aligned to the plasmid library sequences using BWA (Li and Durbin, 2009) with the BWA-backtrack algorithm and default parameters. Read pairs that aligned to more than one test sequence in the plasmid library or had mapping quality scores <20 were discarded. All test sequences that did not have a count of at least 50 in the input were discarded. Relative PPARγ binding scores were generated by calculating the (log2) ratio of ChIP reads to input reads mapping to each test sequence. mRNA and plasmid counts for each tag sequence were calculated from all sequence reads whose first 10 nt or 16 nt perfectly matched one of the tags and remaining nucleotides matched the expected upstream MPRA plasmid sequence (regardless of quality score). All tags that did not have a count of at least 50 in the plasmid pool were discarded. To generate relative activity scores for each test sequence, we calculated the median (log2) ratio of mRNA counts to plasmid counts for all tag corresponding to that sequence.

Gene Expression

To calculate gene expression for each transcription factor, we used mRNA expression data generated using GeneChip arrays (Affymetrix) from (Mikkelsen et al., 2010). For TFs that corresponded to multiple genes, we used the maximum expression value.

Regression models

We fit models to predict quantitative expression in the native (Pool 1 & 3), substituted (Pool 6) and synthetic enhancers (Pool 7). In the native enhancer pools, we predicted enhancer activity ($\log_2$[RNA/DNA]) from motif counts of each correlated TF (38 total). In the synthetic pool, we predicted fold change ($\log_2$) over the template sequences from the motif counts of each tested TF (15 total). In the substituted enhancers, we predicted fold change in activity relative to the wild-type enhancer ($\log_2$) from the change in motif counts between the wild-type and substituted enhancer.

Quantitative enhancer activity was modeled as

$$Y = \beta_0 + \beta_{gc} X_{gc} + \sum_i \beta_i X_i$$

(linear model) or

$$Y = \beta_0 + \beta_{gc} X_{gc} + \sum_i \beta_i X_i + \sum_{i,j} \gamma_{ij} X_i X_j$$

(Lasso model)

where Y are the activities of the enhancers ($log_2 \frac{reads\ in\ RNA}{reads\ in\ DNA}$), X_{gc} is the fraction of G and C bases in the sequence, and X_i are the counts of motif i in each enhancer. To remove redundant variables in the linear model (e.g., TFs in the same structural family (e.g. AP-

1, NF-I) that have similar motifs), we fit used forward and backward stepwise regression using the R package MASS, and selected the model with the minimum AIC. To evaluate the model's performance, we used 10-fold cross validation.

Since we expect only a fraction of the possible interaction terms to contribute to expression, we fit a model with pairwise interaction terms using a Lasso regularized regression, which selects sparse solutions (Tibshirani, 1996), using the R package glmnet. The tuning parameter (λ) for the lasso regularization was chosen by 10-fold cross validation to minimize the MSE. To evaluate the model's performance, we used predictions from the 10-fold cross validation.

Analysis of TF interactions

To identify significant pairwise interactions, we used analysis of variance (ANOVA) to test for a significant interaction between each pair of motifs. Quantitative enhancer activity was modeled as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_{1,2} X_1 X_2$$

where Y are the activities of the enhancers ($log_2 \frac{reads\ in\ RNA}{reads\ in\ DNA}$) and X_i are the counts of motif i in each enhancer. The significance of the interaction term was calculated using an F-test, and the p-values were corrected for multiple testing using a Bonferroni correction.

To calculate the expected number of interactions detected in the synthetic enhancers (Pool 5) that would appear significant in the native enhancers (Pool 1), we simulated interactions in the native enhancers with the effect size observed in the synthetic enhancers. For each pair, we created 1,000 simulated datasets. We first calculated

baseline activities by subtracting the effect of the two individual motifs and the interaction effect

$$Y_{base} = Y - \beta_1 X_1 - \beta_2 X_2 - \gamma_{1,2} X_1 X_2$$

The baseline activities were permuted across sequences, and then the effect of the two individual motifs and the interaction effect were re-added to the permuted activities

$$Y_{sim} = Y_{base,permuted} + \beta_1 X_1 + \beta_2 X_2 + \gamma_{1,2} X_1 X_2$$

For each simulated dataset, we tested the significance of the interaction, and calculated the probability of detecting the interaction across all simulations. The probabilities for each pair were added to determine the expected number of significant interactions that would be detected in the native enhancers.

ACCESSION NUMBERS

The accession number for the sequencing data reported in this paper is GSE84888.

SUPPLEMENTAL NOTE

Role of identified TFs in adipogenesis

Several of the other positively correlated motifs correspond to TFs that are known pro-adipogenic regulators or are otherwise associated with adipocytes. The most strongly correlated motif after PPARγ is C/EBP, a key regulator of adipogenesis that can act as a pioneer factor and tends to co-localize with PPARγ across the genome (Farmer, 2006; Lefterova et al., 2008; Siersbaek et al., 2012). ATF3 and ATF4 interact with C/EBP and are also required for adipogenesis (Maekawa et al., 2010; Mann et al., 2013; Shuman et al., 1997; Vinson et al., 1993). The Fos-Jun family transcription factors (JunB, JunD, Bach1, Fra1, and Jdp2) are induced during adipogenesis, and, when overexpressed in mice, cause lipodystrophy (Luther et al., 2011). ZEB-1, whose motif is

negatively correlated with expression, is also required for adipogenesis (Gubelmann et al., 2014). The list also the NF-I family factors NF-IX and NF-IC, all of which can promote adipocyte differentiation (Li et al., 2009; Nakae et al., 2003; Seo et al., 2004; Waki et al., 2011; Zhu et al., 2010); Foxd1, which regulates adipocyte-specific genes (Dahle et al., 2002); and c-Myc and c-Myb, which are involved in cell-cycle control and play an important role in terminal adipocyte differentiation (Cornelius et al., 1994; Sarruf et al., 2005). Another positively correlated factor, AP-2, represses chondrogenesis, which is an alternative fate for pre-adipocyte fibroblasts (Wenke and Bosserhoff, 2010).

Several of TF motifs negatively correlated with expression correspond to known anti-adipogenic and pro-chondrogenic factors, and transcriptional repressors. Snai1 and Snai2 are transcriptional repressors, with Snai1 having been shown to directly inhibit the adipocyte-specific gene adiponectin by binding to its promoter (Park et al., 2012). Sox family factors are key positive regulators of chondrogenesis (which, as noted above, is an alternative fate for the precursors of adipocytes) (Bi et al., 1999).

TABLES

Table S1. In vitro affinity measurements of PPARY motif sequences

Bound Motif Site	5-flanked sequence	Bound	Kd Bound (nM)
chr1:157643102-157643118	TGTCATTGGGGTGAGAGGTGAGAAGG	6.64	280.835
chr3:41089097-41089113	TGACAACATGCACAAAGGTCAAATTC	4.41	223.426
chr8:90491392-90491408	AAGGCACTAGGCCAGAGTTCAGAGCC	9.00	63.5193
chr3:30881175-30881191	AGCTATCTGGGCAAAAGGGTACAGTG	5.84	318.085
chr7:128189269-128189285	ATTAATGTGGGCAAGAGGTCATCTAT	6.11	55.5924
chr6:73173299-73173315	TGCCAAATGGGGCAAAGTTGAAATTC	3.28	100.598
chr19:38841001-38841017	AACCAGGTAGGGCAAAGGGGAGAGCC	6.53	395.6
chr3:115883631-115883647	CTGAATCTGGGACAAAGGTTATAAAC	3.61	56.4884
chr15:96709419-96709435	TAGAAAGCAGGGCAAAGGGAAGCCTG	6.54	157.007
chr7:135093082-135093098	GAAAATGTGGGTAAAAGTGCAACATA	34.99	93.7941
chr16:59639567-59639583	TAAGAACTAGGGACAGGGTCACACTT	7.27	1154.68

chr12:56575161-56575177	ACTTATTTAGGGCAAAGAGCATTGTG	4.09	131
chr4:86543397-86543413	AGCCTACTAGAGCAAGAGTCACCAGA	4.34	177.457
chr1:129886515-129886531	TGGAAACTGGAGGAAAGGTCACTTTG	3.34	37.8572
chr7:137517925-137517941	TTGAAATCATGTCAAAGTGCAAGATT	2.66	186.2
chr4:45013054-45013070	GAAGCACTGTGATAAAGGTCTTCCCT	2.46	279.805
chr4:59598223-59598239	TGTATACCAGGCAAAGGGGCAAGCCC	5.40	1692.44
chr11:69475159-69475175	TGCCCACCAGGTCAAAGATTCAGTGG	1.56	971.038
chr14:34124086-34124102	TTTTTAGCAAGGCAGAGGGAATCAGG	6.83	5162.82
chr6:24582317-24582333	CAAGCTGTGGGGCAAAGGGTATCTGT	3.26	209.318
chr3:116265904-116265920	CTGTGACCAAGTGAAAAGTCATGTCA	1.46	907.873
chr6:13529561-13529577	TGGTAACTGGGACAAATGTCACAGCA	1.60	87.1463
chr5:30496953-30496969	TGCCAACTAGGTTAAAGGTAAACTCC	0.89	70.105
chr3:145812073-145812089	GTTTAGCTAGGGCAAAGGACAGTTTG	0.62	1186.67

Table S2. TF motifs correlated with PPARγ enhancer activity

			Pool 1			Pool 3	
Factor	TF Expression in Adipocytes	TF Family	Correlation (ρ)	Correlation P-value	Total Sites in Pool	Correlation (ρ)	Correlation P-value
AP-1	+++	AP1	0.11	4.E-05	293	0.08	3.E-06
AP-2	+++		0.09	2.E-04	794	0.02	2.E-01
Zeb1	+++	Repressor	-0.09	6.E-04	176	-0.07	2.E-04
ATF3	+	ATF	0.09	4.E-04	76	0.04	4.E-02
ATF4	+++	ATF	0.09	9.E-04	68	0.08	1.E-05
Bach1	+++	AP1	0.10	6.E-05	75	0.07	2.E-04
C/EBP	+++		0.19	6.E-14	653	0.05	4.E-03
COUP-TFI	+++	NR	0.13	2.E-07	2067	0.07	2.E-04
COUP-TFII	++	NR	0.10	2.E-04	378	0.06	2.E-03
NFIX	+++	NF	0.11	1.E-05	134	0.03	2.E-02
DAX1	-	NR	0.09	2.E-04	308	0.06	2.E-03
E4BP4	++	PAR-bZip	0.09	4.E-04	29	0.09	2.E-07
ERR1	++	NR	0.10	1.E-04	414	0.04	4.E-02
ERR2	-	NR	0.10	7.E-05	198	0.05	7.E-03
ERR3	-	NR	0.10	5.E-05	168	0.05	1.E-02
Fos	+++	AP1	0.09	4.E-04	58	0.07	4.E-05

FOXD1	+	Repressor	0.08	1.E-03	33	0.03	1.E-01
FOXO4	+	fox	0.09	5.E-04	129	0.02	3.E-01
FOXP3	-	NR	0.09	5.E-04	131	0.03	8.E-02
FRA1	+	AP1	0.09	4.E-04	121	0.05	1.E-02
HLF	+	PAR-bZip	0.10	8.E-05	33	0.07	3.E-04
HNF4	+	NR	0.14	3.E-08	2000	0.09	6.E-07
Jdp2	++	AP1	0.11	5.E-05	31	0.05	9.E-03
Jun	+++	AP1	0.11	2.E-05	159	0.08	3.E-05
LXR	++	NR	0.13	3.E-07	742	0.08	4.E-06
Myb	+		0.09	9.E-04	149	0.02	2.E-01
Myc	++		0.10	2.E-04	121	0.04	2.E-02
NFIC	+++	NF	0.16	8.E-10	501	0.06	7.E-04
PPARα	+++	NR	0.18	4.E-12	2149	0.11	5.E-09
PPARγ	+++	NR	0.16	1.E-09	2010	0.09	2.E-07
RAR	+++	NR	0.14	4.E-08	1297	0.09	3.E-07
RXR	+++	NR	0.14	8.E-08	2076	0.12	1.E-10
Snai1	++	Repressor	-0.10	6.E-05	39	-0.05	1.E-02
Snai2	+	Repressor	-0.09	6.E-04	57	-0.06	3.E-03
Sox13	+	Repressor	-0.09	5.E-04	125	-0.05	1.E-02
TR4	++	NR	0.14	8.E-08	529	0.04	1.E-02
VBP	+++	PAR-bZip	0.10	2.E-04	21	0.09	1.E-06
Zfp410	++	Repressor	-0.09	5.E-04	220	-0.01	7.E-01

	Genomic Enrichment		**Deletions**	**Insertions**		**Synthetic**	
Factor	**Enrichment**	**Enrichment P-value**	**P-value (vs. control mutations)**	**Correlation with expression (ρ)**	**Correlation P-value**	**Correlation with expression (ρ)**	**Correlation P-value**
AP-1	1.9	8.E-59	9.E-05	0.14	0.E+00	-	-
AP-2	2.3	2.E-105	7.E-01	0.10	2.E-09	0.15	2.E-03
Zeb1	0.6	3.E-32	3.E-01	-0.13	3.E-15	-	-
ATF3	1.4	3.E-11	3.E-06	0.06	2.E-04	0.33	1.E-11
ATF4	1.5	1.E-18	3.E-03	0.20	0.E+00	-	-
Bach1	1.6	2.E-38	4.E-02	0.05	3.E-03	0.65	0.E+00
C/EBP	3.3	7.E-129	8.E-05	0.01	5.E-01	0.17	7.E-04
COUP-TFI	1.5	4.E-270	4.E-03	0.07	2.E-05	-	-
COUP-TFII	2.6	3.E-162	6.E-04	0.14	0.E+00	0.11	4.E-10
NFIX	2.1	1.E-134	7.E-01	0.04	9.E-03	-	-

DAX1	1.3	3.E-52	9.E-04	0.08	2.E-06	-	-
E4BP4	1.2	8.E-04	1.E-01	0.06	1.E-04	0.34	4.E-12
ERR1	1.3	7.E-30	4.E-03	0.24	0.E+00	-	-
ERR2	1.3	2.E-31	4.E-02	0.07	2.E-05	-	-
ERR3	1.3	5.E-18	1.E-02	0.08	4.E-06	-	-
Fos	1.7	7.E-41	2.E-06	0.08	3.E-06	0.55	0.E+00
FOXD1	1.2	2.E-03	9.E-01	-0.06	3.E-04	0.16	9.E-04
FOXO4	1.0	5.E-01	5.E-01	-0.18	0.E+00	0.15	2.E-03
FOXP3	1.0	3.E-01	3.E-01	-0.14	0.E+00	-	-
FRA1	1.7	5.E-58	3.E-07	0.14	0.E+00	0.57	0.E+00
HLF	2.4	5.E-38	3.E-02	0.06	2.E-04	-	-
HNF4	1.7	0.E+00	6.E-05	0.21	0.E+00	-	-
Jdp2	1.5	2.E-11	4.E-03	0.08	1.E-06	-	-
Jun	1.8	3.E-49	1.E-05	0.14	0.E+00	0.62	0.E+00
LXR	1.5	4.E-128	1.E-04	0.13	2.E-16	0.28	7.E-09
Myb	1.6	1.E-12	4.E-01	0.00	1.E+00	0.11	2.E-02
Myc	1.9	4.E-37	3.E-02	-0.12	2.E-12	0.34	3.E-12
NFIC	2.2	1.E-158	2.E-01	-0.06	5.E-04	-	-
PPARα	2.0	0.E+00	7.E-07	0.35	0.E+00	0.53	0.E+00
PPARγ	1.6	4.E-199	2.E-04	0.21	0.E+00	-	-
RAR	1.5	0.E+00	2.E-06	0.11	8.E-12	-	-
RXR	1.9	2.E-41	1.E-04	0.20	0.E+00	-	-
Snai1	0.7	1.E-16	1.E+00	-0.17	0.E+00	-	-
Snai2	0.8	3.E-09	1.E-01	-0.21	0.E+00	-	-
Sox13	0.6	3.E-26	1.E-01	-0.12	7.E-13	-	-
TR4	1.5	2.E-32	2.E-02	0.11	3.E-12	-	-
VBP	1.4	4.E-08	5.E-02	0.06	1.E-04	0.30	8.E-10
Zfp410	0.7	4.E-24	8.E-01	0.05	5.E-03	-	-

Table S3. Median Expression of Native Enhancers, Conditional on Motif Counts for Each TF

	Median Expression [25th Quantile, 75th Quantile]			
TF	**0 Motif Sites**	**1 Motif Site**	**2 Motif Sites**	**3 Motif Sites**
AP-1	-0.31 [-0.64,0.22]	-0.09 [-0.5,0.8]	-0.06 [-0.54,1.18]	-0.06 [-0.5,0.46]
AP-2	-0.34 [-0.67,0.2]	-0.23 [-0.53,0.38]	-0.11 [-0.59,0.57]	-0.28 [-0.63,0.93]
Zeb1	-0.26 [-0.61,0.3]	-0.46 [-0.76,0.08]	-0.66 [-0.8,-0.51]	

ATF3	-0.29 [-0.63,0.25]	-0.13 [-0.53,1.96]	0.7 [0.03,2.1]	
ATF4	-0.29 [-0.63,0.25]	-0.12 [-0.46,0.89]	3.3 [0.05,3.73]	2.9 [2.9,2.9]
Bach1	-0.29 [-0.63,0.24]	0.12 [-0.37,1.19]	5.32 [5.32,5.32]	
C/EBP	-0.36 [-0.66,0.12]	-0.23 [-0.59,0.31]	0.01 [-0.37,0.77]	0.7 [0.01,1.94]
COUP-TFI	-0.38 [-0.69,0.08]	-0.35 [-0.64,0.09]	-0.24 [-0.6,0.42]	0.01 [-0.45,0.82]
COUP-TFII	-0.32 [-0.65,0.22]	-0.13 [-0.55,0.47]	-0.22 [-0.83,0.83]	-0.29 [-0.29,-0.29]
NFIX	-0.3 [-0.63,0.22]	0.17 [-0.48,1.13]	0.17 [-0.42,1.47]	
DAX1	-0.31 [-0.65,0.24]	-0.19 [-0.52,0.34]	0.19 [-0.31,1.21]	-0.06 [-0.06,-0.06]
E4BP4	-0.29 [-0.63,0.25]	1.67 [-0.26,3.07]	0.66 [0.19,0.98]	
ERR1	-0.33 [-0.66,0.23]	-0.16 [-0.5,0.44]	-0.23 [-0.44,0.3]	-0.26 [-0.37,-0.06]
ERR2	-0.31 [-0.65,0.24]	-0.1 [-0.46,0.48]	-0.15 [-0.33,0.22]	
ERR3	-0.31 [-0.65,0.25]	-0.04 [-0.38,0.47]	-0.49 [-0.74,-0.26]	
Fos	-0.29 [-0.63,0.25]	-0.01 [-0.4,1.36]	1.16 [1.07,1.44]	
FOXD1	-0.29 [-0.63,0.25]	0.34 [-0.22,1.42]		
FOXO4	-0.31 [-0.64,0.25]	-0.07 [-0.34,0.72]	-0.11 [-0.32,1.21]	0.5 [0.12,0.89]
FOXP3	-0.3 [-0.64,0.24]	-0.09 [-0.48,0.73]	-0.36 [-0.76,0.03]	
FRA1	-0.29 [-0.63,0.23]	-0.06 [-0.46,1.22]	-0.09 [-0.1,0.55]	
HLF	-0.29 [-0.63,0.26]	0.41 [-0.17,1.49]	-0.48 [-0.48,-0.48]	
HNF4α	-0.39 [-0.7,0.01]	-0.29 [-0.63,0.23]	-0.26 [-0.6,0.33]	-0.1 [-0.5,0.96]
Jdp2	-0.29 [-0.63,0.25]	-0.32 [-0.44,-0.14]	2.1 [0.7,2.9]	
Jun	-0.3 [-0.63,0.24]	-0.1 [-0.53,0.65]	0.55 [-0.1,2.08]	1.52 [1.52,1.52]
LXR	-0.36 [-0.67,0.15]	-0.23 [-0.56,0.49]	-0.07 [-0.47,0.98]	-0.12 [-0.43,-0.05]
Myb	-0.29 [-0.64,0.23]	-0.09 [-0.48,0.64]	-0.2 [-0.54,1.05]	0.08 [-0.26,0.41]

Myc	-0.3 [-0.64,0.25]	-0.02 [-0.46,0.42]	0.11 [-0.28,0.69]	0.21 [0.21,0.21]
NFIC	-0.34 [-0.65,0.12]	-0.09 [-0.54,0.61]	0.12 [-0.38,1.18]	-0.24 [-0.56,0.27]
PPARα	-0.42 [-0.68,-0.01]	-0.34 [-0.68,0.09]	-0.26 [-0.6,0.39]	-0.03 [-0.4,0.97]
PPARγ	-0.5 [-0.66,-0.07]	-0.35 [-0.66,0.11]	-0.13 [-0.52,0.82]	-0.05 [-0.43,0.59]
RAR	-0.4 [-0.69,0.07]	-0.27 [-0.61,0.33]	-0.13 [-0.51,0.69]	-0.16 [-0.45,0.38]
RXR	-0.44 [-0.64,-0.01]	-0.35 [-0.67,0.15]	-0.22 [-0.56,0.57]	-0.08 [-0.44,0.7]
Snai1	-0.27 [-0.62,0.29]	-0.69 [-0.85,-0.33]		
Snai2	-0.27 [-0.62,0.29]	-0.61 [-0.84,-0.09]	0.31 [0.31,0.31]	
Sox13	-0.27 [-0.62,0.33]	-0.43 [-0.71,-0.13]	-0.34 [-0.59,-0.2]	
TR4	-0.36 [-0.67,0.16]	-0.14 [-0.51,0.51]	-0.1 [-0.41,0.55]	
VBP	-0.28 [-0.63,0.25]	0.86 [-0.29,2]	1.64 [1.64,1.64]	
Zfp410	-0.27 [-0.62,0.32]	-0.36 [-0.68,0.08]	-0.5 [-0.72,-0.01]	-0.5 [-0.5,-0.5]

	Median Expression [25th Quantile, 75th Quantile]			
TF	**4 Motif Sites**	**5 Motif Sites**	**6 Motif Sites**	**7 Motif Sites**
AP-1	0.23 [-0.32,0.94]	-0.09 [-0.09,-0.09]		
AP-2	0.24 [-0.22,0.48]	0.31 [0.03,0.38]		
Zeb1				
ATF3				
ATF4				
Bach1				
C/EBP	0.41 [0,1.26]			
COUP-TFI	-0.27 [-0.58,0.56]	-0.31 [-0.44,0.63]	-0.36 [-0.6,-0.06]	-0.23 [-0.24,-0.23]
COUP-TFII				
NFIX				
DAX1				

E4BP4				
ERR1	-0.94 [-0.94,-0.94]			
ERR2				
ERR3				
Fos				
FOXD1				
FOXO4				
FOXP3				
FRA1				
HLF				
HNF4α	0.13 [-0.36,1.43]	-0.23 [-0.24,0.03]	0.61 [0.46,0.77]	
Jdp2				
Jun				
LXR				
Myb	0.24 [-0.35,0.83]			
Myc	-0.88 [-0.88,-0.88]			
NFIC	-0.37 [-0.52,-0.21]	1.3 [1.3,1.3]	2.5 [2.5,2.5]	
PPARα	0.04 [-0.32,1.2]	2.77 [1.26,2.9]	-0.58 [-0.93,-0.47]	
PPARγ	0.12 [-0.27,0.3]			
RAR	0.22 [-0.32,0.66]	-0.91 [-0.91,-0.91]		
RXR	-0.15 [-0.53,0.86]	2.77 [2.77,2.77]		
Snai1				
Snai2				
Sox13				
TR4				
VBP				
Zfp410				

Table S4. Lasso model coefficients

TF	Coefficient
Intercept	-8.7E-01

TF (cont.)	Coefficient
E4BP4:HNF4	3.7E-03

AP1:AP1	-2.6E-03
AP1:LXR	-8.4E-03
AP1:PPARα	1.6E-04
AP-2:AP-2	8.0E-03
AP-2:ATF4	7.8E-02
AP-2:C/EBP	2.3E-02
AP-2:COUPTFI	-2.1E-02
AP-2:NFIX	2.7E-02
AP-2:LXR	3.5E-02
AP-2:Myb	1.0E-01
AP-2:RAR	-2.4E-03
Zeb1:HNF4	-4.1E-02
Zeb1:PPARα	-7.5E-03
Zeb1:RXR	-5.3E-02
ATF3:C/EBP	2.7E-01
ATF3:PPARα	5.0E-02
ATF4:ATF4	8.4E-02
ATF4:COUPTFI	1.3E-02
Bach1:HNF4	6.4E-02
Bach1:PPARα	2.0E-04
C/EBP:C/EBP	5.2E-02
C/EBP:COUPTFI	1.7E-02
C/EBP:NFIC	1.3E-02
C/EBP:COUPTFII	-2.9E-02
C/EBP:PPARY	8.4E-03
COUPTFI:COUPTFI	-1.2E-02
COUPTFI:FOXO4	1.0E-02
COUPTFI:LXR	-9.4E-03
COUPTFI:Myc	3.1E-02
COUPTFI:NFIC	-1.5E-03
COUPTFI:COUPTFII	8.9E-03
COUPTFI:PPARα	-1.1E-02
COUPTFI:Sox13	-1.1E-02
COUPTFI:TR4	3.7E-02
NFIX:PPARY	5.4E-02
NFIX:RAR	1.7E-05
DAX1:NFIC	-1.7E-02

E4BP4:RXR	2.7E-01
FOXD1:HNF4	1.1E-01
FOXD1:RXR	8.8E-03
FOXO4:HNF4	4.0E-02
FOXO4:PPARα	6.7E-02
FOXP3:HNF4	3.5E-02
FOXP3:PPARα	3.7E-02
FOXP3:RXR	4.1E-03
FRA1:PPARα	5.0E-02
HNF4:NFIC	-7.3E-03
HNF4:PPARα	2.8E-02
Jdp2:Jdp2	1.2E-01
Jun:Jun	1.5E-01
LXR:COUPTFII	-1.7E-03
LXR:PPARα	3.7E-02
LXR:RAR	-5.6E-02
Myb:Myb	2.6E-03
Myb:PPARα	1.4E-02
Myc:RAR	5.0E-02
NFIC:NFIC	7.1E-03
NFIC:PPARα	4.2E-02
NFIC:PPARY	3.8E-02
NFIC:Zfp410	-1.7E-02
LXR:LXR	1.9E-02
LXR:PPARα	7.8E-03
PPARα:PPARα	1.5E-02
PPARα:Snai1	-3.7E-03
PPARα:Sox13	-5.2E-02
PPARα:VBP	4.2E-02
PPARα:Zfp410	-3.4E-02
PPARY:PPARY	9.6E-03
PPARY:RXR	2.1E-02
RAR:TR4	3.1E-03
RXR:Snai1	-1.2E-02
RXR:Sox13	-2.1E-02
RXR:VBP	7.4E-03
RXR:Zfp410	-3.7E-02

Table S5. Pairwise TF interactions

		ANOVA Interaction p-value		Wilcox p-value		Interaction γ	
	Pair	Native	Synthetic	Mutation - Wilcox	Framework - Wilcox	Native	Synthetic
Synthetic	C/EBP:PPARα	1.6E-01	8.0E-03	4.7E-01	1.1E-02	0.04	-0.06
	Bach1:LXR	1.1E-02	2.3E-05	2.3E-05	1.8E-04	-0.36	-0.05
	FRA1:LXR	1.7E-01	5.6E-05	6.7E-05	1.4E-23	-0.14	-0.04
	Jun:LXR	5.2E-02	6.4E-05	1.0E-07	5.6E-23	-0.16	-0.04
	Bach1:PPARα	3.6E-01	1.3E-04	4.9E-03	2.2E-05	0.10	-0.05
	FRA1:PPARα	2.6E-01	3.2E-04	2.5E-02	2.8E-01	0.08	-0.04
	Jun:PPARα	5.9E-01	3.3E-04	3.6E-04	3.0E-01	-0.03	-0.04
	Fos:LXR	4.8E-03	8.6E-03	2.9E-07	3.0E-25	-0.42	-0.04
	LXR:PPARα	1.2E-01	1.8E-05	2.6E-10	3.9E-02	0.04	0.04
	Bach1:Jun	5.9E-02	4.7E-13	1.0E+00	1.2E-01	0.30	0.05
	Bach1:FRA1	9.8E-01	9.5E-13	2.5E-02	2.9E-01	0.00	0.05
	FRA1:Jun	4.8E-01	1.8E-12	3.2E-01	9.6E-02	-0.10	0.04
	Bach1:Fos	3.1E-02	2.1E-12	1.3E-01	1.1E-06	0.50	0.06
	Fos:Jun	4.7E-01	2.8E-12	5.3E-01	5.1E-05	0.10	0.06
	Fos:FRA1	4.0E-01	3.1E-12	1.3E-01	2.0E-05	-0.21	0.06
	Jun:Myb	4.1E-02	6.5E-04	1.3E-01	7.6E-04	-0.26	-0.06
	Bach1:Myb	2.1E-02	9.3E-04	2.8E-01	8.8E-04	-0.41	-0.06
	Fos:Myb	3.7E-01	1.6E-03	3.5E-01	1.8E-03	-0.19	-0.07
	FRA1:Myb	6.0E-02	1.8E-03	6.1E-01	1.3E-03	-0.30	-0.06
	FRA1:Myc	3.2E-01	3.8E-03	2.4E-02	5.9E-03	-0.24	0.06
	Bach1:Myc	2.6E-01	6.0E-03	3.0E-01	3.0E-01	0.24	0.05
Native	E4BP4:RXR	2.8E-05		6.3E-02	3.4E-02	1.13	
	PPARY:VBP	1.6E-04		4.6E-04	1.2E-11	1.99	
	E4BP4:PPARY	3.3E-03		7.4E-01	5.2E-02	1.01	
	E4BP4:LXR	6.5E-03	9.4E-01	1.6E-02	2.3E-01	0.56	0.00
	AP1:LXR	9.1E-04		3.3E-02	1.3E-05	-0.11	
	ATF3:DAX1	1.2E-03		3.8E-02	8.6E-02	-0.59	
	Fos:RXR	1.5E-03		7.5E-03	8.1E-03	-0.61	
	ERR1:Jun	3.1E-03		5.4E-02	2.2E-06	-0.31	
	Jun:RXR	3.4E-03		8.2E-03	1.5E-03	-0.25	
	Fos:LXR	4.8E-03	8.6E-03	2.9E-07	3.0E-25	-0.42	-0.04
	ERR2:Jun	5.6E-03		6.7E-06	3.8E-25	-0.51	
	Fos:COUPTFII	5.7E-03		5.1E-03	1.6E-28	-0.61	
	ERR1:Fos	8.3E-03		2.0E-06	2.7E-05	-0.43	
	ERR1:FRA1	8.6E-03		4.2E-03	1.3E-06	-0.35	
	FOXP3:PPARα	1.4E-03		1.8E-02	2.2E-01	0.23	
	FOXP3:RXR	2.2E-03		3.4E-02	5.5E-03	0.33	
	FOXO4:PPARα	2.6E-03	3.2E-01	7.0E-06	6.5E-01	0.19	0.03
	FOXP3:PPARY	4.7E-03		5.0E-01	1.2E-02	0.39	

	FOXO4:LXR	4.8E-03	9.5E-01	1.0E-04	8.7E-02	0.26	0.00
	FOXP3:RAR	7.4E-03		4.9E-02	2.2E-01	0.25	
	AP-2:ERR2	2.0E-03		3.4E-05	3.1E-23	-0.18	
	AP-2:FOXO4	4.4E-03	2.9E-01	2.3E-02	4.0E-02	-0.22	0.03
	AP-2:ATF4	8.1E-03		4.7E-04	8.1E-04	0.25	
	AP-2:COUPTFII	9.4E-03		1.4E-01	2.3E-17	-0.15	
	LXR:COUPTFII	6.9E-03		2.7E-04	8.3E-03	-0.15	
	LXR:RAR	5.0E-03		4.1E-07	6.3E-05	-0.08	

	Pair	Enrichment <10 bp	Count in Native	Class
Synthetic	C/EBP:PPARα	6.6E-26	469	
	Bach1:LXR	9.3E-01	15	
	FRA1:LXR	5.6E-03	41	
	Jun:LXR	2.2E-02	55	
	Bach1:PPARα	5.5E-02	30	AP-1/NR
	FRA1:PPARα	1.7E-10	73	
	Jun:PPARα	5.9E-09	103	
	Fos:LXR	9.8E-02	18	
	LXR:PPARα	1.5E-01	58	NR/NR
	Bach1:Jun	1.0E+00	1	
	Bach1:FRA1	1.0E+00	0	
	FRA1:Jun	3.6E-01	3	AP-1/AP-1
	Bach1:Fos	4.9E-01	0	
	Fos:Jun	4.6E-01	0	
	Fos:FRA1	8.0E-01	0	
	Jun:Myb	7.9E-01	20	
	Bach1:Myb	1.6E-01	8	Myb
	Fos:Myb	1.6E-01	5	
	FRA1:Myb	1.4E-01	17	
	FRA1:Myc	6.6E-02	7	AP-1/MYC
	Bach1:Myc	9.1E-01	4	
Native	E4BP4:RXR	2.2E-03	13	
	PPARY:VBP	1.5E-02	55	PPARg/PARbZip
	E4BP4:PPARY	1.5E-03	15	
	E4BP4:LXR	9.5E-01	13	
	AP1:LXR	8.2E-10	189	
	ATF3:DAX1	1.0E+00	9	
	Fos:RXR	1.7E-05	35	
	ERR1:Jun	9.9E-01	18	
	Jun:RXR	3.6E-10	120	AP-1/NR
	Fos:LXR	9.8E-02	18	
	ERR2:Jun	2.8E-01	3	
	Fos:COUPTFII	8.3E-01	8	
	ERR1:Fos	4.9E-01	17	
	ERR1:FRA1	1.0E+00	21	

	FOXP3:PPARα	6.1E-06	53	FOX/NR
	FOXP3:RXR	2.3E-15	57	
	FOXO4:PPARα	4.8E-07	54	
	FOXP3:PPARY	1.8E-11	62	
	FOXO4:LXR	6.5E-04	30	
	FOXP3:RAR	1.2E-02	34	
	AP-2:ERR2	6.2E-01	89	AP-2
	AP-2:FOXO4	3.3E-05	21	
	AP-2:ATF4	9.8E-02	63	
	AP-2:COUPTFII	1.0E-01	111	
	LXR:COUPTFII	6.2E-05	29	NR/NR
	LXR:RAR	3.4E-01	36	

Table S6. Changes in expression associated with interactions in synthetic enhancers

			Median Expression			
	TF1	**TF2**	**0 TF1 Sites**	**1 TF1 Sites**	**2 TF1 Sites**	**3 TF1 Sites**
Synthetic	Bach1	- Fos	0.264	0.144	0.978	1.875
		+ Fos	0.119	0.146	0.623	1.630
	Bach1	- Jun	0.263	-0.102		
		+ Jun	0.280	0.153	0.726	1.702
	Bach1	- LXR	0.147	0.042	0.689	1.793
		+ LXR	0.385	0.229	0.814	1.652
	Bach1	- Myb	0.278	0.155	0.799	1.775
		+ Myb	0.203	0.076	0.579	1.354
	Bach1	- Myc	0.245	0.135	0.715	1.661
		+ Myc	0.378	0.389	0.830	2.532
	Bach1	- PPARα	0.190	0.012	0.630	1.861
		+ PPARα	0.295	0.210	0.765	1.680
	Fos	- LXR	0.252	0.220	0.690	1.827
		+ LXR	0.465	0.337	0.807	1.677
	Fos	- Myb	0.412	0.309	0.781	1.831
		+ Myb	0.272	0.102	0.583	1.228
	FRA1	- Bach1	0.268	0.004	0.761	
		+ Bach1	-0.102	0.153	0.647	1.697
	FRA1	- Fos	0.265	0.166	0.947	1.844
		+ Fos	0.119	0.146	0.516	1.635
	FRA1	- LXR	0.123	0.100	0.572	1.769
		+ LXR	0.397	0.209	0.706	1.648
	FRA1	- Myb	0.282	0.156	0.681	1.744

	FRA1	+ Myb	0.203	0.102	0.572	1.311
		- Myc	0.245	0.143	0.637	1.648
		+ Myc	0.382	0.245	0.757	2.553
	FRA1	- PPARα	0.175	0.054	0.571	1.848
		+ PPARα	0.299	0.205	0.681	1.674
	Jun	- Fos	0.263	0.229	0.907	1.829
		+ Fos	0.119	0.149	0.491	1.601
	Jun	- LXR	0.122	0.116	0.559	1.703
		+ LXR	0.388	0.236	0.693	1.648
	Jun	- Myb	0.270	0.218	0.669	1.715
		+ Myb	0.211	0.076	0.541	1.280
	Jun	- PPARα	0.170	0.085	0.571	1.848
		+ PPARα	0.299	0.228	0.645	1.653
	LXR	- AP1	-0.401	-0.236	-0.058	0.120
		+ AP1	-0.055	-0.066	-0.016	-0.264
	PPARα	- C/EBP	0.433	0.339	0.710	1.363
		+ C/EBP	0.260	0.180	0.445	1.193
	PPARα	- LXR	0.399	0.532	0.328	-0.554
		+ LXR	0.609	0.175	0.676	1.367

			Median Expression		
	TF1	**TF2**	**4 TF1 Sites**	**5 TF1 Sites**	**6 TF1 Sites**
Synthetic	Bach1	- Fos	2.087	3.006	2.479
		+ Fos	1.778	2.879	2.995
	Bach1	- Jun			
		+ Jun	1.809	2.896	2.987
	Bach1	- LXR	1.998	2.951	3.188
		+ LXR	1.386	2.634	2.761
	Bach1	- Myb	1.777	2.879	2.929
		+ Myb	1.941	2.913	3.409
	Bach1	- Myc	1.809	2.896	2.987
		+ Myc			
	Bach1	- PPARα	2.224	2.932	2.921
		+ PPARα	1.702	2.829	2.991
	Fos	- LXR	2.018	2.833	3.215
		+ LXR	1.252	2.215	2.761
	Fos	- Myb	1.716	2.569	2.908
		+ Myb	1.923		3.580
	FRA1	- Bach1			
		+ Bach1	1.747		2.922

	FRA1	- Fos	1.828		2.569
		+ Fos	1.669		2.927
	FRA1	- LXR	1.941		3.115
		+ LXR	1.386		2.753
	FRA1	- Myb	1.669		2.914
		+ Myb	1.923		3.237
	FRA1	- Myc	1.747		2.922
		+ Myc			
	FRA1	- PPARα	2.112		2.932
		+ PPARα	1.671		2.922
	Jun	- Fos	1.958		2.569
		+ Fos	1.769		2.927
	Jun	- LXR	1.983		3.115
		+ LXR	1.386		2.753
	Jun	- Myb	1.763		2.914
		+ Myb	1.923		3.237
	Jun	- PPARα	2.112		2.932
		+ PPARα	1.747		2.922
	LXR	- AP1	-0.486	-0.583	
		+ AP1	-0.220		-0.614
	PPARα	- C/EBP	1.333	1.433	1.720
		+ C/EBP	1.001	1.795	
	PPARα	- LXR			
		+ LXR	1.267	1.510	1.720

Table S7. Changes in expression associated with interactions in native enhancers

		Median Expression			
TF1	**TF2**	**0 TF1 Sites**	**1 TF1 Sites**	**2 TF1 Sites**	**3 TF1 Sites**
AP-2	- ATF4	-0.349	-0.238	-0.144	-0.273
	+ ATF4	-0.134	0.272	0.697	-0.280
AP-2	- ERR2	-0.386	-0.234	-0.099	-0.277
	+ ERR2	-0.020	-0.247	-0.133	-0.267
AP-2	- FOXO4	-0.368	-0.234	-0.139	-0.269
	+ FOXO4	-0.025	-0.228	0.224	-0.808
AP-2	- COUPTFII	-0.398	-0.238	-0.144	-0.277
	+ COUPTFII	-0.146	-0.170	-0.102	0.222
ATF3	- DAX1	-0.323	-0.118	1.841	
	+ DAX1	-0.180	-0.133	0.033	

C/EBP	- ATF3	-0.366	-0.247	0.010	0.497
	+ ATF3	-0.149	1.137	2.203	2.393
ERR1	- Fos	-0.336	-0.180	-0.235	-0.258
	+ Fos	0.112	0.221	-0.086	
ERR1	- FRA1	-0.349	-0.174	-0.229	-0.258
	+ FRA1	-0.036	-0.089	-0.232	
Fos	- COUPTFII	-0.335	0.037	1.350	
	+ COUPTFII	-0.144	-0.051	0.483	
Jun	- ERR1	-0.354	-0.099	1.141	1.522
	+ ERR1	-0.187	-0.175	0.191	
Jun	- ERR2	-0.344	-0.099	1.141	1.522
	+ ERR2	-0.105	-0.179	-0.086	
LXR	- E4BP4	-0.368	-0.231	-0.066	-0.238
	+ E4BP4	-0.292	1.027	2.579	1.204
LXR	- Fos	-0.372	-0.229	-0.083	-0.264
	+ Fos	0.227	-0.067	0.879	-0.134
LXR	- FOXO4	-0.388	-0.237	-0.091	-0.232
	+ FOXO4	-0.175	0.298	1.875	-0.914
LXR	- COUPTFII	-0.388	-0.246	-0.011	-0.238
	+ COUPTFII	-0.142	-0.126	-0.101	-0.232
LXR	- Fos	-0.363	-0.236	-0.075	-0.075
	+ Fos	0.060	0.112	0.892	-0.232
PPARα	- FOXO4	-0.428	-0.373	-0.274	-0.082
	+ FOXO4	-0.324	-0.202	0.298	1.488
PPARα	- FOXP3	-0.427	-0.357	-0.269	-0.075
	+ FOXP3	-0.355	-0.247	0.089	1.398
PPARY	- E4BP4	-0.505	-0.353	-0.138	-0.050
	+ E4BP4		0.658	2.579	
PPARY	- FOXP3	-0.505	-0.365	-0.149	-0.100
	+ FOXP3	-0.470	-0.211	0.428	1.220
PPARY	- VBP	-0.505	-0.353	-0.133	-0.050
	+ VBP		0.831	3.749	
RAR	- FOXP3	-0.409	-0.278	-0.148	-0.161
	+ FOXP3	-0.276	0.108	0.501	
RAR	- LXR	-0.421	-0.353	-0.113	-0.029
	+ LXR	-0.270	-0.221	-0.134	-0.170
RXR	- E4BP4	-0.460	-0.357	-0.230	-0.075
	+ E4BP4	0.343	0.658	3.058	
RXR	- Fos	-0.457	-0.367	-0.217	-0.075
	+ Fos	2.040	0.209	-0.232	0.428
RXR	- FOXP3	-0.463	-0.364	-0.236	-0.138
	+ FOXP3	-0.362	-0.247	0.439	0.914
RXR	- Jun	-0.463	-0.372	-0.229	-0.054

	+ Jun	0.193	0.191	-0.131	-0.126

		Median Expression		
TF1	**TF2**	**4 TF1 Sites**	**5 TF1 Sites**	**6 TF1 Sites**
AP-2	- ATF4	0.240	0.310	
	+ ATF4			
AP-2	- ERR2	0.240	0.028	
	+ ERR2		0.347	
AP-2	- FOXO4	0.265	0.310	
	+ FOXO4	-0.874		
AP-2	- COUPTFII	0.169	0.206	
	+ COUPTFII	0.240	0.310	
ATF3	- DAX1			
	+ DAX1			
C/EBP	- ATF3	0.392		
	+ ATF3	3.013		
ERR1	- Fos	-0.945		
	+ Fos			
ERR1	- FRA1	-0.945		
	+ FRA1			
Fos	- COUPTFII			
	+ COUPTFII			
Jun	- ERR1			
	+ ERR1			
Jun	- ERR2			
	+ ERR2			
LXR	- E4BP4	-0.326	-0.583	-0.614
	+ E4BP4			
LXR	- Fos	-0.326	-0.583	-0.614
	+ Fos			
LXR	- FOXO4	-0.326	-0.583	-0.614
	+ FOXO4			
LXR	- COUPTFII	-0.269		-0.614
	+ COUPTFII	-0.589	-0.583	
LXR	- Fos			
	+ Fos			
PPARα	- FOXO4	0.033	2.769	-0.577
	+ FOXO4	1.875		
PPARα	- FOXP3	0.023	2.769	-0.577

	+ FOXP3	1.164		
PPARY	- E4BP4	0.123		
	+ E4BP4			
PPARY	- FOXP3	0.123		
	+ FOXP3			
PPARY	- VBP	0.123		
	+ VBP			
RAR	- FOXP3	0.222	-0.914	
	+ FOXP3			
RAR	- LXR			
	+ LXR	0.222	-0.914	
RXR	- E4BP4	-0.147	2.769	
	+ E4BP4			
RXR	- Fos	-0.147	2.769	
	+ Fos			
RXR	- FOXP3	-0.147	2.769	
	+ FOXP3			
RXR	- Jun	-0.147		
	+ Jun		2.769	

REFERENCES

1. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.

2. Roadmap Epigenomics C, *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317-330.

3. Spitz F & Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 13(9):613-626.

4. Ptashne M & Gann A (1997) Transcriptional activation by recruitment. *Nature* 386(6625):569-577.

5. Badis G, *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935):1720-1723.

6. Badis G, *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* 32(6):878-887.

7. Grove CA, *et al.* (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell* 138(2):314-327.

8. Jolma A, *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1-2):327-339.

9. Zhu C, *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 19(4):556-566.

10. Gerstein MB, *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91-100.

11. Johnson DS, Mortazavi A, Myers RM, & Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830):1497-1502.

12. Ren B, *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science* 290(5500):2306-2309.

13. Robertson G, *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4(8):651-657.

14. Wei CL, *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124(1):207-219.

15. Kheradpour P, *et al.* (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* 23(5):800-811.

16. Kheradpour P & Kellis M (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* 42(5):2976-2987.

17. Kwasnieski JC, Fiore C, Chaudhari HG, & Cohen BA (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome research* 24(10):1595-1602.

18. White MA, Myers CA, Corbo JC, & Cohen BA (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci U S A* 110(29):11952-11957.

19. Whitfield TW, *et al.* (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome biology* 13(9):R50.

20. Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. *Developmental cell* 21(4):611-626.

21. Landolin JM, *et al.* (2010) Sequence features that drive human promoter function and tissue specificity. *Genome Res* 20(7):890-898.

22. Fisher WW, *et al.* (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc Natl Acad Sci U S A* 109(52):21330-21335.

23. Rowan S, *et al.* (2010) Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes & development* 24(10):980-985.

24. Jiang J & Levine M (1993) Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* 72(5):741-752.

25. Gaudet J & Mango SE (2002) Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. *Science* 295(5556):821-825.

26. Siggers T, Duyzend MH, Reddy J, Khan S, & Bulyk ML (2011) Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol Syst Biol* 7:555.

27. Slattery M, *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147(6):1270-1282.

28. Dror I, Golan T, Levy C, Rohs R, & Mandel-Gutfreund Y (2015) A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* 25(9):1268-1280.

29. Gordan R, *et al.* (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* 3(4):1093-1104.

30. Levo M, *et al.* (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome research* 25(7):1018-1029.

31. Barozzi I, *et al.* (2014) Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell* 54(5):844-857.

32. Mirny LA (2010) Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A* 107(52):22534-22539.

33. Thurman RE, *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature* 489(7414):75-82.

34. Raveh-Sadka T, *et al.* (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature genetics* 44(7):743-750.

35. Guertin MJ & Lis JT (2013) Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Curr Opin Genet Dev* 23(2):116-123.

36. John S, *et al.* (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics* 43(3):264-268.

37. Li XY, *et al.* (2011) The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome biology* 12(4):R34.

38. Polach KJ & Widom J (1996) A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J Mol Biol* 258(5):800-812.

39. Heinz S, *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576-589.

40. Buck MJ & Lieb JD (2006) A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nature genetics* 38(12):1446-1451.

41. Zeitlinger J, *et al.* (2003) Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* 113(3):395-404.

42. Blau J, *et al.* (1996) Three functional classes of transcriptional activation domain. *Molecular and cellular biology* 16(5):2044-2055.

43. Han K, Levine MS, & Manley JL (1989) Synergistic activation and repression of transcription by Drosophila homeobox proteins. *Cell* 56(4):573-583.

44. Scholes CD, A.H.; Sanchez, A. (2016) Integrating regulatory information via combinatorial control of the
transcription cycle. *bioRxiv preprint*.

45. Isakova A, Berset Y, Hatzimanikatis V, & Deplancke B (2016) Quantification of Cooperativity in Heterodimer-DNA Binding Improves the Accuracy of Binding Specificity Models. *J Biol Chem* 291(19):10293-10306.

46. Lefterova MI, *et al.* (2008) PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. *Genes & development* 22(21):2941-2952.

47. Nielsen R, *et al.* (2008) Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes & development* 22(21):2953-2967.

48. Mikkelsen TS, *et al.* (2010) Comparative epigenomic analysis of murine and human adipogenesis. *Cell* 143(1):156-169.

49. Melnikov A, *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology* 30(3):271-277.

50. Berman BP, *et al.* (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A* 99(2):757-762.

51. Guertin MJ & Lis JT (2010) Chromatin landscape dictates HSF binding to target DNA elements. *PLoS genetics* 6(9):e1001114.

52. Robertson AG, *et al.* (2008) Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* 18(12):1906-1917.

53. Maerkl SJ & Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315(5809):233-237.

54. Waki H, *et al.* (2011) Global mapping of cell type-specific open chromatin by FAIRE-seq reveals the regulatory role of the NFI family in adipocyte differentiation. *PLoS genetics* 7(10):e1002311.

55. Lupien M, *et al.* (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* 132(6):958-970.

56. Sherwood RI, *et al.* (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature biotechnology* 32(2):171-178.

57. Ballas N, *et al.* (2001) Regulation of neuronal traits by a novel transcriptional complex. *Neuron* 31(3):353-365.

58. Ernst J & Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology* 28(8):817-825.

59. Gelman L, *et al.* (1999) p300 interacts with the N- and C-terminal part of PPARgamma2 in a ligand-independent and -dependent manner, respectively. *J Biol Chem* 274(12):7681-7688.

60. Blanco JC, *et al.* (1998) The histone acetylase PCAF is a nuclear receptor coactivator. *Genes & development* 12(11):1638-1651.

61. Siersbaek R, *et al.* (2011) Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *EMBO J* 30(8):1459-1472.

62. Gotea V, *et al.* (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* 20(5):565-577.

63. Lifanov AP, Makeev VJ, Nazina AG, & Papatsenko DA (2003) Homotypic regulatory clusters in Drosophila. *Genome Res* 13(4):579-588.

64. Rosen ED, *et al.* (2002) C/EBPalpha induces adipogenesis through PPARgamma: a unified pathway. *Genes & development* 16(1):22-26.

65. Yu K, *et al.* (2014) Activating transcription factor 4 regulates adipocyte differentiation via altering the coordinate expression of CCATT/enhancer binding protein beta and peroxisome proliferator-activated receptor gamma. *The FEBS journal* 281(10):2399-2409.

66. Dahle MK, *et al.* (2002) Mechanisms of FOXC2- and FOXD1-mediated regulation of the RI alpha subunit of cAMP-dependent protein kinase include release of transcriptional repression and activation by protein kinase B alpha and cAMP. *J Biol Chem* 277(25):22902-22908.

67. Distel RJ, Ro HS, Rosen BS, Groves DL, & Spiegelman BM (1987) Nucleoprotein complexes that regulate gene expression in adipocyte differentiation: direct participation of c-fos. *Cell* 49(6):835-844.

68. Patel YM & Lane MD (2000) Mitotic clonal expansion during preadipocyte differentiation: calpain-mediated turnover of p27. *J Biol Chem* 275(23):17653-17660.

69. Seo J, *et al.* (2009) Atf4 regulates obesity, glucose homeostasis, and energy expenditure. *Diabetes* 58(11):2565-2573.

70. Lee YH, *et al.* (2013) Transcription factor Snail is a novel regulator of adipocyte differentiation via inhibiting the expression of peroxisome proliferator-activated receptor gamma. *Cell Mol Life Sci* 70(20):3959-3971.

71. Cameron TL, Belluoccio D, Farlie PG, Brachvogel B, & Bateman JF (2009) Global comparative transcriptome analysis of cartilage formation in vivo. *BMC Dev Biol* 9:20.

72. Cuesta I, Zaret KS, & Santisteban P (2007) The forkhead factor FoxE1 binds to the thyroperoxidase promoter during thyroid cell differentiation and modifies compacted chromatin structure. *Molecular and cellular biology* 27(20):7302-7314.

73. Sekiya T, Muthurajan UM, Luger K, Tulin AV, & Zaret KS (2009) Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes & development* 23(7):804-809.

74. Zaret KS, *et al.* (2008) Pioneer factors, genetic competence, and inductive signaling: programming liver and pancreas progenitors from the endoderm. *Cold Spring Harb Symp Quant Biol* 73:119-126.

75. Dusserre Y & Mermod N (1992) Purified cofactors and histone H1 mediate transcriptional regulation by CTF/NF-I. *Molecular and cellular biology* 12(11):5228-5237.

76. Alevizopoulos A, *et al.* (1995) A proline-rich TGF-beta-responsive transcriptional activator interacts with histone H3. *Genes & development* 9(24):3051-3066.

77. Ferrari S, *et al.* (2004) Chromatin domain boundaries delimited by a histone-binding protein in yeast. *J Biol Chem* 279(53):55520-55530.

78. Hebbar PB & Archer TK (2003) Nuclear factor 1 is required for both hormone-dependent chromatin remodeling and transcriptional activation of the mouse mammary tumor virus promoter. *Molecular and cellular biology* 23(3):887-898.

79. Pittenger MF, *et al.* (1999) Multilineage potential of adult human mesenchymal stem cells. *Science* 284(5411):143-147.

80. Joaquin M & Watson RJ (2003) Cell cycle regulation by the B-Myb transcription factor. *Cell Mol Life Sci* 60(11):2389-2401.

81. Huang Z, Xu H, & Sandell L (2004) Negative regulation of chondrocyte differentiation by transcription factor AP-2alpha. *J Bone Miner Res* 19(2):245-255.

82. Erwin GD, *et al.* (2014) Integrating diverse datasets improves developmental enhancer prediction. *PLoS computational biology* 10(6):e1003677.

83. Wang J, *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 22(9):1798-1812.

84. Chinenov Y & Kerppola TK (2001) Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene* 20(19):2438-2452.

85. Kerppola TK & Curran T (1993) Selective DNA bending by a variety of bZIP proteins. *Molecular and cellular biology* 13(9):5479-5489.

86. Shemshedini L, Knauthe R, Sassone-Corsi P, Pornon A, & Gronemeyer H (1991) Cell-specific inhibitory and stimulatory effects of Fos and Jun on transcription activation by nuclear receptors. *EMBO J* 10(12):3839-3849.

87. Herrlich P (2001) Cross-talk between glucocorticoid receptor and AP-1. *Oncogene* 20(19):2465-2475.

88. Oh IH & Reddy EP (1999) The myb gene family in cell growth, differentiation and apoptosis. *Oncogene* 18(19):3017-3033.

89. Vinson CR, Hai T, & Boyd SM (1993) Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design. *Genes & development* 7(6):1047-1058.

90. Vallejo M, Ron D, Miller CP, & Habener JF (1993) C/ATF, a member of the activating transcription factor family of DNA-binding proteins, dimerizes with CAAT/enhancer-binding proteins and directs their binding to cAMP response elements. *Proc Natl Acad Sci U S A* 90(10):4679-4683.

91. Kamei Y, *et al.* (1996) A CBP integrator complex mediates transcriptional activation and AP-1 inhibition by nuclear receptors. *Cell* 85(3):403-414.

92. van der Vos KE & Coffer PJ (2008) FOXO-binding partners: it takes two to tango. *Oncogene* 27(16):2289-2299.

93. Inoue F, *et al.* (2016) A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *bioRxiv preprint.*

94. Simicevic J, *et al.* (2013) Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nat Methods* 10(6):570-576.

95. Matys V, *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34(Database issue):D108-110.

96. Mathelier A, *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42(Database issue):D142-147.

97. Grant CE, Bailey TL, & Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017-1018.

98. Lefterova MI, *et al.* (2010) Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Molecular and cellular biology* 30(9):2078-2089.

99. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.

100. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357-359.

101. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B* 58(1):267-288.

102. Siersbaek R, Nielsen R, & Mandrup S (2012) Transcriptional networks and chromatin remodeling controlling adipogenesis. *Trends in endocrinology and metabolism: TEM* 23(2):56-64.

103. Farmer SR (2006) Transcriptional control of adipocyte formation. *Cell metabolism* 4(4):263-273.

104. Shuman JD, Cheong J, & Coligan JE (1997) ATF-2 and C/EBPalpha can form a heterodimeric DNA binding complex in vitro. Functional implications for transcriptional regulation. *J Biol Chem* 272(19):12793-12800.

105. Mann IK, *et al.* (2013) CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome Res* 23(6):988-997.

106. Maekawa T, Jin W, & Ishii S (2010) The role of ATF-2 family transcription factors in adipocyte differentiation: antiobesity effects of p38 inhibitors. *Molecular and cellular biology* 30(3):613-625.

107. Luther J, *et al.* (2011) Elevated Fra-1 expression causes severe lipodystrophy. *Journal of cell science* 124(Pt 9):1465-1476.

108. Gubelmann C, *et al.* (2014) Identification of the transcription factor ZEB1 as a central component of the adipogenic gene regulatory network. *Elife* 3:e03346.

109. Zhu J, *et al.* (2010) Effects of FoxO4 overexpression on cholesterol biosynthesis, triacylglycerol accumulation, and glucose uptake. *Journal of lipid research* 51(6):1312-1324.

110. Nakae J, *et al.* (2003) The forkhead transcription factor Foxo1 regulates adipocyte differentiation. *Developmental cell* 4(1):119-129.

111. Seo JB, *et al.* (2004) Activated liver X receptors stimulate adipocyte differentiation through induction of peroxisome proliferator-activated receptor gamma expression. *Molecular and cellular biology* 24(8):3430-3444.

112. Li L, *et al.* (2009) The nuclear orphan receptor COUP-TFII plays an essential role in adipogenesis, glucose homeostasis, and energy metabolism. *Cell metabolism* 9(1):77-87.

113. Cornelius P, MacDougald OA, & Lane MD (1994) Regulation of adipocyte development. *Annual review of nutrition* 14:99-129.

114. Sarruf DA, *et al.* (2005) Cyclin D3 promotes adipogenesis through activation of peroxisome proliferator-activated receptor gamma. *Molecular and cellular biology* 25(22):9985-9995.

115. Wenke AK & Bosserhoff AK (2010) Roles of AP-2 transcription factors in the regulation of cartilage and skeletal development. *The FEBS journal* 277(4):894-902.

116. Park YM, *et al.* (2012) Snail, a transcriptional regulator, represses adiponectin expression by directly binding to an E-box motif in the promoter. *Metabolism: clinical and experimental* 61(11):1622-1632.

117. Bi W, Deng JM, Zhang Z, Behringer RR, & de Crombrugghe B (1999) Sox9 is required for cartilage formation. *Nature genetics* 22(1):85-89.

Chapter 3

Positional specificity of different transcription factor classes within enhancers

Parts of this chapter were first published as:

Grossman SR, *et al.* (2018) Positional specificity of different transcription factor classes within enhancers. ***Proc Natl Acad Sci U S A*** 114(7):E1291-E1300.

ABSTRACT

Gene expression is controlled by sequence-specific transcription factors (TFs), which bind to regulatory sequences in DNA. TF binding occurs in nucleosome-depleted regions of DNA (NDRs) (1-3), which generally encompass regions of similar length to those protected by nucleosomes (1, 4, 5). However, less is known about *where* within these regions specific TFs tend to be found. Here, we characterize the positional bias of inferred binding sites for 103 TFs within ~500,000 NDRs across 47 cell types. We find that distinct classes of TFs display different binding preferences: some tend to bind towards the edges, some toward the center, and some at other positions within the NDR. These patterns are highly consistent across cell types, suggesting they reflect TF-specific intrinsic structural or functional characteristics. In particular, TF classes that bind at NDR edges are enriched for those known to interact with histones and chromatin remodelers, whereas TFs with central enrichment interact with other TFs and

cofactors such as p300. Our results suggest distinct regiospecific binding patterns and functions of TF classes within enhancers.

INTRODUCTION

Gene expression is controlled by sequence-specific transcription factors (TFs), which bind to regulatory sequences in DNA. TF binding occurs in nucleosome-depleted regions of DNA (NDRs) (1-3), which generally encompass regions of similar length to those protected by nucleosomes (1, 4, 5). However, less is known about *where* within these regions specific TFs tend to be found.

RESULTS

To investigate the characteristic positions of TF binding sites in distal regulatory elements (enhancers), we identified active regulatory elements across numerous cell types and characterized predicted functional TF binding sites within these elements. We defined putative active regulatory elements by first identifying NDRs in 47 cell types based on DNAseI hypersensitive (DHS) sites defined by the Roadmap Epigenomics project (6) and Assay for Transposase-Accessible Chromatin (ATAC)-seq experiments performed in each cell type. We then further selected those NDRs marked by the active chromatin modification, H3K27ac (using ChIP-seq data from the Roadmap Epigenomics project) (Fig. 1A,B). We and others have previously shown, by massively-parallel reporter assays (MPRA), that genomic sites satisfying these criteria are highly enriched for enhancer activity compared to other genomic sites and random sequences (7-10). Overall, we identified ~40,000-160,000 putative active regulatory elements per cell type, together representing a total of ~500,000 distinct (non-overlapping) elements. The edges of flanking nucleosomes appear to occur at ~120 ± 50 bp from the peak of the DHS/ATAC-seq signal, as assayed by micrococcal nuclease-digestion assays (MNase-seq) (Fig 2A,B). The regions are enriched for transcriptional initiation, consistent with

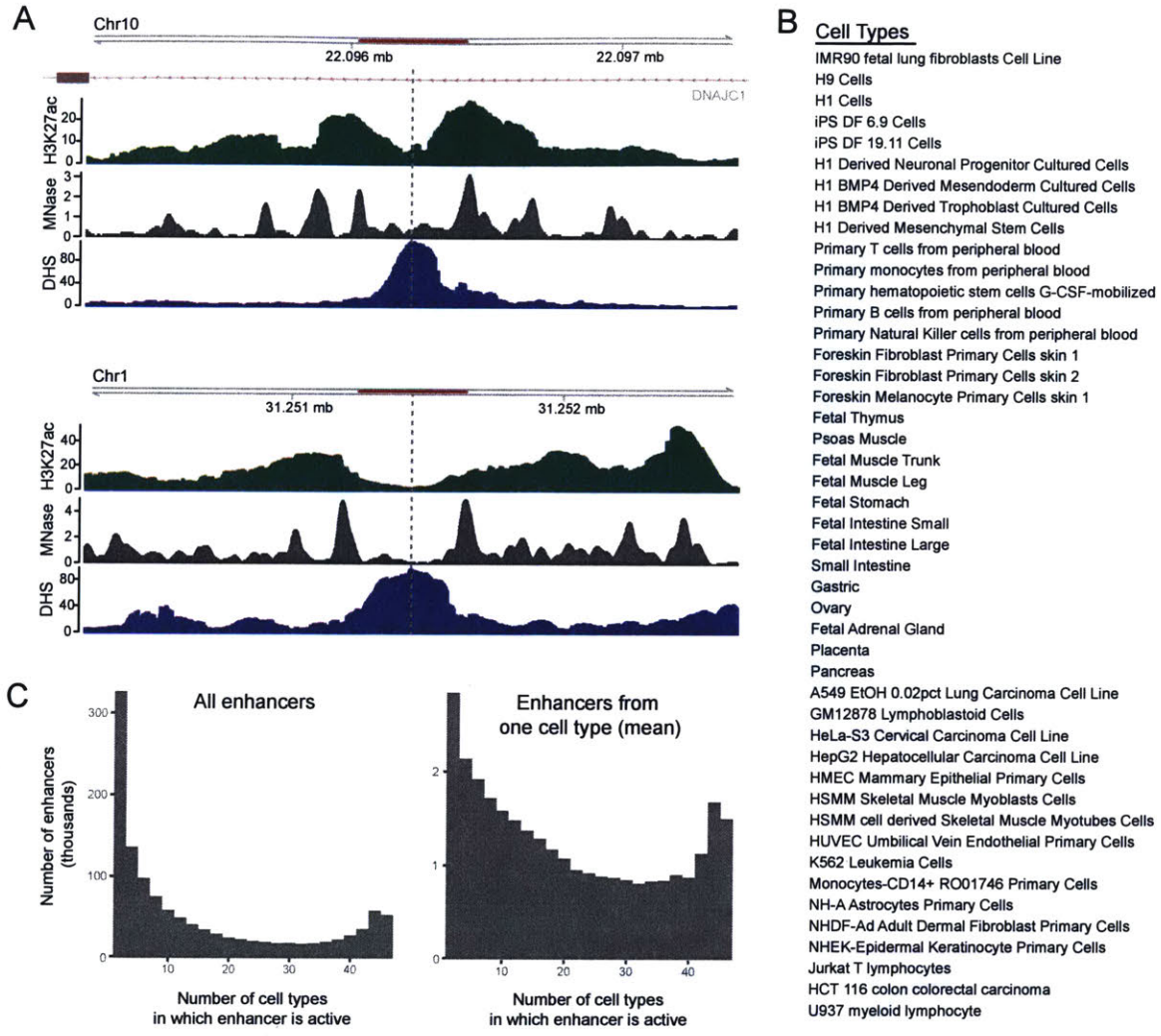


Figure 1. Selection of nucleosome-depleted regions used in analysis. (A) Chromatin structure around two example putative regulatory regions from K562 cells. DNaseI hypersensitive regions (DHS) (blue track) surrounded by regions marked with H3K27ac (green track) centered around the peak of the signal (dotted line) were selected. Micrococcal nuclease accessibility data (MNase-seq, black track) shows these NDRs are flanked by well-positioned nucleosomes. **(B)** List of 47 cell types in which 577,669 putative regulatory NDRs were identified. **(C)** Distribution of number of cell types with activity (defined by significant DHS/ATAC-seq and H3K27ac signal) for all NDRs (left) and average distribution (right) for NDRs active in a cell type (averaged across all cell types).

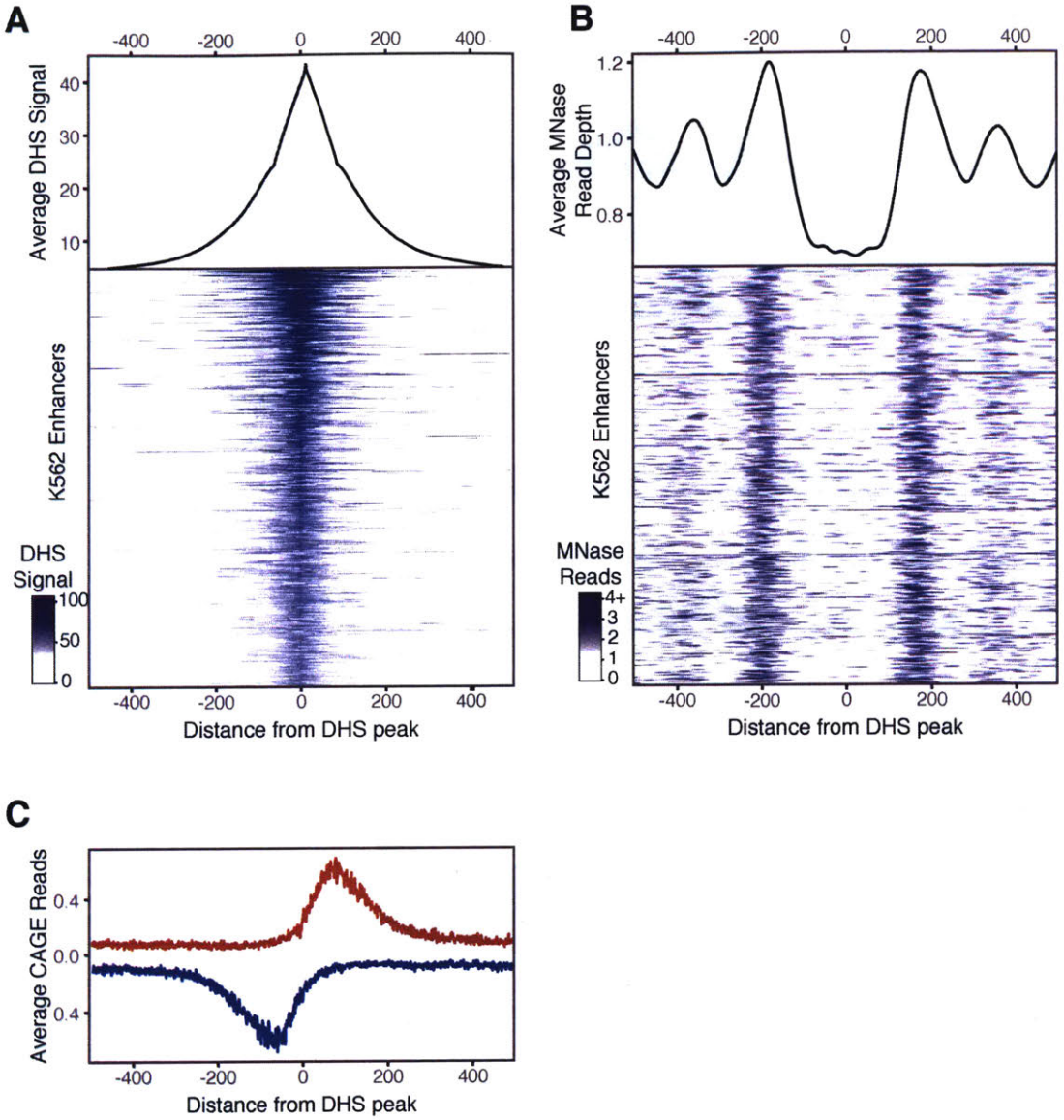


Figure 2. Chromatin structure around putative regulatory NDRs. (A, B) The nucleosome-depleted region at putative regulatory elements tends to span ~200 bp centered around the peak of the DHS signal and is generally flanked by well-positioned nucleosomes centered at around +200 bp and -200 bp. Composite plot (top) and heatmap (bottom) of DHS signal (a) and MNase-seq reads (B) in 1 kb region aligned around peak of DHS signal. 5000 regions from K562 cells sorted by maximum DHS score are shown in heatmaps. **(C)** Composite profile of CAGE reads, indicating transcriptional initiation, on the plus strand (red) and minus strand (blue) from 14 cell types aligned around the peak of the DHS signal in NDRs. Initiation of gene and eRNA transcription peaks ~55 bp away from the peak of the DHS signal and is oriented outwards from the accessible region.

previous reports (11); the peak of transcription is ~55 bp away from the peak of the DHS/ATAC-seq signal and ~45 bp before the position of the flanking nucleosome (Fig. 2C). As expected, cell types with similar anatomical and developmental origins tended to have correlated regulatory elements (Fig. S1). Because developmental enhancers and housekeeping enhancers are typically regulated by distinct sets of transcription factors (12, 13), we distinguished in our analysis between cell-type-restricted enhancers (active in <50% of the cell types) and ubiquitous enhancers (active in >90%)(Fig. 1C).

We next sought to infer functional TF-binding sites within the active regulatory elements. In a recent study (7), we found that TF binding is strongly correlated with the quantitative DNA accessibility of a region. Furthermore, the TF motifs associated with enhancer activity in reporter assays in a cell type corresponded closely to those that are most enriched in the genomic sequences of active regulatory elements in that cell type (7). In these assays, disrupting occurrences of the 20-30 most enriched motifs in such genomic regulatory sequences frequently caused significant changes in enhancer activity, indicating that many represent functional TF binding sites. Together, these results suggest that, for a cell type, occurrences of highly enriched motifs in highly accessible regions are very likely to represent functional TF binding sites.

We used this approach to define a set of candidate functional TF-binding sites. For each of the 47 cell types, we selected the 7500 cell-type restricted NDRs with the strongest DHS/ATAC-seq signals, with an average of 6% being promoter-proximal regions (<1 kb from an annotated TSS) and 94% being distal enhancers. Within these regions, we identified all occurrences of 1796 known motifs (corresponding to 777 TFs),

Figure S1. Cell types clustered by activity of putative regulatory elements. t-SNE plot of 47 cell types clustered based on the activity of the 577,669 putative regulatory elements. Regions were defined as active in a given cell type if they overlapped significant DHS/ATAC-seq and H3K27ac ChIP-seq peaks. Each point represents one cell type. Cell type are colored based on their anatomical origin.

and focused on the 20 most enriched motifs in the cell type (after removing highly similar motifs, see Methods) (Fig. 3A). Overall, these sites corresponded to 103 different TFs across the 47 cell types. As expected, the motif enrichment profiles were correlated among related cell types (Fig. 3B).

We then studied the positions of inferred binding sites for each of the 103 TFs, relative to the peak of the DHS/ATAC-seq signal in the active regulatory elements. Different TFs show strikingly different positional binding patterns (Fig. 4, S2). Some are strongly concentrated at the peak of the DNase/ATAC-seq signal (e.g. CTCF); some are enriched over a more widely distributed central region (e.g. ELF1); some are clustered near the edges of the region (e.g. FOXP1 and ARID3A); and some tend to bind at specific distance from the center of the NDRs (e.g. EPAS1 and RREB1).

To classify these patterns, we calculated the density profiles in ±200 bp regions around the peak and clustered them using k-medoid clustering (see methods). The analysis identified 6 clusters of distinct position patterns (Fig. 5A). The clusters are clearly significant: the mean Kullback-Leibler divergence between density profiles within the same cluster is 1-2 orders of magnitude smaller than the mean between density profiles in different clusters (Fig. 5B) and the density profiles cannot be explained by local sequence composition (Fig. S3A,B). Three of these clusters represent motifs that occur most frequently near the center of NDRs, while the other three clusters tend to occur nearer to the edges (Fig. 5C).

Cluster 1 contains 10 TFs with inferred binding sites that are strongly biased toward the peak of highest DNA accessibility at the middle of the NDR, suggesting that their binding directly shapes local chromatin architecture. For 6 of these TFs (CTCF,

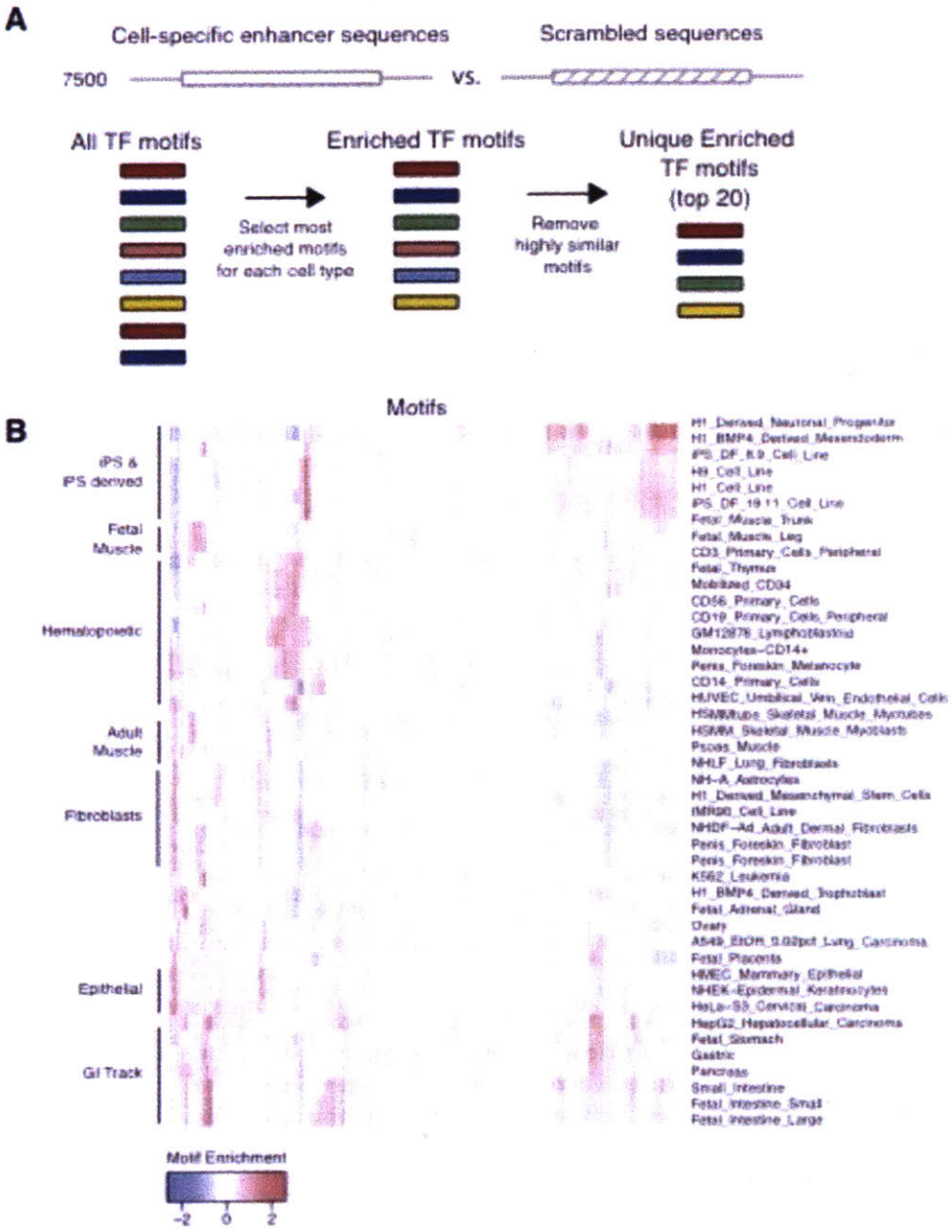


Figure 3. Identification of enriched TF motifs in each cell type. (A) TF motifs from JASPAR, TRANSFAC, and CIS-BP enriched in genomic sequences of NDRs relative to shuffled sequences were identified for each cell type. For TFs with multiple motifs in the combined database, the most enriched motif corresponding to each TF in each cell type was selected, and motifs with high similarity to a more highly enriched motif were removed (see Methods). The 20 most enriched TF motifs from the filtered list were used for positional analysis. **(B)** Enrichment of TF motifs across cell types. Related cell types generally show correlated motif enrichment profiles.

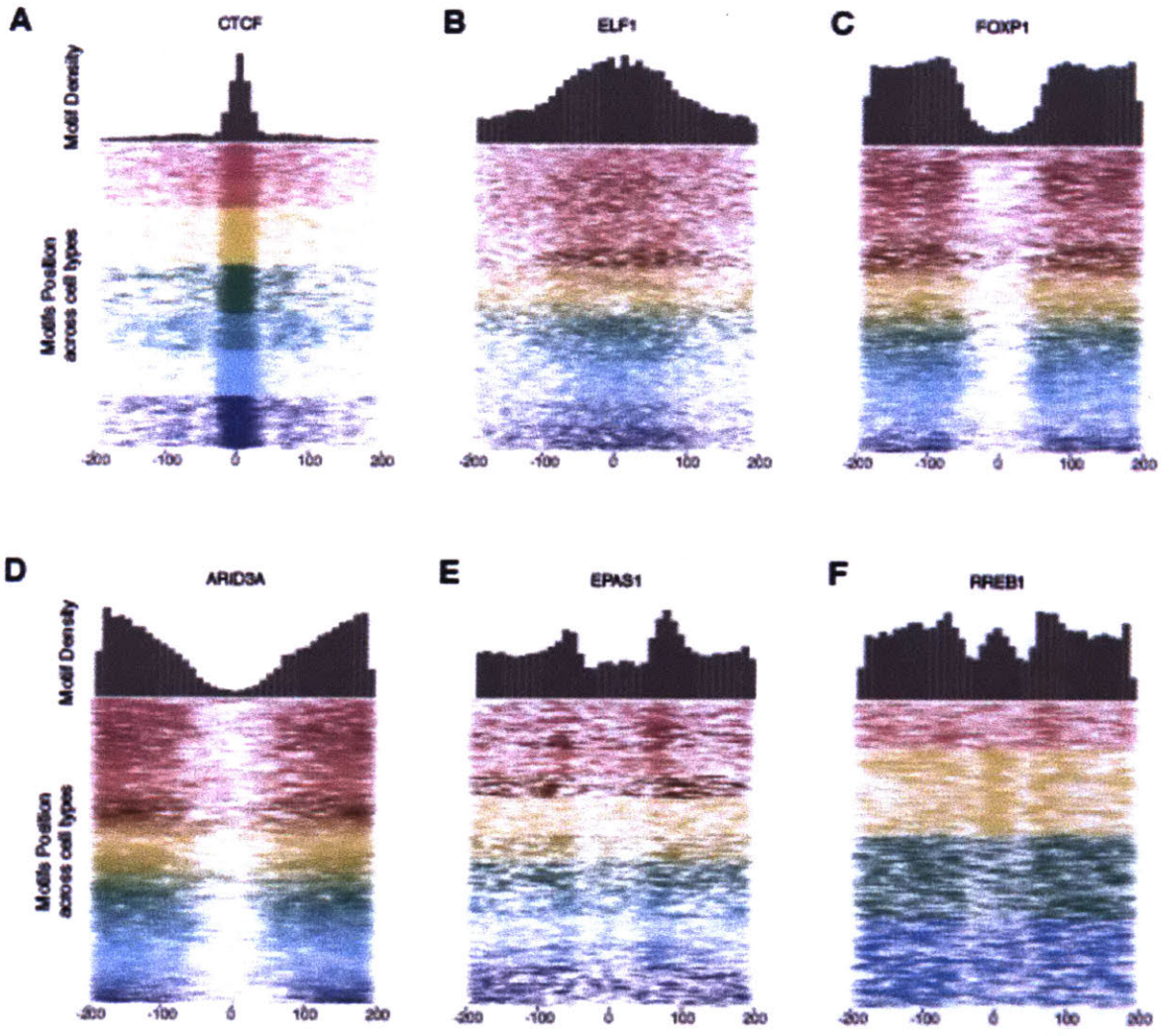


Figure 4. Positional binding patterns of TF motifs show striking differences. Distribution of position of motif sites for FOXP1 **(A)**, ARID3A **(B)**, EPAS1 **(C)**, RREB1 **(D)**, CTCF **(E)** and ELF1 **(F)** across NDR regions, centered around the peak of the DHS signal. Histograms show density of motif sites in 10 bp bins tiled across the NDR (top). Heatmaps show the position of 10,000 motif sites in NDRs, with different colors correspond to motif sites in different cell types.

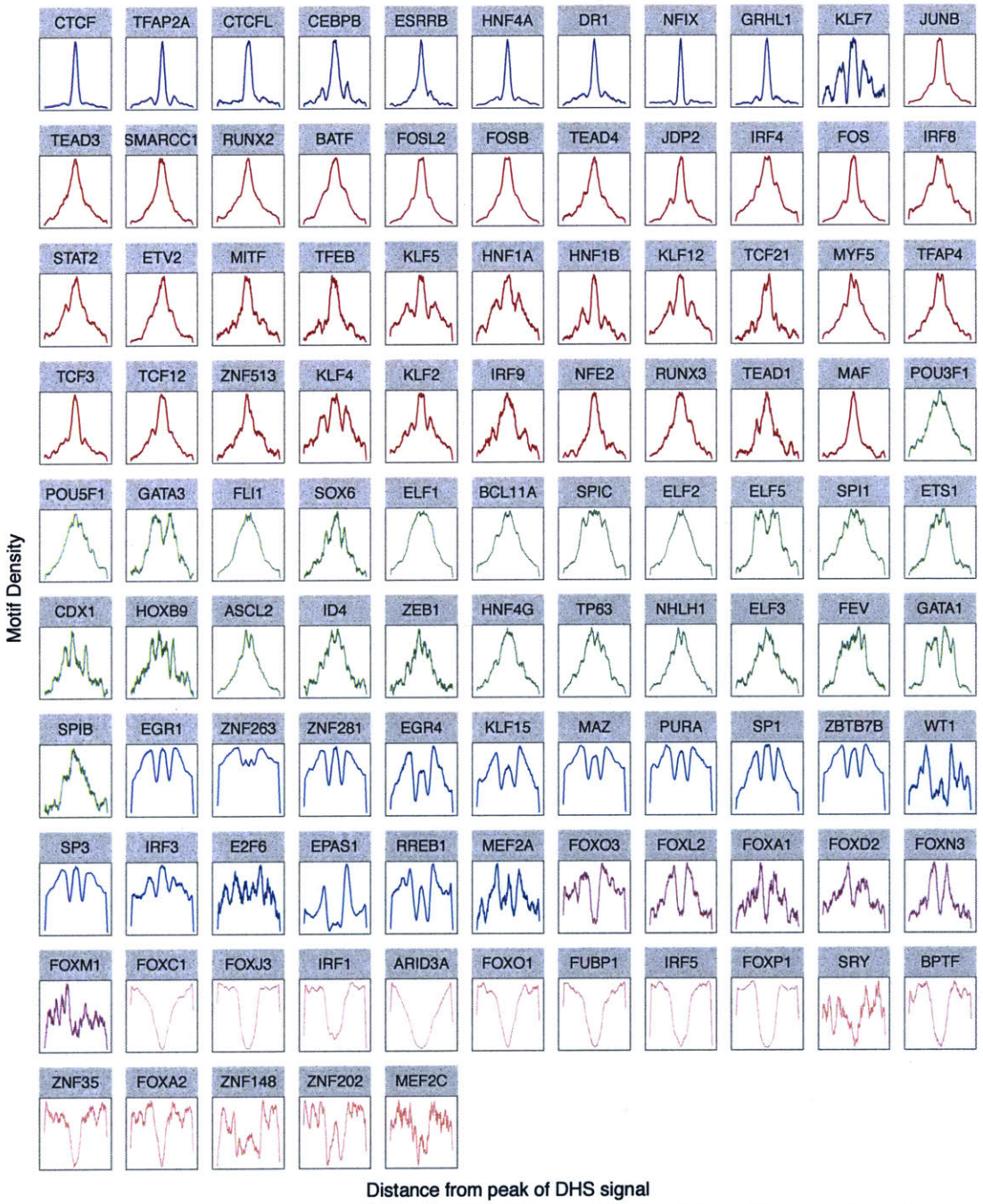


Figure S2. Motif position profiles of 103 TF motifs. Motif position profiles for each TF were calculated using NDRs in all cell types that contained the motif among the top 20 most enriched motifs. Motif sites were collapsed to their central position, and the density of motif sites in 20-bp bins tiled every 1 bp across 400 bp centered around the peak of the DHS signal is shown. Motif position profiles are colored by their cluster (blue - cluster 1; red – cluster 2; green – cluster 3; cyan – cluster 4; purple – cluster 5; orange – cluster 6).

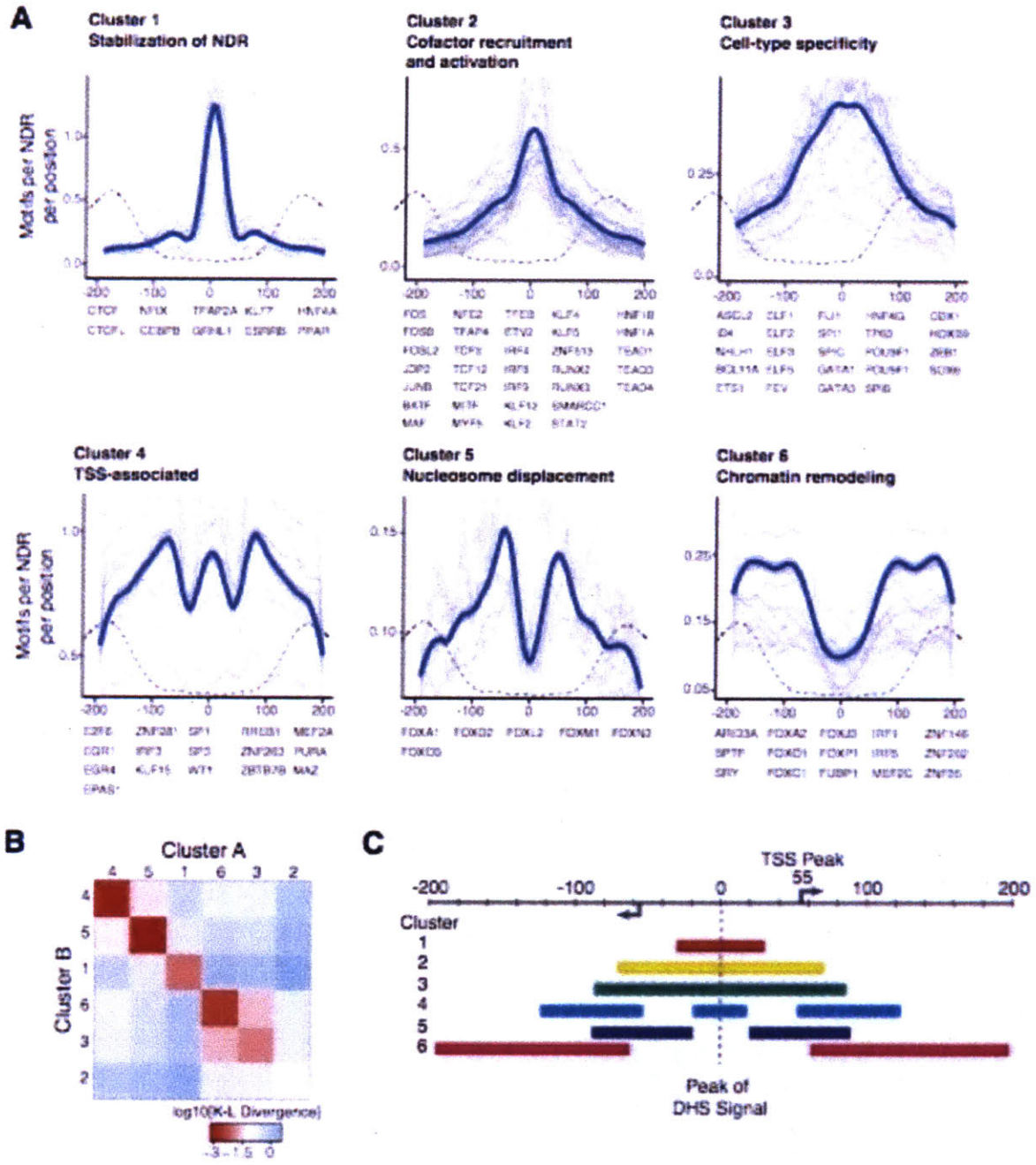


Figure 5. TF motif position patterns fall into 6 distinct clusters. (A) Motif density profiles in 400 bp regions centered around the peak of the DHS signal (gray lines) were clustered using k-mediods clustering with k=6. Density profiles were generated by calculating the frequency of motif occurrences in 20-bp bins tiled every 1 bp in the region. Blue line depicts the smoothed overall density profile of the cluster using the LOESS method. MNase-seq read density (indicating the position of the flanking nucleosomes) is shown in dotted lines for context. **(B)** Average Kullback-Leiber divergence between the motif density profiles of pairs of motifs in the same cluster (diagonal boxes) and different clusters (off-diagonal boxes). Motif density profiles within

the same cluster are substantially more similar than those in different clusters. **(C)** Schematic of NDR structure and motif positions. The arrows indicate the peak of transcriptional initiation estimated from CAGE data. The colored bars represent regions for each cluster with motif densities above the mean. Tick marks occur at 20 bp intervals.

NF-I, C/EBPβ, KLF7, GRHL1, and TFAP2), there is clear functional evidence to support this notion. For example, (i) CTCF induces stably-positioned arrays of nucleosomes around its genomic binding sites (14); (ii) NF-I, C/EBPβ, KLF7 and GRHL1 can function as pioneer factors that can establish and maintain chromatin accessibility (15-20); (iii) a recent systematic analysis of the TF-dependent chromatin accessibility changes induced by binding of 733 TFs identified CTCF, KLF7 and TFAP2 as having some of the strongest effects on local chromatin accessibility during ES cell differentiation (21); (iv) CTCF, NF-I, C/EBPβ and GRHL1 show unusually stable binding to DNA and long residence times (16, 22-24), and (v) motifs in cluster 1 have especially strong DNaseI "footprinting" signals (Fig. S4), a feature associated with a slow DNA-binding off-rate (19, 25). The properties of the 6 TFs may enable them to serve as central anchor points for adapting the surrounding chromatin, stabilizing the NDR and flanking nucleosomes (16).

The remaining 3 TFs in Cluster 1 are nuclear receptors (ESRRB, HNF4A, and PPAR). Unlike the other TFs in the cluster, nuclear receptors are characterized by transient binding to DNA with short residence times (26, 27) and localize almost exclusively to pre-accessible chromatin (18, 27, 28). Nuclear-receptor binding to genomic motif sites is often aided by 'assisted loading' by a partner factor, which binds to a site overlapping or adjacent to the nuclear receptor motif site and opens the chromatin (29). Notably, two of the pioneer TFs in Cluster 1 (C/EBPβ and NF-I) have

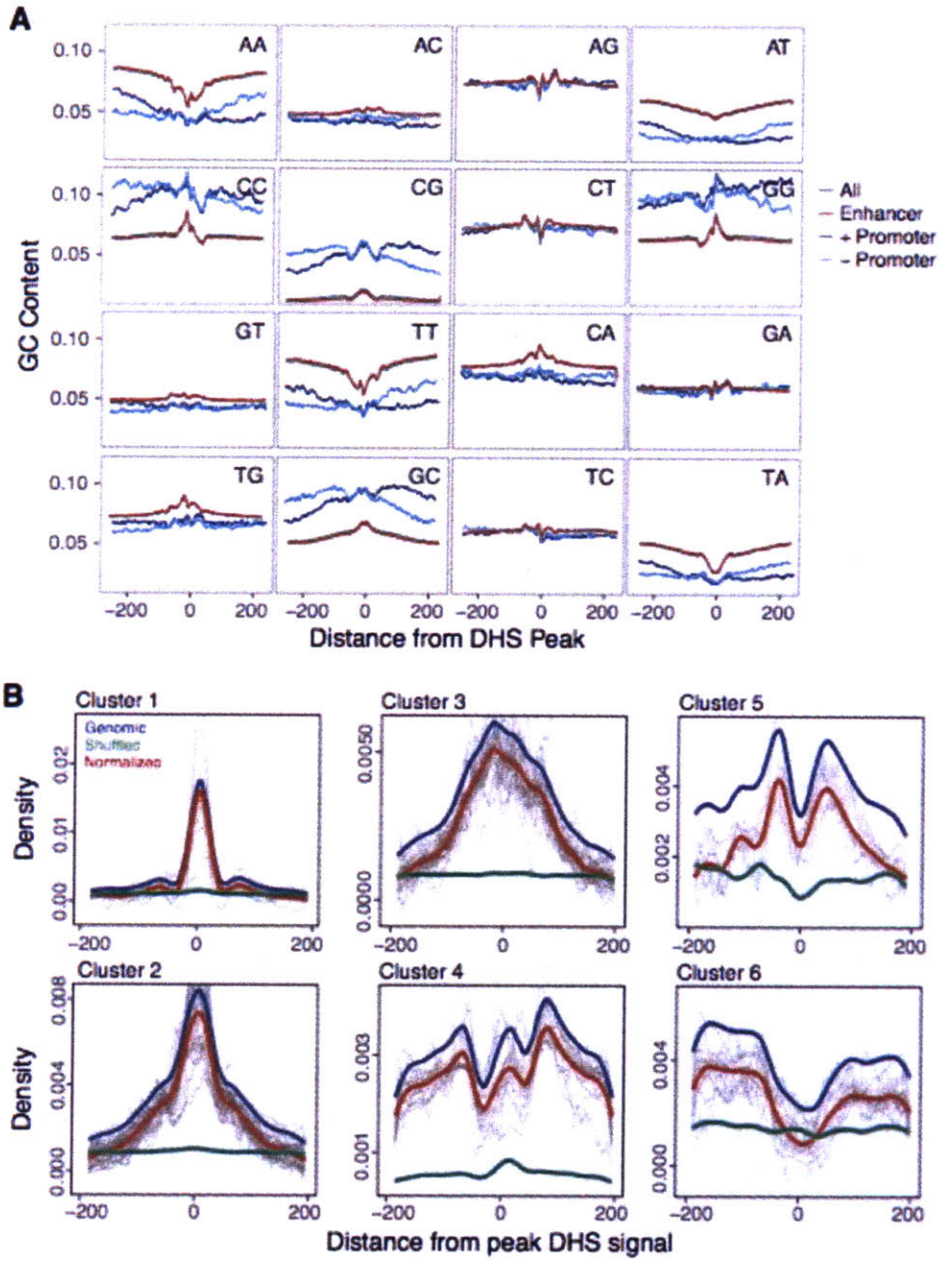


Figure S3. Variation in dinucleotide content does not account for motif position patterns. (A) Average dinucleotide content as a function of distance from the peak of the DHS signal for all NDRs (green lines), distal enhancers (>1kb from annotated TSS; yellow lines), and promoters (<1kb from annotated TSS) on the plus strand (dark blue lines) and minus strand (light blue lines). **(B)** Density profiles were normalized by subtracting the background density profiles in shuffled sequences, holding dinucleotide content at each position constant. The normalized average cluster profiles are plotted for genomic sequences (blue lines, same as Fig. 5A), shuffled sequences (green lines), and the normalized density profiles (red lines).

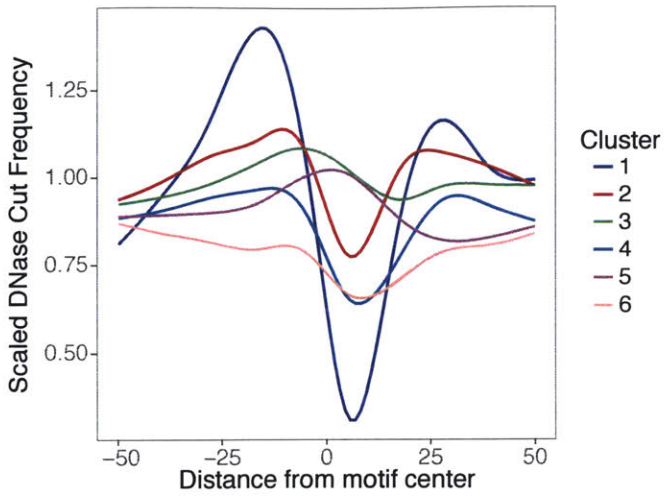


Figure S4. Average DNase footprints of motifs in each TF cluster. Cut count profiles for each TF motif were averaged over all motif sites within NDRs. Cut count profiles for each TF were scaled to a mean of zero, and averaged over all TFs in each cluster.

been shown to catalyze the assisted loading of several nuclear receptors (15, 18, 30-32). The central location of the nuclear receptor motifs may be related to assisted loading by pioneer TFs in Cluster 1.

Cluster 2 contains 31 TFs whose binding sites also are peaked at the center of the NDR, but with a wider distribution than for Cluster 1. The cluster is strongly enriched for transcriptional activators (GO category enrichment, $p_{Benjamini}=3.2x10^{-16}$), such as the activator protein 1 (AP-1) subunits (JUN, FOS, ATF, and MAF factors) and activating factors from the TCF, TEA, RUNX, IRF, and KLF families. Based on known interactions reported in the BIOGRID and INTACT databases (33, 34), these TFs are enriched for interactions with numerous transcriptional coactivators, including p300, CREB binding protein (CBP), YAP1, KDM1A, KAT2B and WWTR1 (Fig. 6, Table S1). Furthermore, the TFs in this cluster interact frequently with each other (average of 1.8 pairwise

Figure 6

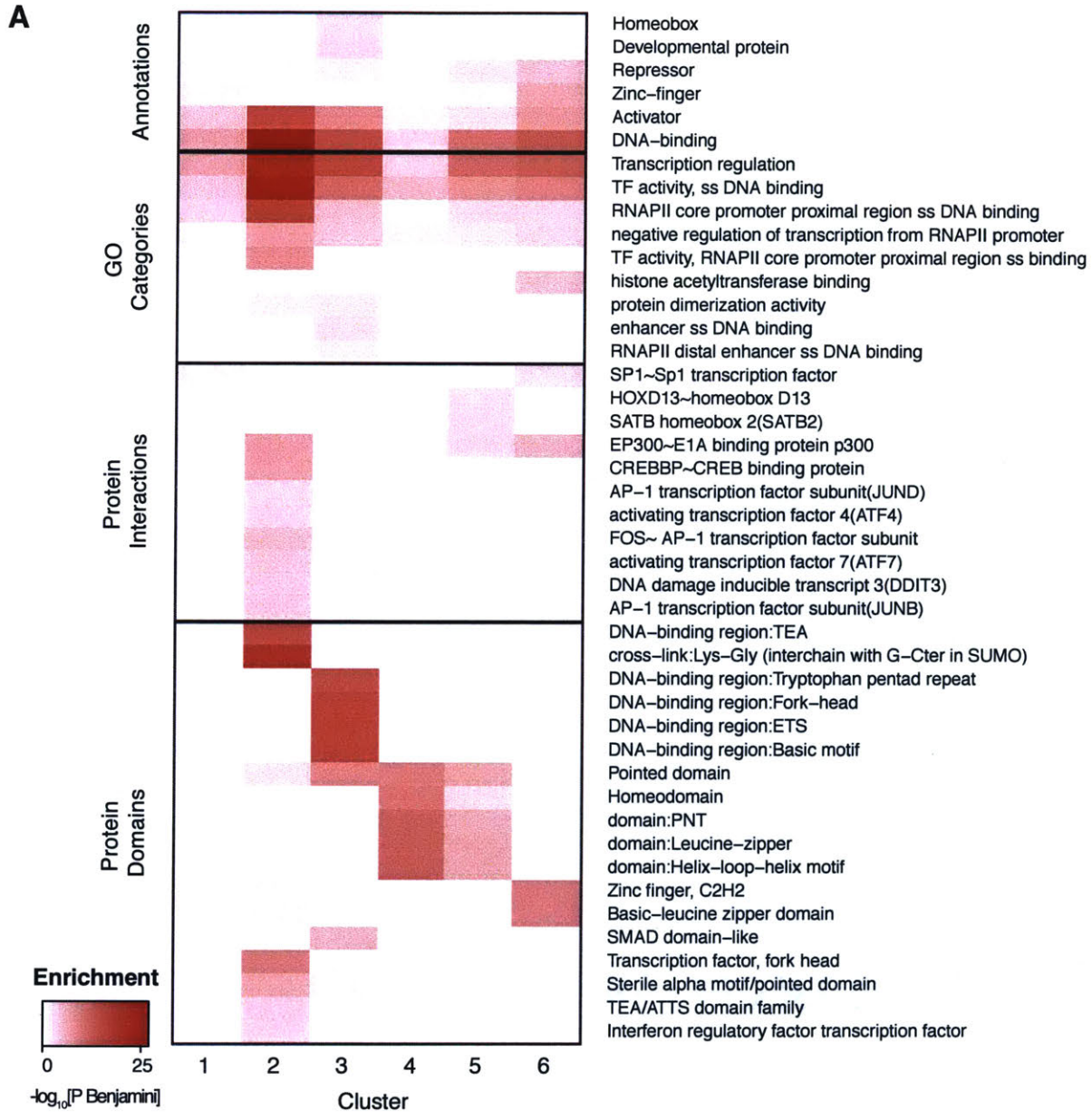


Figure 6. TF Clusters are enriched for distinct functional and structural properties. Selected enrichments for general annotation (Entrez Gene), GO categories, protein-protein interactions, and protein structural domains in the TF clusters. All terms included in the heatmap are significantly enriched ($P_{Benjamini}<0.05$) in at least one cluster. See Table S3 for all significant enrichments.

interactions among the 32 TFs vs. 0.7-1.4 (mean=1.0) among the TFs in other clusters), suggesting they could cooperatively activate transcription. For example, studies of the IFNβ enhancer have shown several TFs from this cluster (IRF3, ATF2 and Jun) bind overlapping motif sites to form a scaffold that recruits CBP/p300 through multidentate interactions (35), leading to synergistic transcriptional activation in response to viral infection (36, 37). Interestingly, TFs in cluster 2 are twice as likely to participate in signaling pathways compared to TFs in other clusters (52% of TFs in cluster 2 vs. 16-30% of TFs in other clusters, based on the KEGG database (38))(Fig. 6, Table S1). The tightly clustered pattern of these motifs in this cluster may therefore promote cooperativity by both facilitating TF-TF interactions and positioning TFs to form complexes that contact multiple sites on cofactors, thereby allowing enhancers to link multiple signaling pathways and respond in a highly synergistic fashion to specific regulatory cues.

Cluster 3, which contains 25 TFs, is the final group that is peaked at the center of the NDR, albeit with a much broader distribution. These TFs are characterized by far greater cell-type specificity in expression and motif enrichment than the TFs in other clusters (Fig. 3B, 6D). Consistent with this observation, Cluster 3 contains numerous TFs that play critical roles in development, including all the Homeobox, POU, SOX, ETS and GATA factors in our dataset (39-44) (Table S1). Furthermore, 20 of the 23 TFs have functional annotations in GO related to differentiation and development in a wide range of tissues (Fig. 6A), including erythrocytes (GATA1, GATA3, ETS1 and SPI1), osteoblasts (TP63 and ID4), keratinocytes (TP63 and POU3F1), blastocysts (SPIC, POU5F1, and ELF3), neurons (ASCL2, FEV and ZEB1), and more. Compared to TFs in

Cluster 2, the TFs in Cluster 3 have fewer annotated interactions with cofactors and other TFs (average 5.2 vs. 10.2 per TF). Together, the broader motif distribution and smaller number of predicted interactions suggest these TFs might participate in fewer physically-mediated cooperative interactions, and instead function more independently or through indirect cooperation with other factors.

Cluster 4, which contains 16 TFs, is unusual in several respects. The motif profiles show both a central peak and flanking peaks, at ~70 bp upstream and downstream. Moreover, all of the motifs in this cluster are asymmetric, with motif occurrences in the flanking peaks showing a clear preferred orientation relative to the center of the NDRs—that is, one of the reverse-complementary sequences defining the motifs preferentially points inward (Fig. S5).

The motifs in Cluster 4 are also strongly enriched in promoter-proximal regions (the 6% of such NDRs contain 12% of occurrences for motifs in Cluster 4 vs. 1-3% for other clusters). (Fig. 7A). ENCODE ChIP-seq data for 39 TFs in our dataset show greater enrichment in promoter regions for TFs in Cluster 4 than for TFs in other clusters (38% of reported peaks within <1 kb of a TSS vs. 8-28% for other clusters; Fig. 7B). One of the TFs in this cluster, SP1, is a well-characterized promoter-proximal factor that binds GC-rich elements in a wide variety of cellular and viral promoters, and many of the other TFs in the cluster (including SP3, EGR1, EPAS1, ZBTB7B, E2F, KLF15, MEF2C, WT1 and PURA) are known to interact with SP1 at promoters (45-55). We compared the motif density profiles in NDRs classified as promoter-proximal versus distal enhancer, but found them to be indistinguishable (Fig. S6).

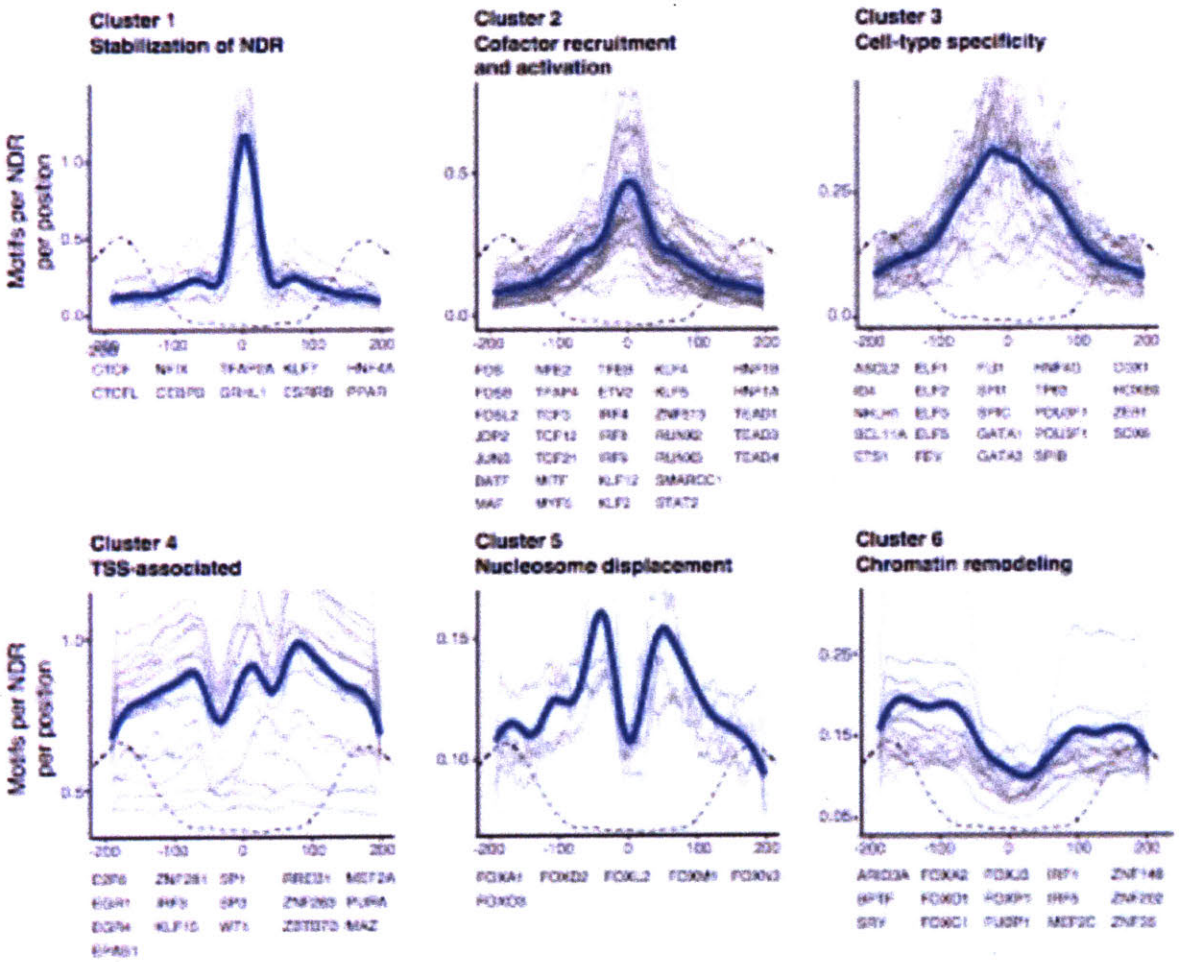


Figure S5. Directional TF motif density for the 6 clusters. Motifs were aligned on the same strand, and motif density profiles in 400 bp regions centered around the peak of the DHS signal (gray lines) were calculated for each TF in each of the 6 clusters. Blue line depicts the overall density profile of the cluster, calculated by fitting a loess model with span=0.2. MNase-seq read density (indicating the position of the flanking nucleosomes) is shown in dotted lines for context.

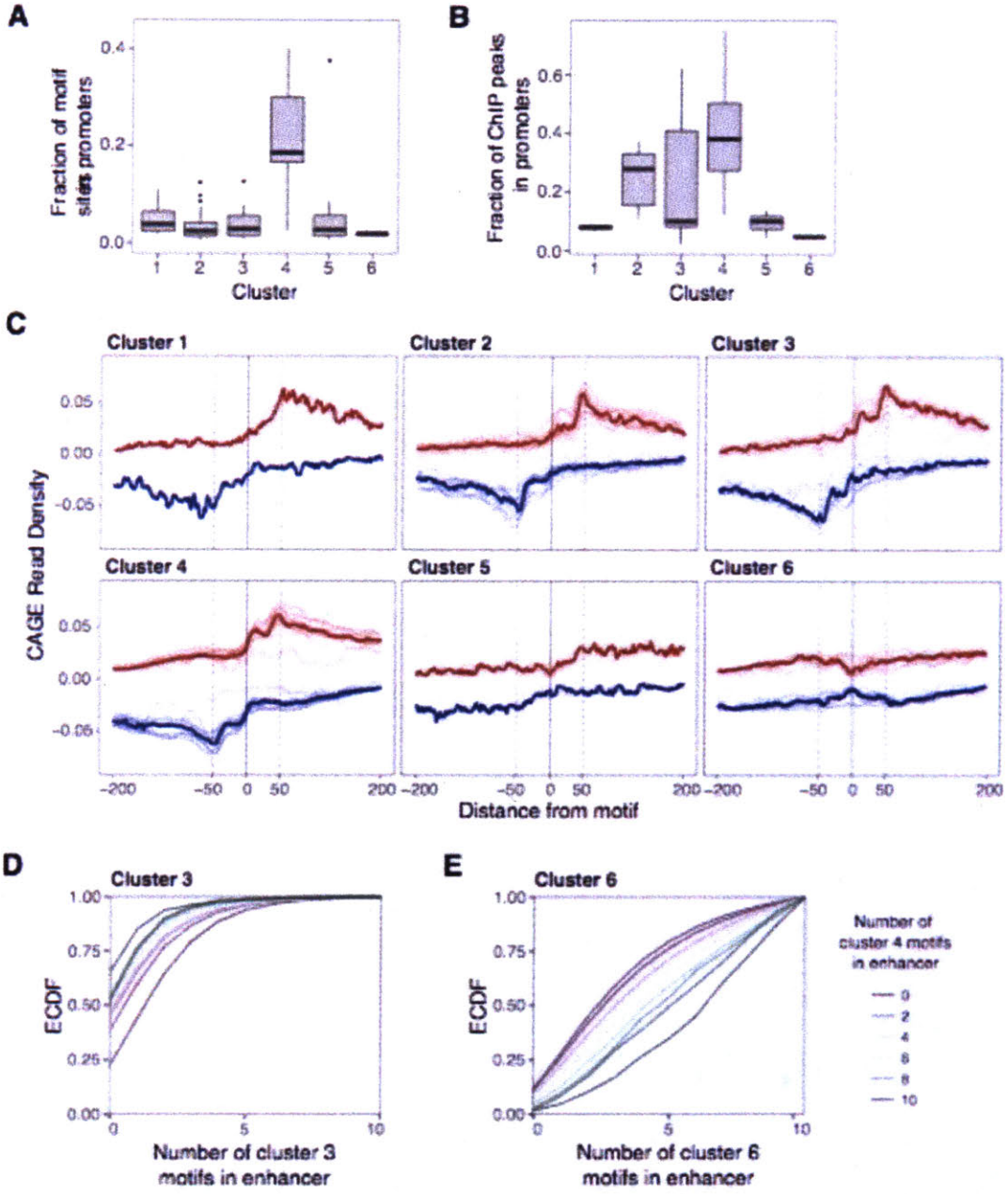


Figure 7. TFs in cluster 4 are enriched in promoters and associated with transcriptional initiation. (A) Fraction of motif sites in NDRs in our analysis that occur in promoters (<1 kb upstream of annotated TSS) for TFs in each cluster. **(B)** Fraction of ChIP-seq peaks for TFs in each cluster that overlap promoter (data for 39 TFs profiled in ENCODE is included). Cluster 4 motif occurrences and TF binding occur in promoters far more frequently than motifs in other clusters. **(C)** Composite of CAGE reads on the plus strand (red) and minus strand (blue) aligned to the center of each TF motif. Thin red and blue lines correspond to CAGE profiles of individual TF motifs, and thick red

and blue lines show the average CAGE profile of all motifs in the cluster. Motifs in clusters 3 and 4 show a peak of transcriptional initiation at the location of the motif site. **(D, E)** Empirical cumulative distribution function (ECDF) of the number of cluster 3 (D) and cluster 6 (E) motif sites in NDRs, conditional on the number of cluster 4 motifs. NDRs with cluster 4 motif sites are co-enriched for cluster 3 motif sites and depleted for cluster 6 motif sites.

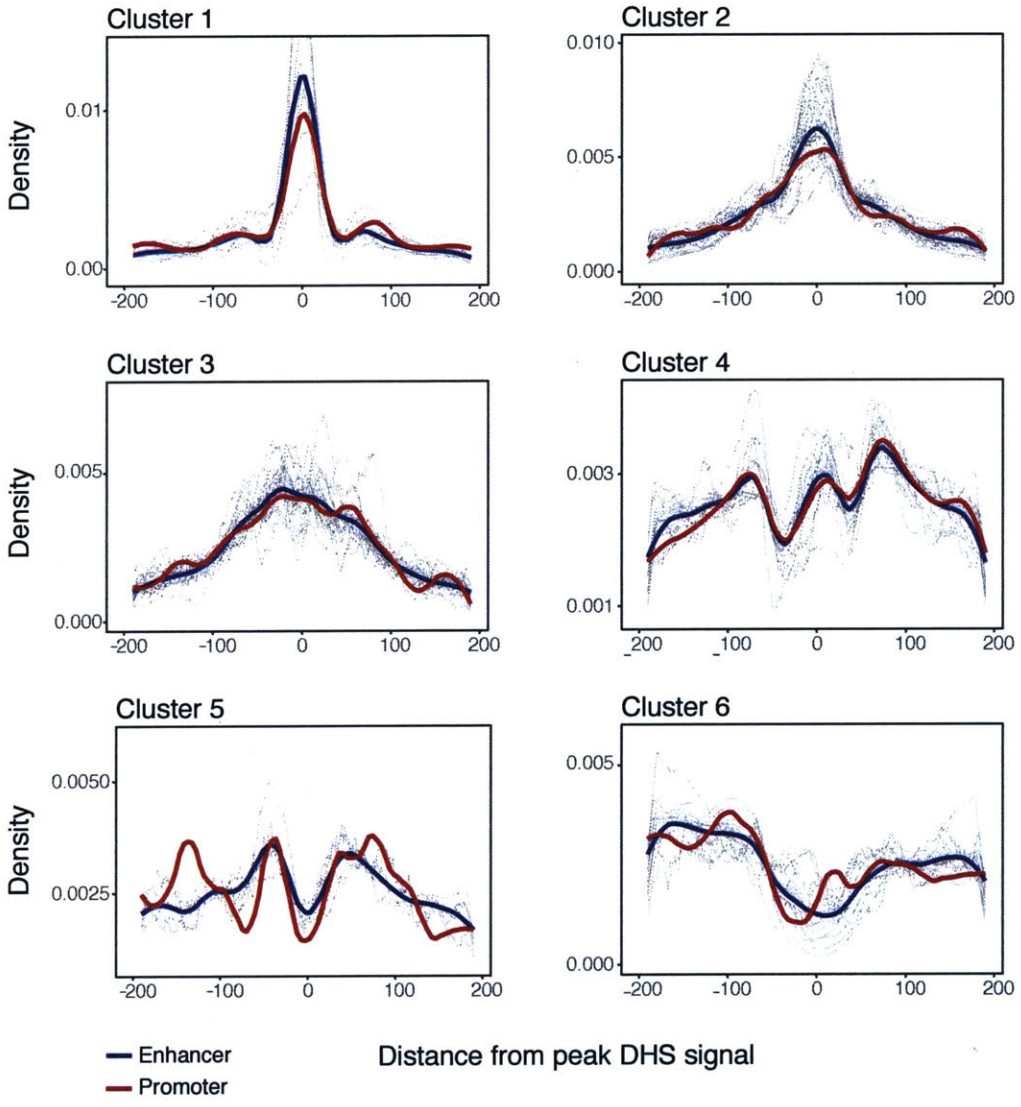


Figure S6. Motif density profiles at promoters versus enhancers. Motif sites for each NDRs were split into promoters (<1 kb from an annotated TSS; red lines) and enhancers (>1 kb from an annotated TSS; blue lines). TF were aligned on the same strand, and the density of motif sites in 20-bp bins tiled every 1 bp in 400 bp regions centered around the peak of the DHS signal was calculated.

TFs in this cluster are also enriched for interactions with p300 (FDR=2.6 x 10^{-6})) and DNMT1, a DNA methyltransferase that methylates CpG dinucleotides in promoter regions (FDR=0.03). Notably, functional studies have demonstrated that SP1 stimulates transcription when bound close to the initiation site but not in distal positions (56, 57), unlike distal enhancer binding factors from Clusters 1-3. These results suggest that SP1 and other TFs in cluster 4 may belong to a distinct functional class of TFs with specialized promoter-associated functions.

Because a key function of promoters is transcript initiation, we hypothesized that the flanking peaks and orientation of TFs in Cluster 4 might reflect a role in establishing or stabilizing transcription start sites (TSS) at both promoters and enhancers. Recent studies have suggested that, in addition to such features as TATA boxes and INR elements, TF binding sites also contribute to determing the position of TSS (11, 58). To examine the relationship of TFs in each cluster with TSS, we examined CAGE (cap analysis of gene expression) data for both enhancer and promoter-proximal NDRs for 14 of the cell lines in our dataset (59). Transcriptional initiation tends to peak at 50-60 bp from the center of the NDRs (as noted above) (Fig. 2C), and ~50 bp away for TF motif occurrences. However, 64% TFs in cluster 4 (vs. 0-14% in other clusters) show an additional peak of transcriptional initiation ~10 bp away from the location of motifs sites (EGR1, EGR4, MAZ, PURA, SP1, SP3, ZBTB7B and ZNF281; Fig. 7C). This observation suggests these TFs play unique roles in positioning the site of initiation.

Cluster 5 contains 6 TFs whose binding sites are not enriched at the center of NDRs, but have peaks at ~60 bp upstream or downstream. The TFs in this cluster all belong to the FOX family of TFs and include the two best-characterized pioneer factors,

FOXA and FOXO. The DNA-binding domain (DBD) of FOX factors structurally resembles the DBD of linker histones H1 and H5 (60, 61), and FOXA factors can compete for binding to linker histone binding sites, which are located near the edges of the core nucleosome, ~65 bp away from its center (60, 62-64). But, whereas linker histone binding leads to compaction of nucleosomal arrays, FOXA binding destabilizes nucleosomes and opens the region for binding by other TFs (65-67). Since enhancer activation typically entails the elimination of a well-positioned central nucleosome (68), motif sites for FOXA and other FOX factors in cluster 5 may be positioned at ±60 bp to displace linker histones and destabilize the central nucleosome, helping other TFs bind their target sites.

Finally, Cluster 6 contains 14 TFs with binding sites enriched near the edges of the accessible region (80-200 bp from the center), suggesting these TFs could interact with the surrounding chromatin. As with Cluster 4, the TFs in cluster 6 have asymmetric motifs and mostly exhibit a preferred orientation relative to the center of the region (Fig. S3), allowing for directional interactions with the surrounding nucleosomes and larger chromatin landscape. Consistent with this notion, 10 of the 14 TFs in Cluster 6 are known to play roles in chromatin remodeling, including BPTF, the DNA-binding subunits of nucleosome remodeling factor (NURF), which recognizes H3K4me3 and facilitates ATP-dependent nucleosome sliding (69-71); ARID3A, which facilitates the opening of the IgH enhancer (72-74); and several FOX factors, which interact with histones and mediate recruitment of chromatin remodeling complexes such as SWI/SNF (67, 75). Many of the motifs in this cluster are A/T-rich (Fig. S3). It is possible that they also recruit additional members of the ARID (A+T rich interaction domain) family that binds

non-specifically to A/T sequences and has been implicated in chromatin remodeling, including ARID1A/BAF250, the DNA-binding subunit of the BAF chromatin remodeling complex (76).

The TFs in Cluster 6 also play roles in nuclear attachment, DNA bending, and DNA unwinding. These TFs are enriched for interactions with the chromatin organizers SATB1 and SATB2, which induce chromatin looping and tether DNA to the nuclear matrix (77, 78). For example, ARID3A binds to sites on the periphery of the IgH enhancer to mediate attachment the nuclear matrix (79). Several of the TFs (ARID3A, SRY and YY1) induce significant DNA bending (73, 80, 81), facilitating TF binding and TF-TF interactions (82, 83). Finally, some (SRY and FUBP1) unwind the DNA double helix, which can promote transcriptional initiation and attachment to the nuclear matrix (81, 84, 85).

To directly test whether TFs in Cluster 6 interact with the surrounding nucleosomes, we used MNase-seq data from two cell types (GM12878 and K562) to infer the position of the flanking nucleosomes for each individual NDR and then aligned the TF motifs. For the TFs present in these cell types, we examined the motif distribution relative to the inferred edge of the flanking nucleosome (rather than to the peak of the DHS signal). Whereas the TFs in Clusters 1-5 did not show peaks of motif sites adjacent the nucleosome edge, 5 of the 8 TFs in Cluster 6 (FOXC1, FOXJ3, FOXO1, FOXP1, and ARID3A) showed a peak (Fig. S7). The remaining 3 TFs (FUBP1, IRF1, IRF5) are not known to play roles in chromatin remodeling.

Finally, we wondered whether certain classes of TFs tend to co-occur in enhancers. To investigate this, we examined whether the distribution of motif sites from

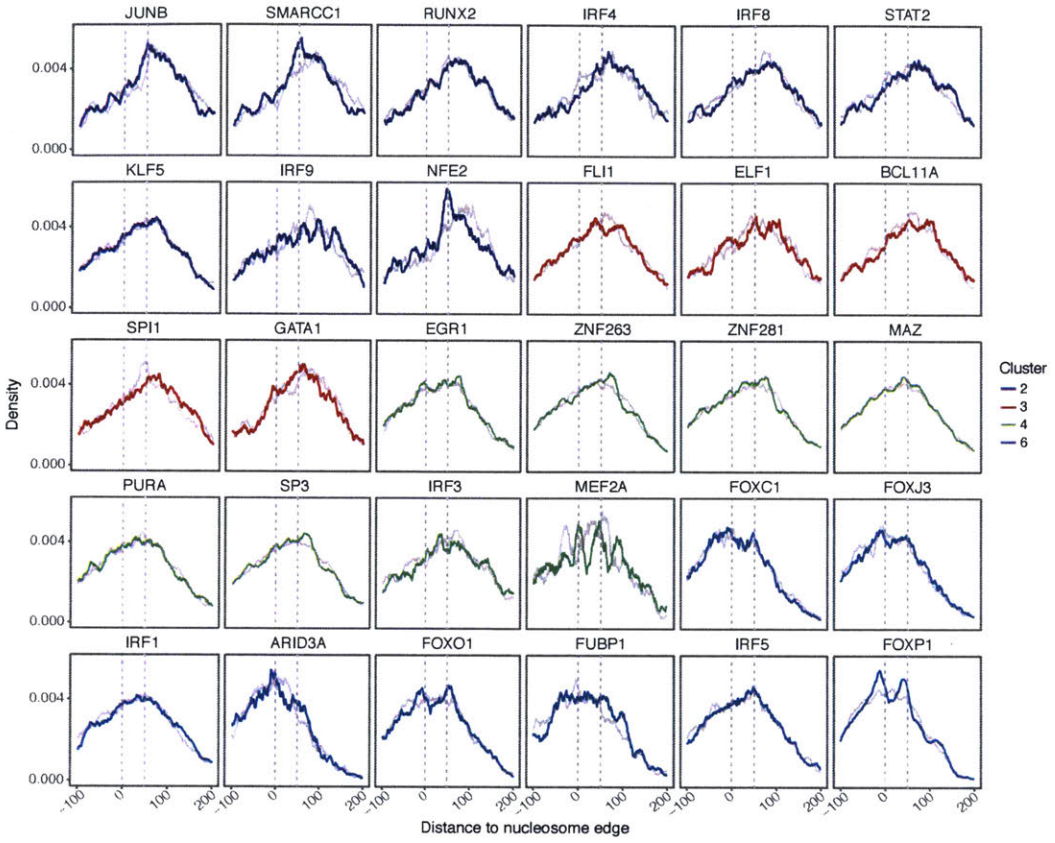


Figure S7. Motif density relative to flanking nucleosome edge. The edges of the nucleosomes flanking each NDR were estimated from MNase-seq data (see Methods), and TF motifs were aligned relative to the nearest flanking nucleosome. Density plots represent the frequency of motif sites in 20-bp bins tiled every 1 bp in genomic NDRs (colored lines) and permuted data (gray line; permuting nucleosome position across motif sites).

each class in the NDRs varied with the presence or absence of motif sites from each of the other classes (Fig. 7D,E, Table S3). We counted the number of non-overlapping motif sites from each cluster in the NDRs and calculated the odds ratio (OR) for co-enrichment between the motif from each pair of clusters. To control for motif similarities, we also calculated the baseline OR for each pair of clusters in shuffled sequences. Significantly coenriched or codepleted cluster pairs were defined as pairs for which the

OR falls outside the 95% confidence interval of the OR in shuffled sequences (Table S3). We found that all six clusters showed significant preferences for co-enrichment and co-depletions with specific other clusters (Table S3). For example, regulatory elements with TF motif sites in cluster 4 (associated with TSS-related functions) contain significantly more TF motif sites from cluster 3 (associated with cell-type-specific activation, Fig. 7D) and significantly fewer motif sites from cluster 5 and 6 (associated with nucleosome remodeling and chromatin architecture, Fig. 7E) than regulatory elements without cluster 4 motifs. Importantly, these cluster associations are consistent across cell types, even though the specific set of TFs active in each cell type differs (Fig. S8). Thus, the TF clusters may constitute a general regulatory code, with different cell types substituting specific TFs to activate different sets of enhancers.

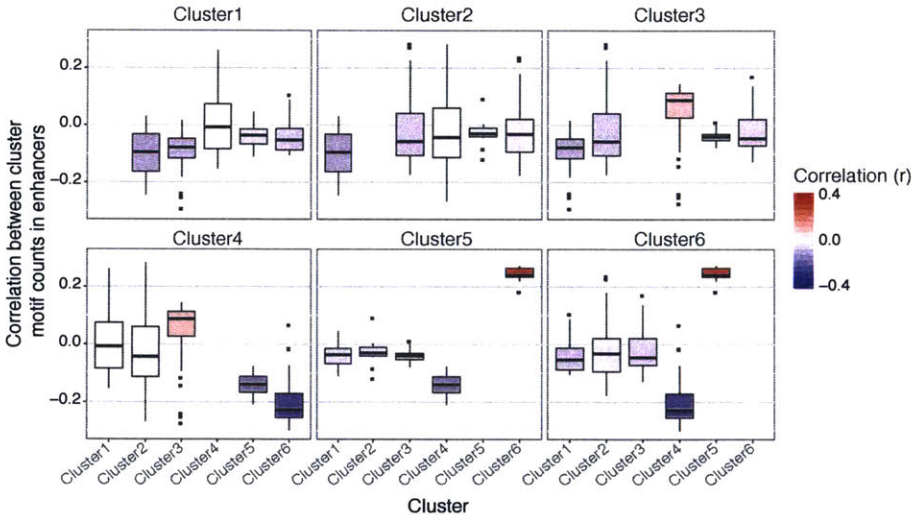


Figure S8. Co-enrichment and -depletion of TF motif clusters is constant across cell types. For each pair of clusters, the correlation (Spearman ρ) between the number of motif sites in NDRs from cluster A and cluster B was calculated in each cell type. Positively correlated cluster pairs tend to occur together in NDRs, whereas negatively cluster pairs are co-depleted. Boxplots depict the range of ρ values for each cluster pair across the 47 cell types. The correlations of motif counts from each pair of clusters is relatively constant across cell types, indicating the clusters may represent general functional classes of TFs that interact similarly in different cellular contexts.

DISCUSSION

It has long been suggested that transcription factors may belong to different functional classes. In some cases, prior biological knowledge of certain transcription factors has been used to categorize TFs into classes such as pioneer factors that have the capability to bind motif sites in closed chromatin versus non-pioneer factors that only bind motif sites in open chromatin, and cell-type specific versus ubiquitous factors. However, there have been few systematic approaches to recognize distinct classes and properties independent of known biological properties of the individual transcription factors. One such functional study was recently performed in Drosophila, in which investigators asked which transcription factors could substitute for each other across a variety of regulatory contexts (12).

Here, we show that solely by looking at the positional distribution of motif sites within NDRs, we are able to recognize six distinct classes of TFs. These classes bring together factors that have a number of similar properties, such as binding stability, interactions with other TFs and cofactors, cell-type specificity, and pioneering ability. Furthermore, the position of motif sites appears to be related to their known functions—for example, localizing pioneer factors to the optimal positions to displace nucleosomes and targeting chromatin remodelers in close proximity to flanking nucleosome.

The degree to which the arrangement of motif sites within regulatory elements determines their function remains an open question. At one end of the spectrum, there are examples of "enhancesomes", such as the IFNβ enhancer, that are exquisitely sensitive to spacing and orientation of the motif sites (35, 86, 87). However, the activity of other regulatory elements, referred to as "billboard" enhancers, appears to be

relatively insensitive to the arrangement of motif sites (88-90). Instead, our work suggests a different kind of constraint, whereby TFs play distinct roles in forming a functional enhancer, facilitated by their position within a regulatory sequence.

The classes identified here also help shed light on properties of some less characterized TFs. For example, they suggest that several other FOX factors in cluster 5 may use a similar mechanism to FOXA1 to displace nucleosomes, and that the uncharacterized zinc finger TFs in cluster 6 (ZNF148, ZNF202 and ZNF35) may have pioneering abilities. In addition, the positional preferences identified may prove useful in building predictors of enhancer activity and recognizing functional enhancers in genomic sequence.

While here we focused on the classes of transcription factors, these results naturally raise the question of whether there are different functional classes of enhancers formed based on these classes of transcription factors. Identifying such enhancer classes may shed light on the classes of transcription factors that must come together to accomplish all the functions necessary to build a functional enhancer. Finally, in addition to helping to understand natural enhancers, better knowledge about the constraints and functional implication of TF positions may aid in creating synthetic enhancers with specific properties that can be used in synthetic biology.

ACKNOWLEDGEMENTS

We would like to thank Karen Adelman, Telmo Henriques, Bradley Bernstein, Aviv Regev, Cigall Kadoch, Seth Cassel, and Kaylyn Williamson for valuable comments and discussion, and Ray Louis for help with DNase footprint analysis. This work was supported by the National Human Genome Research Institute (2U54HG003067-10)

(E.S.L.) and the National Institute of General Medical Sciences (T32GM007753) (S.R.G.).

MATERIALS AND METHODS

ATAC-seq

Jurkat and U-937 cells were either left unstimulated or were stimulated for 1 or 4 hours with 2.5ug/ml anti-human CD3 (Biolegend; Cat# 317315) and 50 ng/ml PMA (Sigma Aldrich; Cat# P1585-1MG) for Jurkat cells and 100 ng/ml LPS (Invivogen; Cat# tlrl-peklps) for U-937. Cells were washed with ice cold FACS Buffer and kept on ice until cell sorting. 25,000 live cells from each condition were sorted in to FACS Buffer and pelleted by centrifugation at 500 RCF for 5 minutes at 4C in a pre-cooled fixed angle centrifuge. Cell lines were then tagmented according to the previously described Fast-ATAC protocol (91). Briefly, all supernatant was removed being careful to not disturb the not visible cell pellet. 50 ul transposase mixture (25 ul of 2x TD, 2.5 ul of TDE1, 0.5 ul of 1% digitonin, 22 ul of nuclease-free water) (Cat# FC-121-1030, Illumina; Cat# G9441, Promega) was added to the cells, the pellet was dissociated by pipetting. Transposition reactions were incubated at 37C for 30 minutes in an Eppendorf ThermoMixer with agitation at 300 RPM. Transposed DNA was purified using a QIAgen MinElute Reaction Cleanup kit (Cat# 28204) and purified DNA was eluted in 12 ul elution buffer (10 mM Tris-HCl, pH 8). Transposed fragments were amplified and purified as described previously (92) with modified primers (93). Libraries were quantified using qPCR prior to sequencing. All Fast-ATAC libraries were sequenced using paired-end, dual-index sequencing on a NextSeq with 76x8x8x76 cycle reads at an average read depth of 30 million reads per sample.

Definition of NDRs

To define NDRs for our analysis, we used DNaseI-seq and H3K27ac ChIP-seq data for 45 cell types in the Epigenomics Roadmap and ENCODE Projects (6, 59), as well as ATAC-seq and H3K27ac ChIP-seq data for Jurkat and U937 cells generated in our lab. To select our initial set of NDRs, we intersected DHS/ATAC-seq narrowPeaks regions and H3K27ac gappedPeaks regions called using MACS2 (94) with the standard parameters used by the Epigenomics Roadmap Project. We then filtered out NDRs that were present in more than 24 (50%) of the cell types in our analysis, and selected the top 7,500 cell-type-restricted NDRs for motif enrichment and positioning analysis. We defined the coordinates in the NDRs relative to the summit called by MACS2 (i.e. position with the the maximum DHS/ATAC-seq signal). For MNase-seq analysis, we used data from GM12878 and K562 generated by the ENCODE project. The center of the nucleosomes flanking the NDRs were estimated by identifying the position with the highest MNase-seq read coverage in the 300 bp upstream and downstream of the peak of the DHS signal.

Motif enrichment analysis

We calculated motif counts for all vertebrate motifs in TRANSFAC (95), JASPAR (96) and CIS-BP (97) in the genomic NDR sequences, as well as scrambled genomic NDR sequences (holding dinucleotide frequencies constant). To identify enriched motifs in each cell type, we used AME (98) with the mhg method to calculate the enrichment of total number of matches of each motif in the genomic sequences compared to the scrambled sequences. In cases where the combined databases contained multiple PWMs corresponding to a single TF, we selected the most enriched motif in each cell

type corresponding to each TF. To remove highly similar motifs, we calculated the pairwise similar of the motifs using the R package PWMEnrich, and removed motifs that had similarity of > 0.8 with a more highly enriched motif. We then selected the top 20 motifs from the filtered list in each cell type for positioning analysis. We called motif sites in the genomic and scrambled sequences using by running FIMO (99) with a p-value threshold of 10^{-4}.

Motif position profiles and clustering

To analyze the positioning of the motifs with NDRs, we collapsed the motif matches to their central position, and calculated the density of each motif in 20-bp windows tiled every 1 bp across the 400 bp centered around the position of maximum DHS/ATAC signal in each NDR. The motif position profiles were then clustered using the pam function from the R package cluster with k=6.

To assess how much each motif position profile is due to the varation in dinucleotide content across the regions,we calculated the background motif density profiles in shuffled sequences, holding the dinucleotide content at each position constant, and normalized the genomic density profiles by subtracting out the background motif frequencies (Fig. S3).

TF co-enrichment analysis

We tested for co-enrichment and -depletion of motifs from the 6 TF motif clusters in genomic NDR sequences using a Fisher exact test. For each pair of cluster A and cluster B, we calculated the odds ratio that a genomic sequence contains a motif from cluster B, conditional on the presence of a motif from cluster A. To control for motif

similarities between motifs in different clusters, we also calculated the same odds ratio in scrambled sequences (holding dinucleotide content constant). To identify significantly co-enriched or -depleted pairs, we selected pairs for which the 95% confidience interval of the genomic OR did not overlap the 95% confidience interval of the shuffled OR.

TABLES

Table S1. TF motifs clusters identified based on the positional distribution of motif sites within NDRs

	Family	TF	Motif Symmetry	Cofactor Binding	AA Bias	Median Enrichment	Number of cell types with enrichment	Median Expression [Range]
Cluster 1	AP2	TFAP2A	S			2.20	16	84.8
	CEBP	CEBPB				3.39	50	n.d.
	CTCF	CTCF				2.43	57	35.2 [30.7-39.8]
		CTCFL		TBP		1.57	1	n.d.
	GRHL1	GRHL1	S			4.39	2	n.d.
	KLF	KLF7				2.66	40	n.d.
	NFI	NFIX	S			2.30	36	54.5
	NR	DR1				2.23	57	21.3
		ESRRB				3.30	1	n.d.
		HNF4A		TBP		3.84	13	132.9 [97.6-170.9]
Cluster 2	AP-1	FOS	S	p300, CBP, SWI/SNF, RB1		4.02	51	17.2 [15.8-214.2]
		FOSB				4.34	29	2.4 [0.1-90.5]
		FOSL2	S			5.79	44	25.8 [12.7-45.2]
		JDP2	S			6.15	26	n.d.
		JUNB	S	p300, SWI/SNF		3.07	54	64.9 [2.2-329.2]
		BATF		KDM1A		2.58	11	0.0 [0.0-0.3]
		MAF		CBP, HDAC		2.86	29	n.d.

		NFE2		p300, CBP	Poly-pro	2.52	5	143.2
		TFAP4	S	SWI/SNF, HDAC	Poly-pro	2.42	42	2.5
	bHLH	TCF3		p300, CBP, HDAC, KAT2B, KDM1A, RB1		2.14	56	n.d.
		TCF12	S	p300, CBP		2.32	57	5.9
		TCF21	S			3.26	13	n.d.
		MITF				2.41	14	63.9
		MYF5	S			2.63	1	2.6 [0.0-558.1]
		TFEB	S		Poly-pro	3.75	19	4.4
	ETS	ETV2				3.76	0	0.5 [0.4-0.6]
		IRF4				2.07	11	0.0 [0.0-229.4]
	IRF	IRF8				2.74	14	104.6 [4.4-204.8]
		IRF9				3.50	46	43.7
		KLF12				3.05	20	2.4 [1.1-6.7]
		KLF2		KAT2B	Poly-ala, Poly-pro	2.66	34	15.5 [0.1-31.0]
	Zinc finger, C2H2	KLF4		p300, CBP, HDAC, SWI/SNF	Poly-pro	2.20	33	18.8 [17.9-19.6]
		KLF5		p300, CBP, YAP1, HDAC		2.23	29	65.3 [0.5-101.7]
		ZNF513				2.50	53	n.d.
	RUNX	RUNX2		p300, RB1	Poly-ala	2.36	8	3.3 [0.0-7.8]
		RUNX3		p300		3.33	12	0.8
	SMARCC	SMARCC1	S	SWI/SNF, KDM1A	Poly-pro, Poly-ala	3.12	57	42.1 [41.5-42.6]
	STAT	STAT2		p300, CBP, SWI/SNF		3.46	56	35.4
	TCF	HNF1B	S			4.50	9	7.9 [6.1-9.6]

		HNF1A	S	p300, CBP, KAT2B		4.13	7	8.4 [4.5-19.1]
	TEAD	TEAD1			Poly -pro	2.34	48	45.7
		TEAD3		YAP1		3.20	45	97.5 [34.8-160.2]
		TEAD4		YAP1		2.95	42	48.6
Cluster 3		ASCL2				3.23	4	0.0 [0.0-0.0]
	bHLH	ID4				2.26	31	24.8
		NHLH1				3.08	2	0.3
	BCL11A	BCL11A				2.59	10	0.0 [0.0-12.9]
	ETS	ETS1				3.41	49	31.3
		ELF1				3.01	53	28.9 [16.2-144.6]
		ELF2				2.47	41	9.3 [4.6-14.0]
		ELF3				2.67	18	0.0
		ELF5				2.86	0	0.0 [0.0-0.1]
		FEV				3.63	3	0.0
		FLI1				2.95	15	1.9 [0.0-52.4]
		SPI1				2.40	18	23.7 [4.0-43.5]
		SPIC				3.84	1	n.d.
	GATA	GATA1				4.54	3	125.2
		GATA3				3.22	19	3.0 [0.1-5.8]
	NR	HNF4G				2.59	7	8.4
	P63	TP63	S			2.98	8	68.3 [45.8-90.8]
	Homeobox	POU3F1				4.69	10	15.7
		POU5F1				4.19	7	204.7 [183.0-226.4]
		POU5F1				4.19	7	204.7 [183.0-226.4]
		CDX1				2.70	5	84.4 [65.7-103.0]
		HOXB9				2.48	10	10.8
		ZEB1				2.26	41	6.5
	SOX	SOX6				2.38	5	3.6
		SPIB				3.10	13	n.d.
Cluster 4	E2F	E2F6		HDAC		1.98	27	7.3
	EGR	EGR1		p300		2.46	50	11.2 [4.8-130.3]
		EGR4				1.87	2	0.0
	EPAS2	EPAS1		p300, CBP		5.26	43	28.5 [0.1-163.0]
	IRF	IRF3		p300, CBP	Poly -pro	2.22	56	32.6 [10.9-44.3]
	KLF	KLF15		p300		2.01	25	5.5 [5.2-5.9]

	MEF	MEF2A		p300, HDAC		3.77	57	28.3 [27.7-28.9]
	PUR	PURA				2.31	46	7.1 [2.1-13.1]
	SP1	SP1		p300, DMT1, HDAC, CBP		2.32	57	25.9 [20.0-38.7]
		SP3		p300, DMT1, HDAC		2.44	53	12.7 [3.6-22.0]
	Zinc finger, C2H2	WT1		DMT1	Poly -pro	2.41	2	0.5
		MAZ			Poly -pro	2.53	57	88.5 [24.2-436.0]
		RREB1		HDAC	Poly -pro	3.03	29	2.4
		ZNF263				2.48	53	10.4 [3.5-20.1]
	ZNF	ZBTB7B		p300, HDAC	Poly -pro	2.41	51	15.0 [4.3-25.0]
		ZNF281			Poly -pro	2.25	43	11.0 [2.7-45.1]
Cluster 5	FOX	FOXA1				2.32	12	14.3
		FOXD2	S			3.67	0	0.1 [0.0-0.2]
		FOXL2				2.39	1	2.1 [0.0-71.0]
		FOXM1				1.91	1	9.1
		FOXN3				1.77	50	19.6 [3.0-36.3]
		FOXO3				2.05	54	14.0 [4.9-42.0]
Cluster 6	ARID	ARID3A			Poly -ala	2.62	27	7.6 [1.4-284.7]
	BPTF	BPTF				1.85	57	14.7 [6.7-24.7]
	FOX	FOXA2		SATB2		2.07	12	0.0 [0.0-0.0]
		FOXC1		SATB2	Poly -ala	3.00	14	1.3 [0.0-31.1]
		FOXJ3				2.98	57	16.6 [9.1-27.0]
		FOXO1		SATB2	Poly -ala	2.17	42	7.2 [0.8-61.8]
		FOXP1		SATB2		6.01	35	6.7 [0.0-15.7]
	FUBP	FUBP1				2.11	57	33.8 [8.9-100.6]
	IRF	IRF1				2.60	22	3.3 [0.3-27.4]
		IRF5				2.08	10	0.4 [0.0-26.7]
	MEF	MEF2C				2.01	21	61.1
	SRY	SRY				2.21	1	0.5 [0.0-1.1]
	ZNF	ZNF148				2.43	43	2.3
		ZNF202				0.44	33	4.0 [3.7-12.9]
		ZNF35				1.64	25	4.8 [1.4-8.7]

Table S2. Enrichment of functional properties in TF motif clusters

	Term	1	2	3	4	5	6
BIOCARTA	h_tertPathway:Overview of telomerase protein component gene hTert Transcriptional Regulation	0.0	0.0	2.4	0.0	0.0	0.0
BIOGRID_INTERACTION	2033:EP300~E1A binding protein p300	0.0	3.1	5.6	0.8	7.2	0.0
	1387:CREBBP~CREB binding protein	0.0	0.0	0.1	0.0	7.1	0.0
	2353:FOS~Fos proto-oncogene, AP-1 transcription factor subunit	0.0	0.0	0.0	0.0	4.8	0.0
	3239:HOXD13~homeobox D13	0.0	3.3	0.0	0.0	0.0	0.0
	23314:SATB2~SATB homeobox 2	0.0	2.9	0.0	0.0	0.0	0.0
	6667:SP1~Sp1 transcription factor	1.8	0.0	2.8	0.0	0.0	0.0
	468:ATF4~activating transcription factor 4	0.0	0.0	0.0	0.0	1.8	0.0
	1386:ATF2~activating transcription factor 2	0.0	0.0	0.0	0.0	1.7	0.0
	6929:TCF3~transcription factor 3	0.0	0.0	0.0	0.0	1.7	0.0
	3725:JUN~Jun proto-oncogene, AP-1 transcription factor subunit	0.0	0.0	1.3	0.0	1.7	0.0
	57154:SMURF1~SMAD specific E3 ubiquitin protein ligase 1	0.0	0.0	0.0	0.0	1.6	0.0
	4088:SMAD3~SMAD family member 3	0.0	0.0	0.0	0.0	1.6	0.2
	10413:YAP1~Yes associated protein 1	0.0	0.0	0.0	0.0	1.5	0.0
	10014:HDAC5~histone deacetylase 5	0.0	0.0	0.0	0.0	1.5	0.0
	1786:DNMT1~DNA methyltransferase 1	0.0	0.0	1.5	0.0	0.0	0.0
	7392:USF2~upstream transcription factor 2, c-fos interacting	0.0	0.0	0.0	0.0	1.5	0.0
	11016:ATF7~activating transcription factor 7	0.0	0.0	0.0	0.0	1.5	0.0
	6605:SMARCE1~SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	0.0	0.0	0.0	0.0	1.5	0.0
	26145:IRF2BP1~interferon regulatory factor 2 binding protein 1	0.0	0.0	0.0	0.0	1.4	0.0
	10538:BATF~basic leucine zipper ATF-like transcription factor	0.0	0.0	0.0	0.0	1.4	0.0
	25937:WWTR1~WW domain containing transcription regulator 1	0.0	0.0	0.0	0.0	1.4	0.0
	6304:SATB1~SATB homeobox 1	0.0	1.3	0.0	0.0	0.0	0.0
COG	Transcription / Cell division and chromosome partitioning	0.0	0.0	3.2	0.0	0.0	0.0
	Transcription	0.0	2.5	0.0	0.0	0.0	0.0
GOTERM_BP_DIRECT	GO:0006357~regulation of transcription from RNA polymerase II promoter	0.1	1.8	0.0	5.5	19.7	0.1
	GO:0045944~positive regulation of transcription from RNA polymerase II promoter	5.1	4.1	2.4	14.9	16.4	1.4
	GO:0006366~transcription from RNA polymerase II promoter	3.8	3.6	4.7	10.7	15.6	1.6

	GO:0045893~positive regulation of transcription, DNA-templated	2.5	3.8	2.6	5.1	9.8	1.9
	GO:0000122~negative regulation of transcription from RNA polymerase II promoter	1.0	2.0	2.9	4.1	6.8	1.7
	GO:0030154~cell differentiation	0.0	0.0	0.0	4.2	0.0	0.1
	GO:0006355~regulation of transcription, DNA-templated	0.0	0.3	4.0	0.0	0.4	0.0
	GO:0006351~transcription, DNA-templated	1.3	3.5	1.7	2.7	3.0	0.1
	GO:0045892~negative regulation of transcription, DNA-templated	0.6	0.6	1.7	3.0	0.1	0.8
	GO:0035019~somatic stem cell population maintenance	0.0	0.0	0.0	2.6	0.0	0.0
	GO:0001824~blastocyst development	0.0	0.0	0.0	2.2	0.4	0.0
	GO:0009887~organ morphogenesis	0.0	0.0	0.0	2.2	0.0	0.0
	GO:1901653~cellular response to peptide	0.0	0.0	0.0	0.0	2.1	0.0
	GO:0060337~type I interferon signaling pathway	0.0	0.2	0.3	0.0	2.1	0.0
	GO:0042832~defense response to protozoan	0.0	0.0	0.0	0.0	1.7	0.0
	GO:0030218~erythrocyte differentiation	0.0	0.0	0.0	1.5	0.0	0.0
	GO:0035329~hippo signaling	0.0	0.0	0.0	0.0	1.5	0.0
	GO:0045595~regulation of cell differentiation	0.0	0.0	0.0	0.0	1.4	0.0
GOTERM_CC_DIRECT	GO:0005634~nucleus	3.6	4.8	6.6	8.2	6.1	1.7
	GO:0005654~nucleoplasm	1.4	5.3	4.0	2.7	6.7	0.0
	GO:0005667~transcription factor complex	0.0	0.0	1.1	6.4	3.9	0.0
	GO:0000790~nuclear chromatin	0.0	0.0	0.4	2.3	5.1	0.0
	GO:0090575~RNA polymerase II transcription factor complex	0.0	0.0	0.0	0.0	2.1	0.0
	GO:0016607~nuclear speck	0.0	0.3	1.3	0.0	0.0	0.0
GOTERM_MF_DIRECT	GO:0003700~transcription factor activity, sequence-specific DNA binding	4.3	7.9	9.0	11.1	23.8	5.2
	GO:0000978~RNA polymerase II core promoter proximal region sequence-specific DNA binding	3.7	2.8	2.7	4.4	17.2	0.0
	GO:0001077~transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	4.3	0.4	4.6	5.3	16.5	0.6
	GO:0043565~sequence-specific DNA binding	5.6	7.4	4.5	14.5	7.9	6.3
	GO:0000981~RNA polymerase II transcription factor activity, sequence-specific DNA binding	0.0	7.4	1.5	9.5	1.9	4.2
	GO:0000982~transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding	0.0	0.0	0.0	0.0	8.1	0.0
	GO:0003677~DNA binding	2.3	2.8	2.7	2.5	5.4	1.4
	GO:0070888~E-box binding	0.0	0.0	0.0	2.4	5.4	0.0
	GO:0044212~transcription regulatory region DNA binding	4.4	1.3	2.3	2.1	4.8	0.0
	GO:0035035~histone acetyltransferase binding	0.0	0.0	4.6	0.0	0.0	0.0
	GO:0046872~metal ion binding	0.0	0.3	4.5	0.0	0.1	0.0
	GO:0043425~bHLH transcription factor binding	0.0	0.0	0.0	0.0	4.2	0.0
	GO:0008134~transcription factor binding	1.4	2.0	1.1	1.8	4.2	0.0

	GO:0001078~transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding	0.8	0.6	1.8	4.0	1.3	0.0
	GO:0001085~RNA polymerase II transcription factor binding	0.0	0.0	0.0	3.7	0.7	0.0
	GO:0005515~protein binding	0.3	0.5	1.7	0.4	3.6	0.0
	GO:0046982~protein heterodimerization activity	0.4	0.0	0.2	0.0	3.1	0.0
	GO:0000975~regulatory region DNA binding	0.0	1.3	0.0	0.0	3.0	0.0
	GO:0001228~transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding	0.8	0.6	0.0	2.8	0.5	0.0
	GO:0001158~enhancer sequence-specific DNA binding	0.0	0.0	0.0	2.8	0.0	0.0
	GO:0000979~RNA polymerase II core promoter sequence-specific DNA binding	2.6	0.0	2.2	2.1	1.8	0.0
	GO:0003705~transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding	2.5	2.0	0.8	0.0	1.7	0.0
	GO:0046983~protein dimerization activity	0.0	0.0	0.0	2.4	2.0	0.0
	GO:0003713~transcription coactivator activity	1.5	0.0	0.0	1.0	2.4	0.0
	GO:0000977~RNA polymerase II regulatory region sequence-specific DNA binding	0.0	1.3	2.3	2.0	0.9	0.0
	GO:0003676~nucleic acid binding	0.2	0.1	2.2	0.0	0.1	0.0
	GO:0000976~transcription regulatory region sequence-specific DNA binding	0.0	0.0	0.0	2.1	0.0	0.0
	GO:0000980~RNA polymerase II distal enhancer sequence-specific DNA binding	0.0	0.0	0.0	2.0	0.0	0.0
	GO:0044729~hemi-methylated DNA-binding	0.0	0.0	1.9	0.0	0.0	0.0
	GO:0010385~double-stranded methylated DNA binding	0.0	0.0	1.8	0.0	0.0	0.0
	GO:0003682~chromatin binding	0.4	1.7	0.3	1.5	1.8	0.0
	GO:0042826~histone deacetylase binding	0.0	0.0	0.7	0.0	1.4	0.0
INTACT	11016:activating transcription factor 7(ATF7)	0.0	0.0	0.0	0.0	4.0	0.0
	3726:JunB proto-oncogene, AP-1 transcription factor subunit(JUNB)	0.0	0.0	0.0	0.0	3.9	0.0
	1649:DNA damage inducible transcript 3(DDIT3)	0.0	0.0	0.0	0.0	3.8	0.0
	468:activating transcription factor 4(ATF4)	0.0	0.0	0.0	0.0	3.4	0.0
	23314:SATB homeobox 2(SATB2)	0.0	3.3	0.0	0.0	0.0	0.0
	3727:JunD proto-oncogene, AP-1 transcription factor subunit(JUND)	0.0	0.0	0.3	0.0	3.0	0.0
	9935:MAF bZIP transcription factor B(MAFB)	0.0	0.0	0.0	0.0	1.8	0.0
	3229:homeobox C13(HOXC13)	0.0	1.7	0.0	0.0	0.0	0.0
	4779:nuclear factor, erythroid 2 like 1(NFE2L1)	0.0	0.0	0.0	0.0	1.6	0.0
	6304:SATB homeobox 1(SATB1)	0.0	1.6	0.0	0.0	0.0	0.0
	6886:TAL bHLH transcription factor 1, erythroid differentiation factor(TAL1)	0.0	0.0	0.0	1.6	0.2	0.0
	1386:activating transcription factor 2(ATF2)	0.0	0.0	0.0	0.0	1.5	0.0
INTEPD	IPR000418:Ets domain	0.0	0.0	0.0	16.2	0.0	0.0

	IPR001766:Transcription factor, fork head	0.0	6.1	0.0	0.0	0.0	12.3
	IPR018122:Transcription factor, fork head, conserved site	0.0	2.5	0.0	0.0	0.0	9.9
	IPR004827:Basic-leucine zipper domain	0.0	0.0	0.0	0.0	9.8	0.0
	IPR011991:Winged helix-turn-helix DNA-binding domain	0.0	6.4	0.1	8.7	2.4	9.3
	IPR015880:Zinc finger, C2H2-like	0.7	0.3	9.2	0.1	0.9	0.0
	IPR013087:Zinc finger C2H2-type/integrase DNA-binding domain	0.7	0.3	9.2	0.1	1.0	0.0
	IPR007087:Zinc finger, C2H2	0.8	0.3	9.2	0.1	0.9	0.0
	IPR000837:Fos transforming protein	0.0	0.0	0.0	0.0	7.6	0.0
	IPR011598:Myc-type, basic helix-loop-helix (bHLH) domain	0.0	0.0	0.0	1.5	6.2	0.0
	IPR003118:Pointed domain	0.0	0.0	0.0	5.6	0.0	0.0
	IPR013761:Sterile alpha motif/pointed domain	0.0	0.0	0.0	4.1	0.0	0.0
	IPR016361:Transcriptional enhancer factor	0.0	0.0	0.0	0.0	3.6	0.0
	IPR000818:TEA/ATTS	0.0	0.0	0.0	0.0	3.6	0.0
	IPR019471:Interferon regulatory factor-3	0.0	0.0	0.0	0.0	3.2	0.0
	IPR001346:Interferon regulatory factor DNA-binding domain	0.0	1.3	0.0	0.0	3.0	0.0
	IPR019817:Interferon regulatory factor, conserved site	0.0	1.3	0.0	0.0	3.0	0.0
	IPR001356:Homeodomain	0.0	0.0	0.0	2.9	0.0	0.0
	IPR008917:Eukaryotic transcription factor, Skn-1-like, DNA-binding	0.0	0.0	0.0	0.0	2.6	0.0
	IPR017855:SMAD domain-like	0.0	0.0	0.0	0.0	2.6	0.0
	IPR009057:Homeodomain-like	0.0	0.0	0.0	2.5	0.1	0.0
	IPR017970:Homeobox, conserved site	0.0	0.0	0.0	2.2	0.0	0.0
	IPR013088:Zinc finger, NHR/GATA-type	0.8	0.0	0.0	2.0	0.0	0.0
	IPR008967:p53-like transcription factor, DNA-binding	0.0	0.0	0.0	0.0	1.8	0.0
	IPR022084:Transcription factor Elf, N-terminal	0.0	0.0	0.0	1.8	0.0	0.0
	IPR008984:SMAD/FHA domain	0.0	0.0	0.0	0.0	1.7	0.0
	IPR013711:Runx, C-terminal domain	0.0	0.0	0.0	0.0	1.6	0.0
	IPR013524:Runt domain	0.0	0.0	0.0	0.0	1.6	0.0
	IPR000040:Acute myeloid leukemia 1 protein (AML1)/Runt	0.0	0.0	0.0	0.0	1.6	0.0
	IPR027384:Runx, central domain	0.0	0.0	0.0	0.0	1.6	0.0
	IPR021802:Basic helix-loop-helix leucine zipper transcrition factor MiT/TFE	0.0	0.0	0.0	0.0	1.5	0.0
	IPR012346:p53/RUNT-type transcription factor, DNA-binding domain	0.0	0.0	0.0	0.0	1.4	0.0
KEGG_PATHWAY	hsa04380:Osteoclast differentiation	0.0	0.0	0.0	0.0	4.6	0.0
PFAM	PF00178:Ets-domain	0.0	0.0	0.0	16.3	0.0	0.0
	PF00250:Fork head domain	0.0	6.1	0.0	0.0	0.0	12.4
	PF00170:bZIP transcription factor	0.0	0.0	0.0	0.0	7.6	0.0

	PF00010:Helix-loop-helix DNA-binding domain	0.0	0.0	0.0	1.6	6.4	0.0
	PF02198:Sterile alpha motif (SAM)/Pointed domain	0.0	0.0	0.0	5.7	0.0	0.0
	PF00096:Zinc finger, C2H2 type	0.0	0.0	4.6	0.0	0.8	0.0
	PF01285:TEA/ATTS domain family	0.0	0.0	0.0	0.0	3.8	0.0
	PF10401:Interferon-regulatory factor 3	0.0	0.0	0.0	0.0	3.4	0.0
	PF00605:Interferon regulatory factor transcription factor	0.0	1.2	0.0	0.0	3.2	0.0
	PF00046:Homeobox domain	0.0	0.0	0.0	3.1	0.0	0.0
	PF12310:Transcription factor protein N terminal	0.0	0.0	0.0	1.8	0.0	0.0
	PF00853:Runt domain	0.0	0.0	0.0	0.0	1.6	0.0
	PF08504:Runx inhibition domain	0.0	0.0	0.0	0.0	1.6	0.0
	PF11851:Domain of unknown function (DUF3371)	0.0	0.0	0.0	0.0	1.6	0.0
	PF00320:GATA zinc finger	0.0	0.0	0.0	1.4	0.0	0.0
	PF00157:Pou domain - N-terminal to homeobox domain	0.0	0.0	0.0	1.3	0.0	0.0
PIR_SUPERFAMILY	PIRSF002603:transcriptional enhancer factor	0.0	0.0	0.0	0.0	4.3	0.0
SMART	SM00413:ETS	0.0	0.0	0.0	14.7	0.0	0.0
	SM00339:FH	0.0	5.2	0.0	0.0	0.0	11.7
	SM00338:BRLZ	0.0	0.0	0.0	0.0	9.0	0.0
	SM00355:ZnF_C2H2	0.9	0.1	6.8	0.0	0.5	0.0
	SM00353:HLH	0.0	0.0	0.0	1.3	5.4	0.0
	SM00251:SAM_PNT	0.0	0.0	0.0	5.2	0.0	0.0
	SM00426:TEA	0.0	0.0	0.0	0.0	3.7	0.0
	SM01243:SM01243	0.0	0.0	0.0	0.0	3.3	0.0
	SM00348:IRF	0.0	1.1	0.0	0.0	3.1	0.0
	SM00389:HOX	0.0	0.0	0.0	2.3	0.0	0.0
	SM00401:ZnF_GATA	0.0	0.0	0.0	1.3	0.0	0.0
UP_KEYWORDS	DNA-binding	7.5	10.8	13.4	15.7	27.0	3.6
	Transcription regulation	7.3	11.7	12.9	16.1	21.7	3.6
	Transcription	7.3	11.8	12.9	16.1	21.5	3.7
	Nucleus	4.5	7.3	7.9	11.5	17.3	2.2
	Activator	4.5	2.4	7.3	8.6	16.3	1.3
	Zinc-finger	1.5	1.0	6.5	0.6	0.2	0.0
	Ubl conjugation	1.0	3.4	3.5	1.7	6.4	0.7
	Zinc	1.2	0.7	5.4	0.7	0.1	0.0
	Repressor	1.0	2.4	4.9	1.9	0.7	0.4
	Isopeptide bond	1.3	2.5	3.4	2.4	4.6	0.3
	Metal-binding	0.8	0.4	3.8	0.3	0.0	0.0
	Developmental protein	0.0	0.2	0.5	3.7	0.1	0.0
	Phosphoprotein	0.6	2.5	1.0	0.4	3.1	1.4
	Homeobox	0.0	0.0	0.0	3.1	0.0	0.0

	Differentiation	0.0	0.7	1.4	1.8	0.6	0.0
	Coiled coil	0.0	0.0	0.0	0.0	1.7	0.0
UP_SEQ_FEATURE	DNA-binding region:Basic motif	0.0	0.0	0.0	0.2	19.5	0.0
	DNA-binding region:ETS	0.0	0.0	0.0	16.3	0.0	0.0
	domain:Leucine-zipper	0.0	0.0	0.0	0.0	15.6	0.0
	DNA-binding region:Fork-head	0.0	5.5	0.0	0.0	0.0	11.9
	zinc finger region:C2H2-type 1	0.7	0.2	10.3	0.1	1.1	0.0
	zinc finger region:C2H2-type 3	0.8	0.3	10.1	0.0	1.0	0.0
	zinc finger region:C2H2-type 2	0.7	0.2	10.1	0.0	1.0	0.0
	domain:Helix-loop-helix motif	0.0	0.0	0.0	1.2	6.3	0.0
	domain:PNT	0.0	0.0	0.0	5.3	0.0	0.0
	cross-link:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO)	1.6	0.3	1.3	4.1	0.2	0.0
	DNA-binding region:TEA	0.0	0.0	0.0	0.0	3.6	0.0
	DNA-binding region:Tryptophan pentad repeat	0.0	0.8	0.0	0.0	2.9	0.0
	DNA-binding region:Homeobox	0.0	0.0	0.0	1.8	0.0	0.0

Table S3. Co-enrichment and -depletion of motifs from different clusters

Cluster Pair	Odds Ratio (Genomic)	Odds Ratio (Shuffled)	p-value (Genomic)	p-value (Shuffled)
1_2	0.36 [0.36-0.37]	0.54 [0.53-0.55]	0.0E+00	0.0E+00
1_3	0.76 [0.75-0.77]	0.77 [0.75-0.78]	0.0E+00	9.2E-193
1_4	1.29 [1.26-1.32]	1.74 [1.70-1.77]	4.9E-98	0.0E+00
1_5	1.51 [1.49-1.54]	1.43 [1.40-1.46]	0.0E+00	1.0E-217
1_6	0.81 [0.79-0.83]	0.69 [0.68-0.71]	1.0E-61	0.0E+00
2_3	1.04 [1.03-1.06]	1.08 [1.06-1.09]	8.9E-08	5.3E-26
2_4	0.92 [0.90-0.94]	1.06 [1.05-1.08]	4.5E-12	9.0E-16
2_5	0.56 [0.55-0.57]	0.71 [0.70-0.72]	0.0E+00	2.9E-261
2_6	1.33 [1.30-1.36]	1.08 [1.06-1.10]	2.6E-114	3.0E-26
3_4	1.28 [1.26-1.31]	1.16 [1.14-1.18]	1.9E-112	7.3E-75
3_5	0.64 [0.64-0.65]	0.83 [0.81-0.84]	0.0E+00	8.8E-73
3_6	1.02 [1.00-1.05]	1.12 [1.10-1.13]	4.3E-02	2.0E-45
4_5	0.93 [0.91-0.95]	0.81 [0.79-0.83]	2.4E-09	1.2E-86
4_6	0.44 [0.41-0.46]	0.47 [0.46-0.48]	1.9E-275	0.0E+00
5_6	1.84 [1.79-1.89]	1.72 [1.68-1.75]	0.0E+00	0.0E+00

REFERENCES

1. Felsenfeld G, Boyes J, Chung J, Clark D, & Studitsky V (1996) Chromatin structure and gene expression. *Proc Natl Acad Sci U S A* 93(18):9384-9388.

2. Gerstein MB, *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91-100.

3. Polach KJ & Widom J (1996) A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J Mol Biol* 258(5):800-812.

4. Scruggs BS, *et al.* (2015) Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol Cell* 58(6):1101-1112.

5. Gross DS & Garrard WT (1988) Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57:159-197.

6. Roadmap Epigenomics C, *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317-330.

7. Grossman SR, *et al.* (2017) Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci U S A* 114(7):E1291-E1300.

8. Kheradpour P, *et al.* (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* 23(5):800-811.

9. Kwasnieski JC, Fiore C, Chaudhari HG, & Cohen BA (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome research* 24(10):1595-1602.

10. White MA, Myers CA, Corbo JC, & Cohen BA (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci U S A* 110(29):11952-11957.

11. Core LJ, *et al.* (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics* 46(12):1311-1320.

12. Stampfel G, *et al.* (2015) Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* 528(7580):147-151.

13. Zabidi MA, *et al.* (2015) Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518(7540):556-559.

14. Fu Y, Sinha M, Peterson CL, & Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* 4(7):e1000138.

15. Hebbar PB & Archer TK (2007) Chromatin-dependent cooperativity between site-specific transcription factors in vivo. *J Biol Chem* 282(11):8284-8291.

16. Denny SK, *et al.* (2016) Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. *Cell* 166(2):328-342.

17. Plachetka A, *et al.* (2008) C/EBPbeta induces chromatin opening at a cell-type-specific enhancer. *Molecular and cellular biology* 28(6):2102-2112.

18. Grontved L, *et al.* (2013) C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *EMBO J* 32(11):1568-1583.

19. Soufi A, *et al.* (2015) Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* 161(3):555-568.

20. Naval-Sanchez M, Potier D, Hulselmans G, Christiaens V, & Aerts S (2015) Identification of Lineage-Specific Cis-Regulatory Modules Associated with Variation in Transcription Factor Binding and Chromatin Activity Using Ornstein-Uhlenbeck Models. *Mol Biol Evol* 32(9):2441-2455.

21. Sherwood RI, *et al.* (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature biotechnology* 32(2):171-178.

22. Cao Z, Umek RM, & McKnight SL (1991) Regulated expression of three C/EBP isoforms during adipose conversion of 3T3-L1 cells. *Genes & development* 5(9):1538-1552.

23. Nakahashi H, *et al.* (2013) A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* 3(5):1678-1689.

24. Nevil M, Bondra ER, Schulz KN, Kaplan T, & Harrison MM (2017) Stable Binding of the Conserved Transcription Factor Grainy Head to its Target Genes Throughout Drosophila melanogaster Development. *Genetics* 205(2):605-620.

25. Sung MH, Guertin MJ, Baek S, & Hager GL (2014) DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell* 56(2):275-285.

26. Mazza D, Abernathy A, Golob N, Morisaki T, & McNally JG (2012) A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Res* 40(15):e119.

27. Sharp ZD, *et al.* (2006) Estrogen-receptor-alpha exchange and chromatin dynamics are ligand- and domain-dependent. *Journal of cell science* 119(Pt 19):4101-4116.

28. Siersbaek R, *et al.* (2011) Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *EMBO J* 30(8):1459-1472.

29. Voss TC, *et al.* (2011) Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell* 146(4):544-554.

30. Chavez S & Beato M (1997) Nucleosome-mediated synergism between transcription factors on the mouse mammary tumor virus promoter. *Proc Natl Acad Sci U S A* 94(7):2885-2890.

31. Hebbar PB & Archer TK (2003) Nuclear factor 1 is required for both hormone-dependent chromatin remodeling and transcriptional activation of the mouse mammary tumor virus promoter. *Molecular and cellular biology* 23(3):887-898.

32. Lefterova MI, *et al.* (2008) PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. *Genes & development* 22(21):2941-2952.

33. Chatr-Aryamontri A, *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 45(D1):D369-D379.

34. Orchard S, *et al.* (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42(Database issue):D358-363.

35. Thanos D & Maniatis T (1995) Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* 83(7):1091-1100.

36. Kim TK & Maniatis T (1997) The mechanism of transcriptional synergy of an in vitro assembled interferon-beta enhanceosome. *Mol Cell* 1(1):119-129.

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

38. Kanehisa M, Furumichi M, Tanabe M, Sato Y, & Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45(D1):D353-D361.

39. Charron F & Nemer M (1999) GATA transcription factors and cardiac development. *Semin Cell Dev Biol* 10(1):85-91.

40. Mallo M, Wellik DM, & Deschamps J (2010) Hox genes and regional patterning of the vertebrate body plan. *Dev Biol* 344(1):7-15.

41. Maroulakou IG & Bowe DB (2000) Expression and function of Ets transcription factors in mammalian development: a regulatory network. *Oncogene* 19(55):6432-6442.

42. Rosenfeld MG (1991) POU-domain transcription factors: pou-er-ful developmental regulators. *Genes & development* 5(6):897-907.

43. Sarkar A & Hochedlinger K (2013) The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell Stem Cell* 12(1):15-30.

44. Ting CN, Olson MC, Barton KP, & Leiden JM (1996) Transcription factor GATA-3 is required for development of the T-cell lineage. *Nature* 384(6608):474-478.

45. Karlseder J, Rotheneder H, & Wintersberger E (1996) Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F. *Molecular and cellular biology* 16(4):1659-1667.

46. Khachigian LM, Williams AJ, & Collins T (1995) Interplay of Sp1 and Egr-1 in the proximal platelet-derived growth factor A-chain promoter in cultured vascular endothelial cells. *J Biol Chem* 270(46):27679-27686.

47. Koizume S, *et al.* (2012) HIF2alpha-Sp1 interaction mediates a deacetylation-dependent FVII-gene activation under hypoxic conditions in ovarian cancer cells. *Nucleic Acids Res* 40(12):5389-5401.

48. Krainc D, *et al.* (1998) Synergistic activation of the N-methyl-D-aspartate receptor subunit 1 promoter by myocyte enhancer factor 2C and Sp1. *J Biol Chem* 273(40):26218-26224.

49. Kypriotou M, *et al.* (2007) Human collagen Krox up-regulates type I collagen expression in normal and scleroderma fibroblasts through interaction with Sp1 and Sp3 transcription factors. *J Biol Chem* 282(44):32000-32014.

50. Li J, *et al.* (2010) Sp1 and KLF15 regulate basal transcription of the human LRP5 gene. *BMC Genet* 11:12.

51. Minc E, *et al.* (1999) The human copper-zinc superoxide dismutase gene (SOD1) proximal promoter is regulated by Sp1, Egr-1, and WT1 via non-canonical binding sites. *J Biol Chem* 274(1):503-509.

52. Nenoi M, Ichimura S, Mita K, Yukawa O, & Cartwright IL (2001) Regulation of the catalase gene promoter by Sp1, CCAAT-recognizing factors, and a WT1/Egr-related factor in hydrogen peroxide-resistant HP100 cells. *Cancer Res* 61(15):5885-5894.

53. Parks CL & Shenk T (1997) Activation of the adenovirus major late promoter by transcription factors MAZ and Sp1. *J Virol* 71(12):9600-9607.

54. Tretiakova A, Steplewski A, Johnson EM, Khalili K, & Amini S (1999) Regulation of myelin basic protein gene transcription by Sp1 and Puralpha: evidence for association of Sp1 and Puralpha in brain. *J Cell Physiol* 181(1):160-168.

55. Yamamoto J, *et al.* (2004) A Kruppel-like factor KLF15 contributes fasting-induced transcriptional activation of mitochondrial acetyl-CoA synthetase gene AceCS2. *J Biol Chem* 279(17):16954-16962.

56. Courey AJ, Holtzman DA, Jackson SP, & Tjian R (1989) Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. *Cell* 59(5):827-836.

57. Seipel K, Georgiev O, & Schaffner W (1992) Different activation domains stimulate transcription from remote ('enhancer') and proximal ('promoter') positions. *EMBO J* 11(13):4961-4968.

58. Valen E & Sandelin A (2011) Genomic and chromatin signals underlying transcription start-site selection. *Trends Genet* 27(11):475-485.

59. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.

60. Clark KL, Halay ED, Lai E, & Burley SK (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* 364(6436):412-420.

61. Ramakrishnan V, Finch JT, Graziano V, Lee PL, & Sweet RM (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* 362(6417):219-223.

62. Cirillo LA, *et al.* (1998) Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J* 17(1):244-254.

63. Chaya D, Hayamizu T, Bustin M, & Zaret KS (2001) Transcription factor FoxA (HNF3) on a nucleosome at an enhancer complex in liver chromatin. *J Biol Chem* 276(48):44385-44389.

64. Cirillo LA, *et al.* (2002) Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* 9(2):279-289.

65. Iwafuchi-Doi M, *et al.* (2016) The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol Cell* 62(1):79-91.

66. Taube JH, Allton K, Duncan SA, Shen L, & Barton MC (2010) Foxa1 functions as a pioneer transcription factor at transposable elements to activate Afp during differentiation of embryonic stem cells. *J Biol Chem* 285(21):16135-16144.

67. Lalmansingh AS, Karmakar S, Jin Y, & Nagaich AK (2012) Multiple modes of chromatin remodeling by Forkhead box proteins. *Biochim Biophys Acta* 1819(7):707-715.

68. He HH, *et al.* (2010) Nucleosome dynamics define transcriptional enhancers. *Nature genetics* 42(4):343-347.

69. Tsukiyama T & Wu C (1995) Purification and properties of an ATP-dependent nucleosome remodeling factor. *Cell* 83(6):1011-1020.

70. Barak O, *et al.* (2003) Isolation of human NURF: a regulator of Engrailed gene expression. *EMBO J* 22(22):6089-6100.

71. Wysocka J, *et al.* (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* 442(7098):86-90.

72. Lin D, *et al.* (2007) Bright/ARID3A contributes to chromatin accessibility of the immunoglobulin heavy chain enhancer. *Mol Cancer* 6:23.

73. Kaplan MH, Zong RT, Herrscher RF, Scheuermann RH, & Tucker PW (2001) Transcriptional activation by a matrix associating region-binding protein. contextual requirements for the function of bright. *J Biol Chem* 276(24):21325-21330.

74. Webb CF, Das C, Eneff KL, & Tucker PW (1991) Identification of a matrix-associated region 5' of an immunoglobulin heavy chain variable region gene. *Molecular and cellular biology* 11(10):5206-5211.

75. Riedel CG, *et al.* (2013) DAF-16 employs the chromatin remodeller SWI/SNF to promote stress resistance and longevity. *Nat Cell Biol* 15(5):491-501.

76. Wilsker D, Patsialou A, Dallas PB, & Moran E (2002) ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development. *Cell Growth Differ* 13(3):95-106.

77. Cai S, Han HJ, & Kohwi-Shigematsu T (2003) Tissue-specific nuclear architecture and gene expression regulated by SATB1. *Nat Genet* 34(1):42-51.

78. Yasui D, Miyano M, Cai S, Varga-Weisz P, & Kohwi-Shigematsu T (2002) SATB1 targets chromatin remodelling to regulate genes over long distances. *Nature* 419(6907):641-645.

79. Herrscher RF, *et al.* (1995) The immunoglobulin heavy-chain matrix-associating regions are bound by Bright: a B cell-specific trans-activator that describes a new DNA-binding protein family. *Genes & development* 9(24):3067-3082.

80. Natesan S & Gilman MZ (1993) DNA bending and orientation-dependent function of YY1 in the c-fos promoter. *Genes & development* 7(12B):2497-2509.

81. Werner MH, Huth JR, Gronenborn AM, & Clore GM (1995) Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell* 81(5):705-714.

82. Giese K, Cox J, & Grosschedl R (1992) The HMG domain of lymphoid enhancer factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. *Cell* 69(1):185-195.

83. Giese K, Kingsley C, Kirshner JR, & Grosschedl R (1995) Assembly and function of a TCR alpha enhancer complex is dependent on LEF-1-induced DNA bending and multiple protein-protein interactions. *Genes & development* 9(8):995-1008.

84. Duncan R, *et al.* (1994) A sequence-specific, single-strand binding protein activates the far upstream element of c-myc and defines a new DNA-binding motif. *Genes & development* 8(4):465-480.

85. Bode J, *et al.* (1992) Biological significance of unwinding capability of nuclear matrix-associating DNAs. *Science* 255(5041):195-197.

86. Arnosti DN & Kulkarni MM (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* 94(5):890-898.

87. Panne D, Maniatis T, & Harrison SC (2007) An atomic model of the interferon-beta enhanceosome. *Cell* 129(6):1111-1123.

88. Arnosti DN, Barolo S, Levine M, & Small S (1996) The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* 122(1):205-214.

89. Liu F & Posakony JW (2012) Role of architecture in the function and specificity of two Notch-regulated transcriptional enhancer modules. *PLoS Genet* 8(7):e1002796.

90. Rastegar S, *et al.* (2008) The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol* 318(2):366-377.

91. Corces MR, *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics* 48(10):1193-1203.

92. Buenrostro JD, *et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523(7561):486-490.

93. Buenrostro JD, Wu B, Chang HY, & Greenleaf WJ (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* 109:21 29 21-29.

94. Zhang Y, *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9(9):R137.

95. Matys V, *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34(Database issue):D108-110.

96. Mathelier A, *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42(Database issue):D142-147.

97. Weirauch MT, *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158(6):1431-1443.

98. McLeay RC & Bailey TL (2010) Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* 11:165.

99. Grant CE, Bailey TL, & Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017-1018.

Chapter 4

Regiospecific binding of TF classes within enhancers occurs *in vivo* and facilitates TF regulatory activity

This chapter presents an early draft of a paper with contributions from Jesse Engreitz, Michael Kane, and Eric Lander. We discuss preliminary results as more data is currently being generated for the full study.

ABSTRACT

Transcription factors (TFs) control gene expression by binding to target sites in regulatory sequences in nucleosome-depleted regions of DNA, but relatively little is known about the functional role of binding site organization. We recently found that inferred binding sites for 103 human TFs across 47 cell types display "regiospecificity"—that is, different positional biases within nucleosome depleted regions (1). These patterns separate TFs into distinct classes that share functional properties and appear to be optimally positioned to facilitate their known activities—for example, binding sites for pioneer factors are localized to the logical position for nucleosome displacement. Here, we sought to experimentally test the functional significance of the position of TF binding sites in enhancers. Using massively parallel reporter assays (MPRA), we functionally characterized the activity of TF motifs in native and perturbed regulatory NDRs from five human cell lines. We found that different classes of TF binding sites are associated with different effects on expression output. Moreover, we found that

TF binding sites in "optimal" positions are associated with greater transcriptional activity and likely to be bound by TFs *in vivo* (using ChIP-seq data for 61 TFs from the ENCODE project). Overall, these results confirm the biological importance of the regiospecific binding patterns and provide support for distinct functional roles of the TF classes.

INTRODUCTION

Transcription factors (TFs) control when, where, and to what level each gene is transcribed, by binding to target sites in regulatory sequences such as promoters and enhancers. Enhancer- and promoter-bound TFs together mediate transcriptional control of target genes through a variety of mechanisms, including recruitment of cofactors and the general transcriptional machinery and remodeling of the local chromatin environment. Since the discovery of enhancers 40 years ago, deciphering the "regulatory code"—the rules for how TF synergy is encoded in enhancer sequences and how TFs achieve combinatorial enhancer control—has been an area of intense interest. Such a code could shed light on core biological processes such as development, and could enable prediction of the effect on expression of an arbitrary regulatory sequence and the engineering synthetic enhancers for use in research or medicine.

Combinatorial TF regulation is encoded in regulatory sequences in the form of arrays of specific recognition sites for distinct combinations of transcription factors (TFs). Very few TFs can activate transcription on their own; they instead typically function cooperatively with partner TFs bound at the same enhancers. This combinatorial property of enhancers allows a relatively small set of TFs to implement varied expression outputs for 20,000 genes across thousands of cell types and conditions by acting in different groupings (2-4). Furthermore, it enables enhancers to be specifically activated by the convergence of multiple signaling pathways, such as in developmental and immune enhancers (5-9).

Functional studies of various regulatory elements have revealed that both binding-site composition (identity and number of different TF binding sites) and organization (position, order and orientation of TF binding sites) can affect expression

output (10). However, the general rules and underlying mechanisms that determine the transcriptional output of a particular regulatory sequence remain poorly understood. Not every potential TF binding site in an enhancer is bound *in vivo* (11, 12), and for all but a few enhancers the functional TF binding sites have not been experimentally defined. Furthermore, the relationship between the input (composition and arrangement of TF binding sites) and the output (gene expression) has proven to be complex; many different relationships (or "computations") exist (2-4). Finally, the functions of most TFs in transcriptional control are largely unknown (13), making it hard to generalize across TFs or infer universal principles of regulatory grammar. Hence, a universal code at the level of TF binding sites has remained elusive.

In recent years, new high-throughput and quantitative technologies that enable us to synthesize and simultaneously measure the activities of thousands of regulatory sequences have made it possible to evaluate regulatory rules more systematically. Studies from our lab (see Chapter 2) and others have used pools of native and synthetic regulatory sequences to test various features of enhancer architecture and build models to predict expression from the motif-site composition (i.e. number and TF identity) of regulatory sequences. However, these models only partially explained the variation in expression output, suggesting additional factors, such as the organization of the sites, also play an important role. Consistent with this notion, permuting the order and orientation of sites in synthetic enhancers in liver cells resulted in significant variance in transcriptional output (14). However, the number of possible arrangements of TF binding sites is enormous. Consequently, it has been difficult to test how the organization of sites in regulatory sequences affects their activity.

In our previous work, we developed a framework for studying active TF binding sites in a network of enhancers in a well-characterized model system (PPARγ response elements [PPREs] in mouse adipocytes). Our method used iterative rounds of massively parallel reporter assays (15, 16) to identify and characterize functional motif sites in native, perturbed and synthetic enhancer sequences (17). We identified 20-30 active TFs (i.e. TFs whose binding sites were significantly correlated with enhancer activity). Notably, enrichment of the TF binding sites within the genomic enhancers accurately predicted TF activity in the functional assays. Examining the synergy between different TF binding sites, we identified "equivalence groups" of TFs that interacted with TF in other groups in the same way. The existence of these groups supports a model that separates TFs into functional classes that play different roles in transcriptional activation.

Instead of exhaustively testing permutations of TF binding sites in regulatory sequences, we sought to study the positions of naturally occurring TF binding sites within the NDR. To infer functional TF binding sites, we drew on an earlier observation that TF binding sites most strongly associated with enhancer activity correspond to the most highly enriched TF motifs in genomic regulatory elements (17). We showed that inferred binding sites for 103 human TFs across 47 cell types display distinct positional biases within NDRs (1), separating TFs into six positional classes: binding sites for some classes occur near the center, while others occur towards the edges or at specific locations within NDRs (Fig. 1A,B). Strikingly, the positional classes comprise TFs with shared functional properties, such as binding stability, interactions with other TFs and cofactors, cell-type specificity, and pioneering ability. Furthermore, the positional bias of

the TFs appears to be connected to their known functions—for example, binding sites for TFs that recruit chromatin modifiers occur towards the edges of the NDR, and binding sites for pioneer factors are localized to the optimal position for nucleosome displacement. These results suggest that distinct regiospecific binding patterns within enhancers reflect functional classes of TFs positioned to optimize their role in enhancer activation.

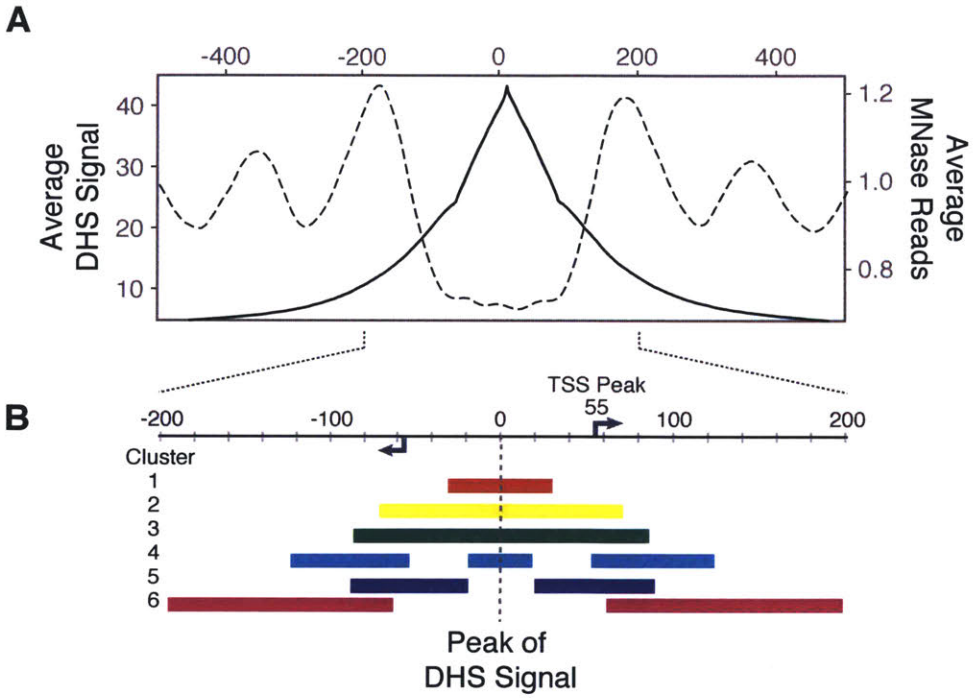


Figure 5. TF motif position patterns fall into 6 distinct clusters. (A) The nucleosome-depleted region at putative regulatory elements tends to span ~200 bp centered around the peak of the DHS signal and is generally flanked by well-positioned nucleosomes centered at around +200 bp and -200 bp. Composite plot of DHS signal (solid line) and MNase-seq reads (dashed line) in 1 kb region aligned around peak of DHS signal. **(B)** Schematic of NDR structure and motif positions. The arrows indicate the peak of transcriptional initiation estimated from CAGE data. The colored bars represent regions for each cluster with motif densities above the mean. Tick marks occur at 20 bp intervals.

We sought to extend our observation that TF motifs show regiospecific distribution by asking whether TF motifs in preferred positions show greater TF binding and enhancer activity *in vivo*. We utilized the MPRA-based approach we developed previously to functionally characterized the activity of TF motifs in native and perturbed regulatory NDRs from six human cell lines. We assessed the generality of the association between motif enrichment and activity that we observed in PPREs, and classified the positional preferences of TF motifs that were either (1) correlated with enhancer activity in an MPRA experiment or (2) enriched in regulatory NDR in at least one cell line. Using the MPRA-based approach, we functionally characterized the activity of TF motifs in native and perturbed regulatory NDRs from six human cell lines and examined the correlation between activity and enrichment. We found that positional classes show different activities in the plasmid-based reporter assay, consistent with functional differences. Finally, we tested the hypothesis that the positions of TF binding sites in NDR contributes to their activity. Strikingly, we found that binding sites in "optimal" positions were more likely to be bound by TFs *in vivo* (using ChIP-seq data for 61 TFs from the ENCODE project) and associated with greater regulatory activity in reporter assays. Thus, the regiospecific binding of TF classes within NDRs occurs *in vivo* and contributes to promoter activation.

RESULTS

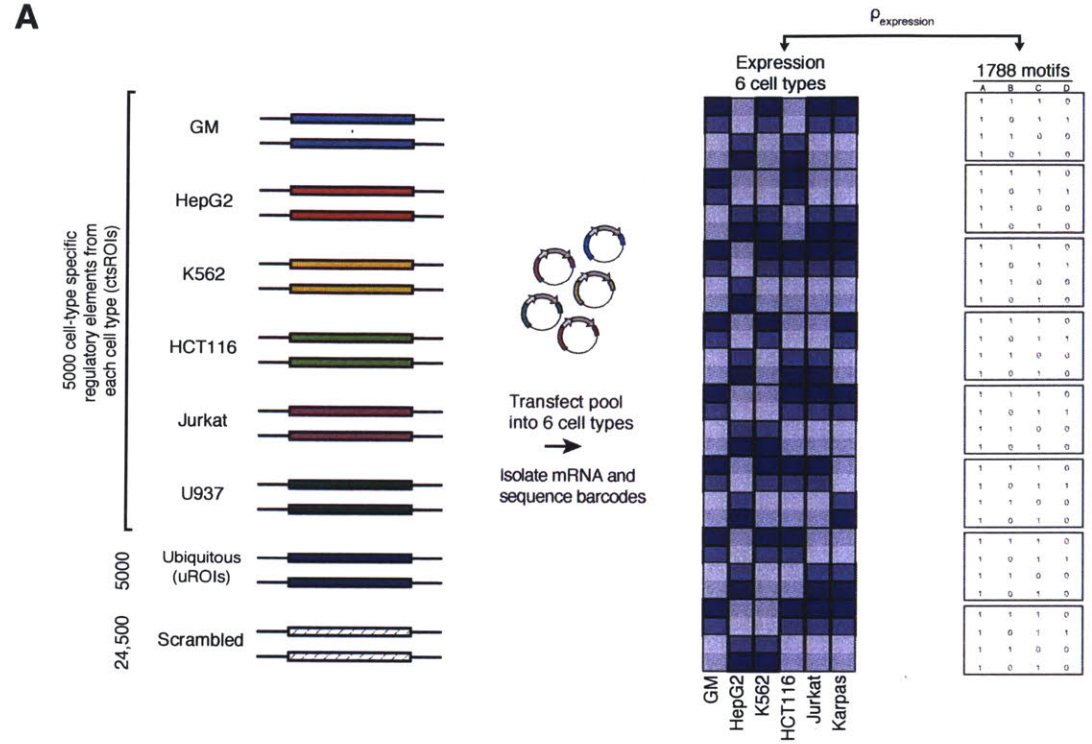
Enhancer selection and MPRA design

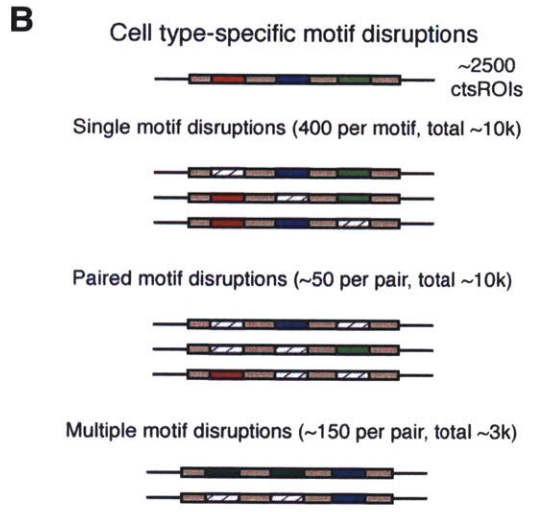
To evaluate the function and activity of binding sites for different classes of TFs in a wide range of regulatory contexts, we used MPRA to measure the activity of 30,000 genomic regions of interest (ROIs) and 21,000 control sequences in five cell lines:

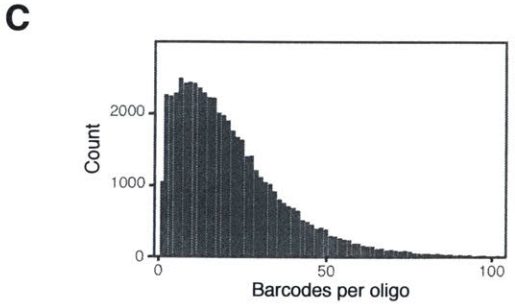
erythroleukemia (K562), liver carcinoma (HepG2), lymphoblast (GM12878), colon cancer (HCT116), and lymphocytes (Jurkat). We selected the ROIs from putative active regulatory elements in the studied cell types, which were defined as NDRs in one or more cell type marked by the active chromatin modification, H3K27ac (using DNAseI hypersensitive (DHS) sites defined by the Roadmap Epigenomics project (18), Assay for Transposase-Accessible Chromatin (ATAC)-seq experiments, and ChIP-seq data from the Roadmap Epigenomics project). The list included 5,000 regions that had cell type-specific active chromatin in each of the five cell lines (ctsROIs), as well as 5,000 ubiquitous regions with active chromatin marks in >90% of the cell types profiled in the Roadmap Epigenomics project (uROIs).

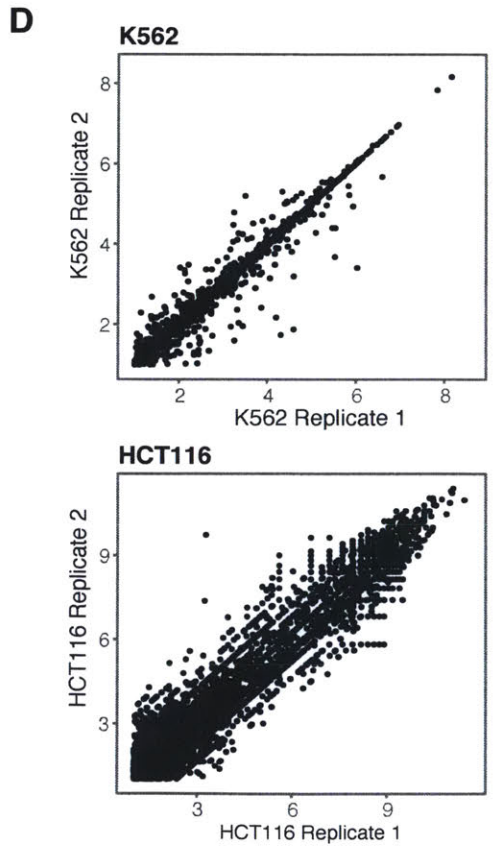
We synthesized a library of DNA oligonucleotides representing 170 bp segments from each of the 30,000 ROIs, centered at the peak of the DHS/ATAC-seq signal (oligos; Fig. 2A). We also synthesized 21,000 negative-control constructs, which were generated by randomly permuting dinucleotides in 3,500 ctsROIs from each cell type and 3,500 uROIs. This 51k "cross-cell type type" pool contains hundreds to thousands of occurrences of most known TF motifs, allowing us to characterize TF motifs and positions correlated with enhancer activity in an unbiased way.

Figure 2 (next page). MPRA design and validation. (A) The cross-cell type MPRA pool included 5,000 regulatory elements with cell type-restricted activity from 6 cell lines, 5,000 ubiquitously active regulatory elements, and 24,500 matched scrambled sequence as controls. The pool was transfected into 6 cell lines and mRNA was sequenced to determine the expression driven by each construct in each cell line. **(B)** Design of cell-specific motif mutation pools. For each cell line, we selected 2,500 endogenously active regulatory elements from (A) and mutated the 20 most enriched TF motifs individually (top) and in pairs (bottom). **(C)** Distribution of number of unique barcodes per construct in the cross-cell type pool. Cell-specific pool distributions were similar (not shown).**(D)** Correlation between experimental replicates of MPRA with cross-cell type pool in K562 and HCT116 cells.


A
5000 cell-type specific regulatory elements from each cell type (ctsROIs)
GM
HepG2
K562
HCT116
Jurkat
U937
5000
Ubiquitous (uROIs)
24,500
Scrambled
Transfect pool into 6 cell types
Isolate mRNA and sequence barcodes
ρexpression
Expression 6 cell types
1788 motifs
GM
HepG2
K562
HCT116
Jurkat
Karpas


B
Cell type-specific motif disruptions
~2500 ctsROIs
Single motif disruptions (400 per motif, total ~10k)
Paired motif disruptions (~50 per pair, total ~10k)
Multiple motif disruptions (~150 per pair, total ~3k)


C
Count
2000
1000
0
0
50
100
Barcodes per oligo


D
K562
K562 Replicate 2
K562 Replicate 1
2
4
6
8
HCT116
HCT116 Replicate 2
HCT116 Replicate 1
3
6
9

As a complementary approach, we also designed cell type-specific pools to directly assess the regulatory contributions of binding sites for the TFs included in our original analysis of positional binding preferences (Fig. 2B). This analysis focused on the 20 most enriched TF motifs in each of the cell lines—which we and others have found often show activity in functional assays (17, 19)—and classified these TFs based on the positional distribution of their motif site into six distinct classes. For each enriched motif, we disrupted (by scrambling and/or reversing the motif site) a single motif occurrence of ~400 ctsROIs from each cell type and all occurrences of ~150 ctsROIs with multiple copies of the motif. Finally, we also disrupted ~100 instances of each pairwise combination of motifs in the ctsROIs. Together, these perturbations allow us to estimate the individual TF activities, identify interaction effects, and investigate role of binding site position. Including the ctsROIs and negative-control sequences, each cell-specific library comprised ~20,000 oligos originating from ~2,500 unique regulatory elements. This 51k cross-cell type oligo pool and the 24k cell-specific pools were cloned using our previously-developed MPRA protocol (17, 20), with a few modifications to accommodate the larger library size (See Methods, Fig. 2C).

To measure the activity of regulatory elements and TF motif sites, we transfected the cross-cell type pool into the five selected cell lines (K562, HepG2, GM12878, HCT116 and Jurkat) and a B-cell lymphoma cell line (Karpas422), and the cell-specific pools into four cell lines (K562, HepG2, GM12878, and Jurkat; the HCT116 oligo pool was contaminated during cloning and dropped from further analysis). mRNA was isolated and reverse transcribed, PCR amplification was performed on the 3' end surrounding the barcode, and the PCR products were deep sequenced to determine

barcode abundances in the transcriptional output. Barcodes in the plasmid library were also PCR amplified and sequenced to determine their relative concentrations in the library. To quantify the enhancer activity, we calculated the log ratio of the mRNA/DNA read counts aggregated over all barcodes associated with each construct. Constructs captured by few barcodes (<5) or with low abundance (<5 counts per million [cpm]) in the DNA pool were excluded, yielding 44,907 constructs (88%) for analysis in the cross-cell type pool and 13,745-20,918 constructs (62-79%) in the cell type-specific pools. Biological replicates were performed for the cross-cell type pool in two of the cell lines (K562 and HCT116), and the activity estimates were highly correlated between biological replicates in both samples (R=0.89-0.96; Fig. 2D).

Evaluating the regulatory activity of the enhancer segments

We first examined the activity of the genomic regulatory elements across the different cell types. The regulatory elements showed greater activity than the scrambled controls in all cell lines (Fig. 3A). Cell lines with more efficient transfections (K562, HepG2) showed stronger signal-to-noise ratios than cell lines that were more difficult to transfect (Jurkat, GM12878, HCT116), as had been observed in previous studies (19, 21).

To estimate the proportion of ROIs that were active in each cell line, we compared the fraction of ROI sequences whose expression decreased upon scrambling to the fraction that would be expected by chance (19). We found that between 55-77% of the ctsROI had higher activity than their scrambled control in their cognate cell type. Since we expect that all active ROI and half the inactive ROI should have higher expression than their scrambled control, we estimate that 8-53% of ctsROIs were active

in their cognate cell type (Table 1). For highly transfectable cell lines, such as K562, Karpas422, and HepG2, the fraction of ctsROIs that were active in cognate cell types was comparable to our previous observations for PPARγ enhancers (17) Not all putative regulatory elements are expected to show activity in MPRA for various reasons, including (1) key functional elements missing from the 180 bp window tested, (2) poised regulatory elements activated by specific environmental cues, or (3) biochemical incompatibility between the enhancer element and the minimal promoter (TATA) in the reporter assay. Cell lines with low transfection efficiency, such as GM12878 and Jurkat cells, showed fewer active constructs due to increased noise in the data. As expected, a larger fraction of the uROI and ctsROIs in cognate cell types were active in each sample (Table 1; Fig. 3A), confirming the assay is specific. The class of uROI had the largest population of active constructs and the highest quantitative expression output in all cell types (Table 1; Fig. 3B). We also found that promoter-proximal uROI and ctsROI (<5kb from a TSS) were associated with higher activity (median of 1.2-fold; Fig. 3C).

	uROIs	**Cognate cell type ctsROIs**	**Other ctsROIs**	**All ROIs**
K562	79%	53%	22%	33%
Karpas422	69%	27%	8%	25%
HepG2	40%	23%	7%	13%
HCT116	36%	19%	2%	7%
GM12878	17%	10%	8%	8%
Jurkat	38%	8%	5%	8%

Table 1. Percentage of ROIs active in each cell line

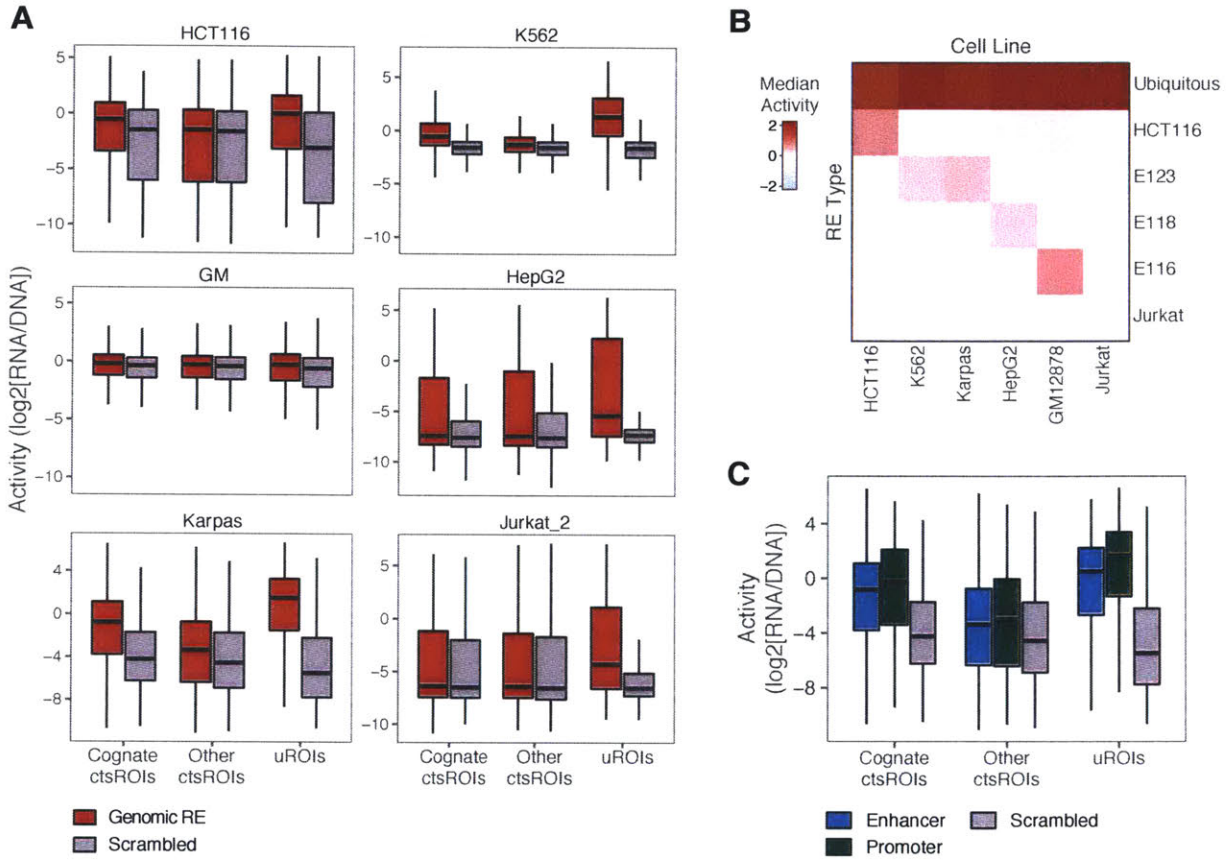


Figure 3. Activity of genomic regulatory elements in MPRA. (A) Boxplot showing the distribution of activities measured by MPRA for ROIs (red) and scrambled control sequences (gray). **(B)** Heatmap showing the median normalize activity of each class of regulatory elements in each sample (cell line). **(C)** Distribution of normalized activities for distal enhancers (blue) and proximal promoters (green) versus control sequences (gray).

Identifying TF motif sites correlated with enhancer activity

We next set out to identify motif sites that were significantly associated with enhancer activity in each cell type. We scanned the 51,000 sequences in the MPRA pool for 1490 known TF-binding motifs corresponding to 612 TFs, and counted the number of non-overlapping occurrences of each motif in each sequence. Across the

30,000 regulatory elements sequences in the pool, the number of occurrences per motif ranged from 90 (HSFY2) to 23,205 (SP4). A total of 467 motifs, corresponding to 409 distinct TFs, were correlated with enhancer activity in at least one of the six cell types ($p_{Bonferroni}<10^{-4}$; Table S1; Fig. 4A). The number of significantly correlated motifs ranged from 33 (in GM12878) and 343 (in HepG2). Only a few motifs reached significance in GM12878 and Jurkat cells due to the small population of ROIs that showed activity above the experimental noise; these numbers would likely increase if transfection efficiency were greater. About half of the 409 TFs identified were significantly correlated with activity in only one cell type (218), while the rest were identified in two or more cell types. Overall, 72% of the TFs identified as significantly correlated with activity in a particular cell type were detectably expressed in that cell type, far greater than the number expected by chance (522 of 724 TF-cell type pairs; $p_{binomial}=4.4x10^{-37}$; Fig. 4A). For many of the TFs that were not expressed, their function suggests that they act as regulators in developmental precursors to the cell type, and most (80%) showed greater motif enrichment in the cell type in which they were identified as active than the other cell types. These motifs thus show significant correlation with activity because they distinguish ctsROIs in cognate cell types from the other ctsROIs. Moreover, we identified known cell type-specific regulators (Fig. 4A), such as GATA factors in erythrocytes (K562) and NFκB and RELA in lymphoblasts (GM12878). These results indicate the motifs identified represent functional TF binding sites in the corresponding cell type.

Since different TFs have been found to activate housekeeping and developmental genes (22), we wondered if different TFs were active in uROIs and

ctsROIs. We therefore analyzed the two classes separately, and found that 226 of the 409 TFs were positively correlated with expression in ctsROIs but not uROIs, including the majority (182) of the TFs active in only one of the cell types. In contrast, only 8 TFs were positively correlated in uROIs but not ctsROIs. Most of the remaining TFs (107) were identified as active in both uROIs and ctsROIs; however, the majority (72%) of these showed greater activity in uROIs. The remaining 68 only reached the threshold for significance in the combined analysis. The TFs identified in both uROIs and ctsROIs are enriched for promoter-binding factors such as SP1, GABPA, and the ELF family of TFs, which are known to operate on CpG island promoters (23) and regulate housekeeping gene expression (24-26), while those identified only in ctsROIs include TFs that tend to bind distal enhancers such as JUN, MAF, IRF4 and ATF3 (27). Overall, our data suggest that a dedicated set of TFs regulates cell type-specific enhancers, while promoter-binding factors tend to operate in both cell type-specific and ubiquitous settings.

Combinatorial regulation by multiple synergistic TFs, none of which are sufficient individually to drive transcription, underlies many examples of cell-specific gene expression. We therefore hypothesized that TFs might be correlated with more robust expression in ctsROIs in cognate cell types that have binding sites for partner TFs compared with other ctsROIs. Indeed, we found that most active TFs (83%) were correlated with greater increases in expression in ctsROI in cognate cell types than other ctsROIs, suggesting pervasive cooperative TF-TF interactions.

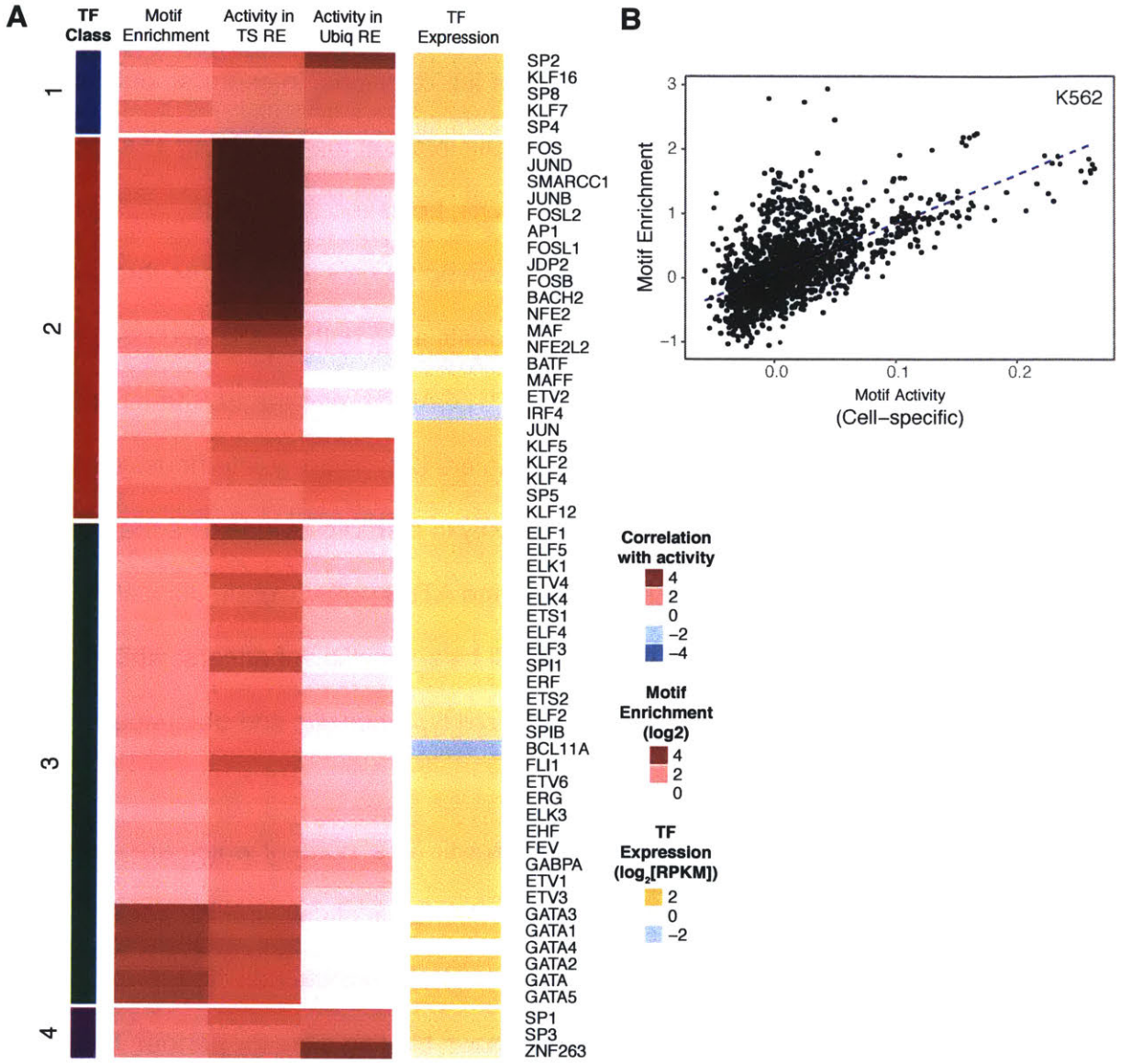


Figure 4. TF motifs correlated with enhancer activity in K562. (A) TF motifs that are significantly correlated with expression in K562 MPRA experiment ($p_{Bonferroni}<10^{-4}$). Heatmap shows enrichment in K562-specific active regulatory elements , Spearman correlation (ρ) of motif counts and enhancer activity in K562-specific ctsROIs and uROIs, and TF expression in K562 cells. TFs are separated by positional class (colored bars). **(B)** Correlation between genomic enrichment and motif activity (Spearman ρ) for all known motifs in K562 cells (r=0.60).

Correlation between TF activity and genomic motif enrichment

We previously observed in adipocytes that motifs that were strongly enriched in genomic enhancers were highly correlated with enhancer activity in functional assays. To test the generality of this observation, we compared TF correlation with motif enrichment in active regulatory elements and activity in ctROIs. We found that cell type-specific activity was strongly associated with genomic enrichment in all the cell types (Spearman ρ=0.20 [GM12878] - 0.60 [K562]). Furthermore, the majority of TF motifs with cell type-specific activity were highly enriched in regulatory elements (72% - 96% [median=89%] of significantly correlated motifs in the top quartile of enrichment; Fig. 4B).

Similarly, most of the enriched motifs were significantly correlated with enhancer activity in the cell type-restricted regulatory elements (55% - 85% of top 20 motifs by enrichment are in the top quartile of activity). Interestingly, four of the six most enriched motifs in K562 cells that are not significantly associated with enhancer activity correspond to TFs that are normally inactive but are induced in response to cellular stresses, including EPAS1 (induced by hypoxia), CREB3L2 (induced by ER stress), ARNT (induced by presence of xenobiotic substances), and MLXIP (induced in response to high glucose levels). Similarly in HepG2 cells, the most enriched motifs that are not identified as correlated with transcriptional output are EPAS1, RREB1 (induced by Ras signaling), and CREB5 (induced by cAMP). These examples suggest that some of the enriched motifs that appear to be inactive may be functional in different conditions not tested in our experiment.

Top enriched motif sites directly regulate transcriptional output

Observing a correlation between the occurrences of a TF binding motif and expression output does not prove that the TF binding site plays a causal role in transcriptional activation. The examples above, for instance, suggest that some TF binding sites may be present in regulatory elements in anticipation of various cellular signals. Conversely, TF binding sites that occur pervasively in regulatory elements might not show a strong correlation with activity, but in fact play an essential role in driving transcription.

To directly determine contribution of highly enriched TF binding sites to the activity of ctsROIs in cognate cell types, we analyzed the effect of motif disruptions in our cell type-specific MPRAs. Because only mutations in constructs that are active regulatory elements would be expected to show an effect, we restricted our analysis to sequences in the top half of wild-type activity. We measured the change in enhancer activity from the wild-type enhancer.

Strikingly, we found that disrupting all but one of the enriched motifs caused significant decreases in expression output (median fold change of 1.5 [IRF8] to 2.4 [SMARCC1]; $p_{Wilcox} = 7.4x10^{-11}$ to $2.9x10^{-166}$; Fig. 5A). The exception—MAF in K562 cells—acts as a repressor in erythrocytes (28) and shows a small median increase in expression upon disruption in our assay. Interestingly, disrupting multiple motif occurrences in regulatory sequences did not reduce expression significantly more than disrupting a single motif site (Fig. 5B), indicating the mutations reduce expression almost to baseline levels. Indeed, disrupting these enriched motifs in active regulatory elements reduced expression about as much as scrambling the whole sequence

(median of 1.7-fold to 3.6-fold). Thus, nearly all highly enriched motifs appear to have strong direct effects on transcription.

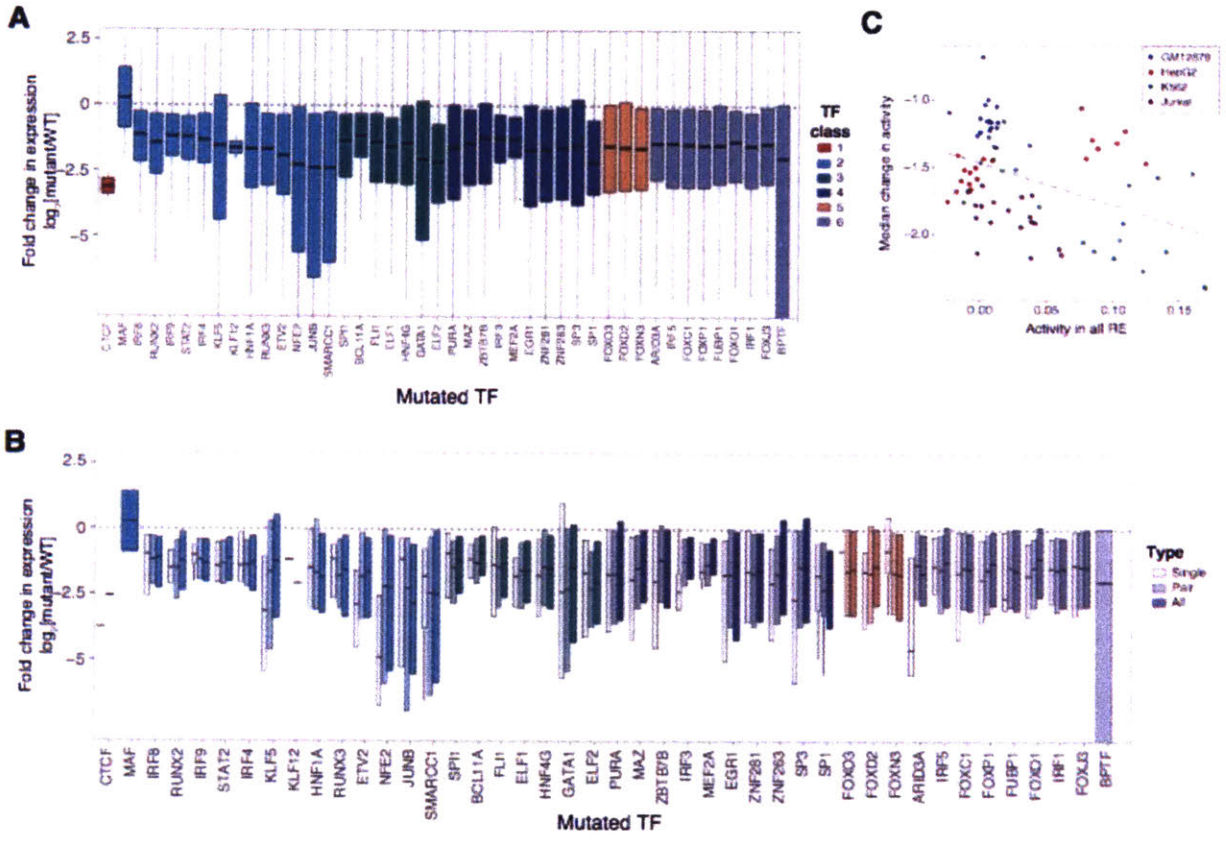


Figure 5. Enriched TF binding sites directly contribute to transcription in active regulatory elements. (A) Change in expression due to mutations in each motif across ~500 regulatory elements in each enriched cell type. TFs are colored by their positional class. **(B)** Change in expression due to mutations of a single motif site (light boxes), pairs of motif site that include the motif in question (medium boxes), and all occurrences of the motif (dark boxes). **(C)** Relationship between the median change in expression due to mutation in each motif and the correlation between motif counts and expression across 5,000 cell type-restricted active regulatory elements. Each point represent one TF in one cell type, and points are colored by sample cell type.

Finally, we examined how closely the relative effects of TF motif disruptions matched the strength of the correlation with enhancer activity in wild-type regulatory elements. We found that the two quantities were very closely related in K562 cells ($\rho_{Spearman}$=0.74) and showed significant but weaker correlation in the other cell types ($\rho_{Spearman}$=0.22 overall; likely due to noise in MPRAs). These results suggest that for TFs that directly regulate transcription in the experimental conditions, the effect of motif perturbations can be accurately estimated from the correlation of the motif occurrences with activity in native regulatory sequences.

Positional binding patterns of 265 TFs fall into six distinct classes

We next investigated the positional binding patterns for the 409 TFs identified as active in one or more MPRA experiment. We restricted our analysis to the 265 motifs that were also enriched in at least one cell type ($p_{ame}<10^{-4}$) and therefore occurred frequently enough to accurately estimate positional distributions. The majority of these TFs (162) were not included in our original positional binding analysis (1). We scanned active regulatory elements from the relevant cell types and identified the position of inferred binding sites relative to the peak of the DHS/ATAC-seq signal. We calculated the density profiles for each TF in ±200 bp regions around the peak and clustered them using k-medoids clustering (1).

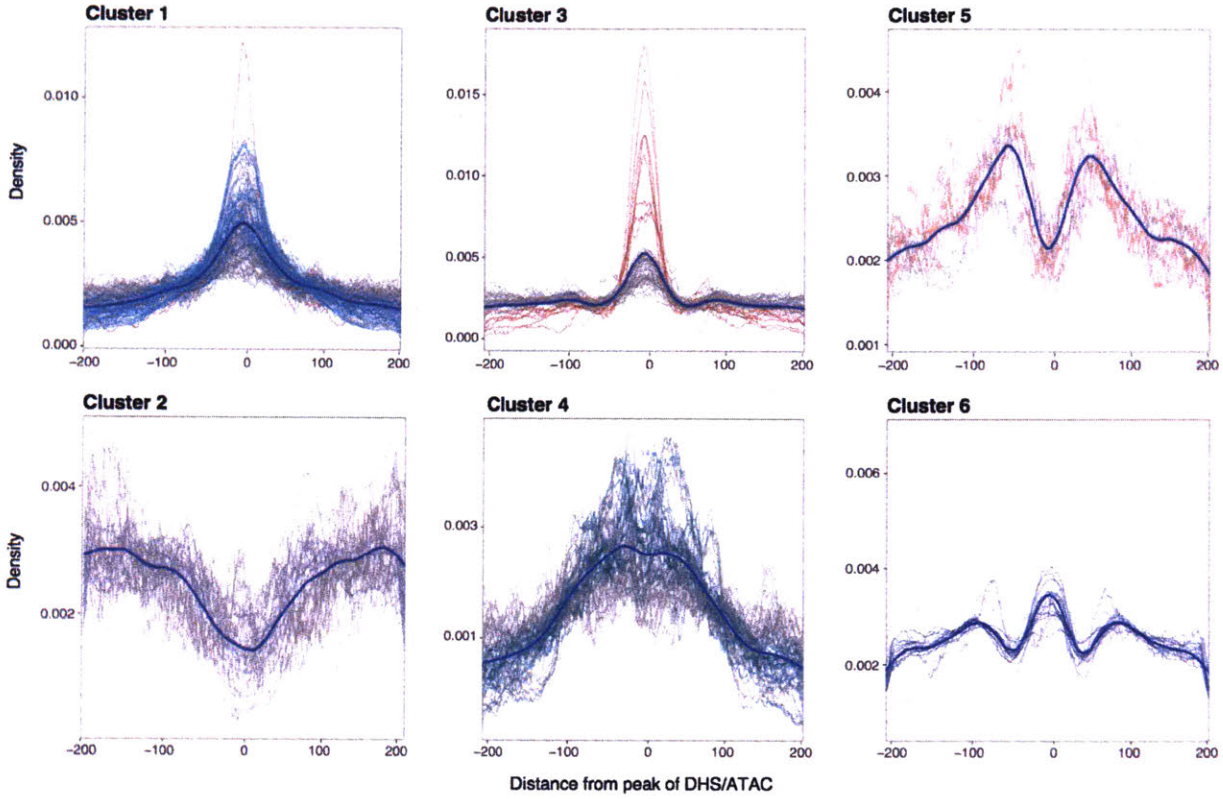


Figure 6. Functional TF motif sites in MPRA display distinct positional binding site patterns. Density profiles of motifs correlated with enhancer activity and enriched in active NDRs in 400 bp regions centered around the peak of the DHS signal (gray lines) were clustered using k-mediods clustering with k=6. Density profiles were generated by calculating the frequency of motif occurrences in 20-bp bins tiled every 1 bp in the region. Blue line depicts the smoothed overall density profile of the cluster using the LOESS method. TFs classified in previous analysis are colored according to their original class assignments.

The analysis identified 6 clusters of distinct regiospecific binding patterns that were extremely similar to those identified with the smaller set of 103 TFs (Fig. 6), with three clusters concentrated near the center of the NDR, and three that tend to occur near the edges or at particular locations within the NDR. Nearly all of the 103 original TFs remained in the same cluster, although a few (13) shifted between the three centrally-located clusters. The new members of the TF classes appear to be consistent

with their hypothesized functional characteristics. For example, new cluster 2 (activation, signaling and cofactor recruitment) factors include ten additional activator protein 1 (AP-1) subunits associated with strong transcriptional activation and five STAT family TFs involved in signal transduction, and new cluster 6 (chromatin remodeling) factors include HMG20B, a component of the LSD1/CoREST complex that demethylates histones (29-31), PRDM1, which mediates gene silencing and chromatin reorganization in multiple tissues (32-34), and two DMRT factors, which regulate chromatin reorganization of sex chromosomes (35).

Binding sites in preferred positions are more likely to be bound by TFs

We next explored whether the positioning of TFs contributes to their role in transcriptional regulation. Since a common observation about mammalian TFs is that only a minority of their potential binding sites are actually occupied *in vivo* (36, 37), we first examined whether motif sites in preferred positions are more likely to be occupied by TFs using ChIP-seq data for 61 TFs in 16 cell types from the ENCODE project. Since ChIP-seq signal is too broad to distinguish occupancy at multiple motif sites within a single ROI, we focused our analysis on ROI with only one motif site for the TF in question. Positions with motif density above the mean were defined as "optimal" for each class, and motif sites for each TF were split into groups with optimal and non-optimal positions.

Strikingly, motif sites in optimal positions were more likely to overlap regions with significant ChIP enrichment for TFs in all six positional classes (median of 1.2- to 2-fold; Fig. 7A). Furthermore, among regions with significant TF binding, NDRs that had motif sites in optimal positions had quantitively higher ChIP enrichment than those with motif

sites in non-optimal positions (median of 1.3- to 4-fold; Fig. 7B). TFs in cluster 1 show the greater effect, perhaps reflecting their role in nucleating and stabilizing the NDRs. This analysis is independent of the number of motif sites in optimal and nonoptimal positions, implying that motif sites in optimal positions are more likely to be functional binding sites.

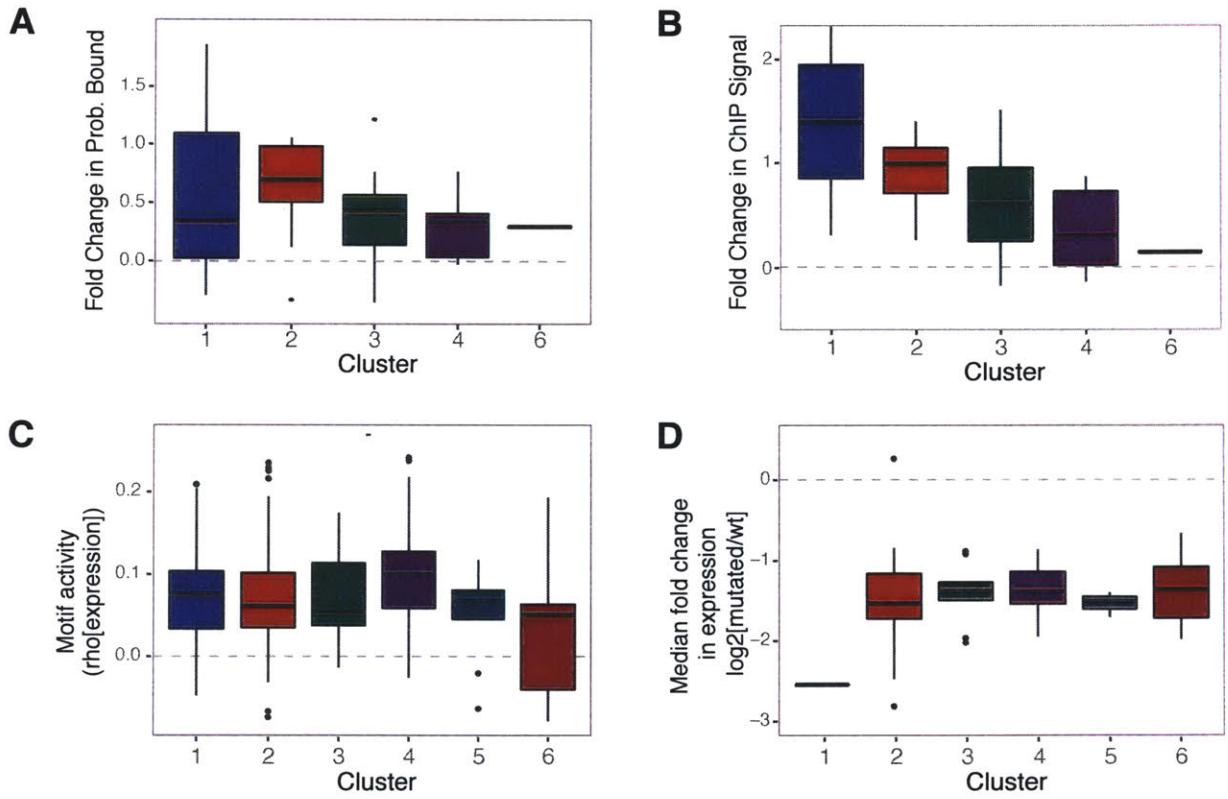


Figure 3. TF motif sites in preferred binding positions are more likely to be bound *in vivo* and drive stronger expression. (A) Relative probabilities that motif sites in optimal and nonoptimal positions overlap regions with significant TF ChIP enrichment for TFs in each class, based on ENCODE ChIP datasets for 61 TFs in 16 cell types. **(C)** Relative quantitative ChIP enrichment in TF-bound regions that have motif sites in optimal and nonoptimal positions. **(C)** Fold change in the correlation between motif counts and enhancer activity for motif sites in optimal and nonoptimal positions across 265 TFs whose motifs are significantly correlated with expression of 30,000 regulatory elements in at least one cell line. **(D)** Median relative change in expression due to

mutations in motifs in optimal and nonoptimal positions across 43 TF motifs tested in 4 cell lines.

Optimally-positioned binding site contribute to greater transcriptional activity

We next studied the how the position of TF binding sites in NDR contributes to their role in transcriptional regulation. We first examined whether TFs in preferred binding positions were associated with greater marginal effects on enhancer activity. For each of the 265 TFs identified above, we calculated the relative correlations of motif sites in optimal and nonoptimal positions with expression output. For all TF classes except one, placing a binding site in an optimal position was associated with a significant increase in transcriptional activity (median of 1.2- to 2-fold per class; Fig. 7C). The final class, Cluster 6, contains TFs associated with chromatin regulation that did not might not contribute to expression in the plasmid MPRA context. Similarly, disrupting TF binding sites in preferred positions caused greater reductions in transcriptional output (Fig. 7D). We note that is not clear from this analysis whether the difference in binding site activity reflects simply the differential TF occupancy or an additional effect of binding site position on TF function.

DISCUSSION

The expression output of a regulatory element is determined by the composition and arrangement of bound TFs, which cooperatively regulate transcription. Because enhancer-bound TFs act collaboratively to control transcription, identifying the full network of active TFs in any enhancers of interest is essential to predict their expression output. We developed a simple paradigm to identify and validate active TFs and binding sites in sets of co-regulated enhancers by (1) measuring the activity of

thousands of genomic regulatory elements using high-throughput reporter assays, (2) identifying all known motifs whose counts significantly correlate with enhancer activity, and (3) characterizing regulatory contributions of correlated motifs using perturbations. Using this approach, we measured the enhancer activity of 30,000 genomic ROIs and 20,000 control sequences, and identified 409 active TFs in developmentally-regulated and ubiquitous regulatory elements in 6 human cell lines.

The 409 TFs can be separated into three groups that show different patterns of activity and motif enrichment across cell types and different types of regulatory elements. The composition of TFs from these three groups determine the inducibility and scope of activity of each genomic regulatory element:

(i) Promoter-biased TFs that strongly activate ubiquitous elements. These TFs tend to be expressed and enriched in regulatory elements in most or all cell types. Many Cluster 4 TFs belong to this class, such as prototypic promoter-binding TFs such as SP1 and GABPA.

(ii) Broadly-expressed TFs that exclusively act in cell type-specific elements, such as AP-1 factors and MYC. This group tends to bind distal enhancers, and partners with cell type-specific TFs to select target enhancers (38, 39). Motifs for these TFs are enriched in cell type-specific elements from many cell types but not in ubiquitous elements. Many Cluster 2 TFs belong to this group, including strong transcriptional activators that can amplify signals from cell- or signal-specific TFs. Finally, a subset of the TFs in this group are activated in response to specific conditions, such as STATs and NF-κB, and mediate signal response pathways.

(iii) TFs with cell type-restricted expression and enrichment that control differentiation and expression of cell type-specific genes. Examples of this type of TF in our dataset include GATA1 in K562 erythroid cells (40) and HNF4α in HepG2 hepatocytes (41). Some of these TFs have pioneering activity, making cell-specific enhancers accessible to other TFs (42).

Importantly, we found that TF motifs associated with enhancer activity in functional assays are highly correlated with motifs that are highly enriched in the active regulatory elements. In the four cell lines tested, the nearly all of the TF binding sites we predicted to be functional based on their enrichment in cell type-restricted enhancers were found to directly contribute to enhancer activity assayed by mutational perturbation. Since chromatin states can be readily mapped, this approximation enables the prediction of active TFs and functional binding sites in a wide range of cellular contexts and conditions.

The prevalence and nature of organizational constraints on motifs in regulatory sequences is an area of active investigation and great interest. Examples from the regulatory sequences that have extensively characterized span a wide range. At one extreme are enhanceosomes such as the IFNβ enhancer, which requires the binding of multiple TFs to overlapping sites. Enhanceosome activity depends on extensive protein-protein interactions between TFs, making them exquisitely sensitive to shifts or rearrangements of motifs (43). At the other extreme are so-called "billboard" enhancers, which contain independent motif sites that contribute additively to expression and can be freely rearranged (44-46). Other examples lie between these two extremes, allowing rearrangement of some elements but not others (7, 47). Evolutionary analyses indicate

enhancer sequences diverge rapidly between species but often retain the ancestral function through compensatory TF binding site turnover (48-50), which would seem to support more flexible organizational principles.

We previously observed that the positional distributions of 103 enriched motifs in NDRs fall into six distinct classes. Here, we show that these enriched motifs constitute functional TF binding sites that contribute to enhancer activity. Furthermore, 162 newly-identified TFs associated with enhancer activity fall into the same six positional classes. Excitingly, we find that motif sites in preferred positions show increased binding and regulatory effects. Across all classes, these motif sites were about twice as likely to be bound by TFs, and had 3-fold higher activity relative to motif sites in non-optimal positions. These results demonstrate that regiospecific binding constraints play a functional role in enhancer activity.

Based on our studies, we propose a novel architectural model combining elements of enhanceosomes and billboard enhancers within an overarching organizational structure based on the position of binding sites within the NDR. The model provides a theoretical framework that can be translated to any cellular context, and makes concrete predictions that can be tested through comparison of enhancer conservation between species and functional studies.

ACKNOWLEDGEMENTS

We would like to thank Chris Burge, Bradley Bernstein, Aviv Regev, Karen Adelman, Telmo Henriques, Cigall Kadoch, Seth Cassel, and Kaylyn Williamson for valuable comments and discussion. This work was supported by the National Human

Genome Research Institute (2U54HG003067-10) (E.S.L.) and the National Institute of General Medical Sciences (T32GM007753) (S.R.G.).

MATERIALS AND METHODS

ATAC-seq

Jurkat and U-937 cells were either left unstimulated or were stimulated for 1 or 4 hours with 2.5ug/ml anti-human CD3 (Biolegend; Cat# 317315) and 50 ng/ml PMA (Sigma Aldrich; Cat# P1585-1MG) for Jurkat cells and 100 ng/ml LPS (Invivogen; Cat# tlrl-peklps) for U-937. Cells were washed with ice cold FACS Buffer and kept on ice until cell sorting. 25,000 live cells from each condition were sorted in to FACS Buffer and pelleted by centrifugation at 500 RCF for 5 minutes at 4C in a pre-cooled fixed angle centrifuge. Cell lines were then tagmented according to the previously described Fast-ATAC protocol (51). Briefly, all supernatant was removed being careful to not disturb the not visible cell pellet. 50 ul transposase mixture (25 ul of 2x TD, 2.5 ul of TDE1, 0.5 ul of 1% digitonin, 22 ul of nuclease-free water) (Cat# FC-121-1030, Illumina; Cat# G9441, Promega) was added to the cells, the pellet was dissociated by pipetting. Transposition reactions were incubated at 37C for 30 minutes in an Eppendorf ThermoMixer with agitation at 300 RPM. Transposed DNA was purified using a QIAgen MinElute Reaction Cleanup kit (Cat# 28204) and purified DNA was eluted in 12 ul elution buffer (10 mM Tris-HCl, pH 8). Transposed fragments were amplified and purified as described previously (52) with modified primers (53). Libraries were quantified using qPCR prior to sequencing. All Fast-ATAC libraries were sequenced using paired-end, dual-index sequencing on a NextSeq with 76x8x8x76 cycle reads at an average read depth of 30 million reads per sample.

Definition of NDRs

To define NDRs for our analysis, we used DNaseI-seq and H3K27ac ChIP-seq data for 45 cell types in the Epigenomics Roadmap and ENCODE Projects (18, 54), as well as ATAC-seq and H3K27ac ChIP-seq data for Jurkat and U937 cells generated in our lab. To select our initial set of NDRs, we intersected DHS/ATAC-seq narrowPeaks regions and H3K27ac gappedPeaks regions called using MACS2 (55) with the standard parameters used by the Epigenomics Roadmap Project. We then filtered out NDRs that were present in more than 24 (50%) of the cell types in our analysis, and selected the top 7,500 cell-type-restricted NDRs for motif enrichment and positioning analysis. We defined the coordinates in the NDRs relative to the summit called by MACS2 (i.e. position with the the maximum DHS/ATAC-seq signal). For MNase-seq analysis, we used data from GM12878 and K562 generated by the ENCODE project. The center of the nucleosomes flanking the NDRs were estimated by identifying the position with the highest MNase-seq read coverage in the 300 bp upstream and downstream of the peak of the DHS signal.

MPRA Cloning

MPRA libraries were clones as in (17) with the following optimizations for the larger pool size. First, we used emulsion PCR to add random 20 nucleotide barcodes to the oligos, which yielded more uniform construct representation than classic PCR (56). The barcoded oligos were cloned into a plasmid background and deep sequenced to map barcodes to constructs, and an inert open reading frame (ORF) with a minimal promoter (TATA) was inserted between the 3' end of the constructs and the 5' end of the barcodes to produce the final pooled library of constructs. To optimize barcode mapping

and reproducibility, we titrated the number of unique transformants during cloning to ensure that each construct was represented by ~10-50 tags in the final plasmid library (average = 17; Fig. 1C).

Cell culture and transfection

All cell lines were grown with 100 units/ml streptomycin and 100 mg/ml penicillin.

K562 cells

We maintained K562 (ATCC) cells a density between 100K and 1M per mL in RPMI-1640 (Thermo Fisher Scientific, Waltham, MA) with 10% heat-inactivated FBS (HIFBS, Thermo Fisher Scientific) and 2mM L-glutamine. We transfected 100M K562 cells using the Lonza (Cologne, Germany) Amaxa 96-well Shuttle according to the manufacturer's instructions for this cell type (except transfecting all 500,000 cells in a single well) with 250 ng of MPRA pool DNA.

HCT116 cells

We maintained HCT116 (ATCC) cells between 20 and 80% confluence in McCoy's 5A Medium with 10% HIFBS, 1 mM sodium pyruvate and 2mM L-glutamine. On the day of transfection media was replaced with 30 mL fresh MEM/FBS followed by transfection with 87.5 µL of Lipofectamine 3000 (Life Technologies, L3000015) and 35 µg of DNA using the manufacturer's protocol. Cells were incubated with transfection reagents for 24 hours, then washed with 15 mL of PBS followed by dissociation with 0.05% trypsin-EDTA (Life Technologies, 25300), centrifugation, PBS wash and a final collection at 300x g prior to storage at -80°C.

HepG2 cells

We maintained HepG2 (ATCC) cells between 20 and 80% confluence in DMEM with

10% HIFBS. On the day of transfection media was replaced with 30 mL fresh MEM/FBS followed by transfection with 87.5 µL of Lipofectamine 3000 (Life Technologies, L3000015) and 35 µg of DNA using the manufacturer's protocol. Cells were incubated with transfection reagents for 24 hours, then washed with 15 mL of PBS followed by dissociation with 0.05% trypsin-EDTA (Life Technologies, 25300), centrifugation, PBS wash and a final collection at 300x g prior to storage at -80°C.

GM12878 cells

We maintained GM12878 (Corriell) cells between $2x10^5$ and $1x10^6$ cell/ml in RPMI-1640 with 15% qualified FBS (QFBS, Invitrogen) and 2mM L-glutamine. For transfections cells were grown to a density of ~$1x10^6$ cells/mL prior to the removal of $1x10^8$ cells. Cells were collected by centrifugation at 120x g and suspended in 1 mL of RPMI with 100 µg of the MPRA library. Electroporation was performed in 100 µL volumes with the Neon transfection system (Life Technologies) applying 3 pulses of 1200 V for 20 ms each. Cells were allowed to recover in 180 mL in RPMI with 15% FBS for 24 hours then collected by centrifugation, washed once with PBS, collected and frozen at -80°C.

Jurkat cells

We maintained Jurkat (ATCC) cells between $1x10^5$ and $3x10^6$ cell/ml in RPMI-1640 with 10% qualified FBS (QFBS, Invitrogen) and 2mM L-glutamine. using the Lonza (Cologne, Germany) Cells were transfected using the Amaxa 96-well Shuttle using the manufacturer's SE kit following the manufacturer's instructions with the program 96-CM-150. Each well contained 1 ug MPRA pool DNA.

Karpas422 cells

We maintained Karpas422 (Bernstein lab) cells between $5x10^5$ and $2x10^6$ cell/ml in RPMI-1640 with 10% qualified FBS (QFBS, Invitrogen) and 2mM L-glutamine. Cells were collected by centrifugation at 120x g and suspended in 1 mL of RPMI with 100 µg of the MPRA library. Electroporation was performed in 100 µL volumes with the Neon transfection system (Life Technologies) applying 3 pulses of 1200 V for 20 ms each. Cells were allowed to recover in 180 mL in RPMI with 10% FBS for 24 hours then collected by centrifugation, washed once with PBS, collected and frozen at -80°C.

Motif enrichment analysis

We calculated motif counts for all vertebrate motifs in TRANSFAC (57), JASPAR (58) and CIS-BP (59) in the genomic NDR sequences, as well as scrambled genomic NDR sequences (holding dinucleotide frequencies constant). To identify enriched motifs in each cell type, we used AME (60) with the mhg method to calculate the enrichment of total number of matches of each motif in the genomic sequences compared to the scrambled sequences. In cases where the combined databases contained multiple PWMs corresponding to a single TF, we selected the most enriched motif in each cell type corresponding to each TF. To remove highly similar motifs, we calculated the pairwise similar of the motifs using the R package PWMEnrich, and removed motifs that had similarity of > 0.8 with a more highly enriched motif. We then selected the top 20 motifs from the filtered list in each cell type for positioning analysis. We called motif sites in the genomic and scrambled sequences using by running FIMO (61) with a p-value threshold of 10^{-4}.

Motif position profiles and clustering

To analyze the positioning of the motifs with NDRs, we collapsed the motif matches to their central position, and calculated the density of each motif in 20-bp windows tiled every 1 bp across the 400 bp centered around the position of maximum DHS/ATAC signal in each NDR. The motif position profiles were then clustered using the pam function from the R package cluster with k=6.

TABLES

Table S1. TF motifs significantly correlated with MPRA enhancer activity

Motif	TF	Sample	ρ	p	Genomic Enrichment	TF Expression	Cluster
M2278_1.02	FOS	HCT116	0.09	0.0E+00	3.23	+++	2
V_AP1_Q4_01	AP1	HCT116	0.08	0.0E+00	2.95	++	2
M2292_1.02	JUND	HCT116	0.08	0.0E+00	3.19	+++	
M4623_1.02	JUNB	HCT116	0.08	0.0E+00	3.26	+++	2
M4565_1.02	FOSL2	HCT116	0.08	0.0E+00	3.09	+++	2
M4526_1.02	SMARCC1	HCT116	0.08	0.0E+00	2.99	+++	2
M4619_1.02	FOSL1	HCT116	0.08	0.0E+00	3.12	+++	2
JDP2_full_1	JDP2	HCT116	0.07	0.0E+00	3.28	+++	2
M6228_1.02	FOSB	HCT116	0.07	0.0E+00	2.47	++	2
V_BACH2_01	BACH2	HCT116	0.06	0.0E+00	2.68	+	2
GABPA_full	GABPA	HCT116	0.06	0.0E+00	0.69	++	3
NFE2_DBD	NFE2	HCT116	0.06	0.0E+00	2.87	++	2
ELK3_DBD	ELK3	HCT116	0.06	0.0E+00	0.65	++	3
ETS1_DBD_1	ETS1	HCT116	0.06	0.0E+00	0.92	++	3
ELK1_full_1	ELK1	HCT116	0.06	0.0E+00	0.64	++	3
ERG_full_1	ERG	HCT116	0.06	0.0E+00	0.82	++	3
EHF_full	EHF	HCT116	0.06	0.0E+00	0.82	++	3
M5420_1.02	ETV1	HCT116	0.06	0.0E+00	0.59	++	3
FEV_DBD	FEV	HCT116	0.06	0.0E+00	0.76	++	3
ELK4_DBD	ELK4	HCT116	0.06	0.0E+00	0.68	++	3
ELF1_DBD	ELF1	HCT116	0.06	0.0E+00	0.79	++	3
FLI1_full_1	FLI1	HCT116	0.06	0.0E+00	0.81	++	3
ETV4_DBD	ETV4	HCT116	0.06	0.0E+00	0.63	++	3
M5422_1.02	ETV3	HCT116	0.06	0.0E+00	0.48	++	3
M5398_1.02	ERF	HCT116	0.06	0.0E+00	0.83	++	3
ETV6_full_2	ETV6	HCT116	0.06	0.0E+00	0.64	++	3
M4452_1.02	BATF	HCT116	0.05	0.0E+00	1.68	-	2
ELF4_full	ELF4	HCT116	0.05	0.0E+00	0.81	++	3
Elf5_DBD	ELF5	HCT116	0.05	0.0E+00	0.98	++	3
ETV5_DBD	ETV5	HCT116	0.05	0.0E+00	0.20	++	-
V_MAF_Q6_01	MAF	HCT116	0.05	0.0E+00	1.52	++	2
ELF3_DBD	ELF3	HCT116	0.05	0.0E+00	0.90	++	3
ETV2_DBD	ETV2	HCT116	0.04	0.0E+00	0.99	++	2
M5209_1.02	SP5	HCT116	0.04	0.0E+00	1.48	++	2
M6360_1.02	NFE2L2	HCT116	0.04	0.0E+00	1.44	++	2
M2301_1.02	NFYB	HCT116	0.04	0.0E+00	-0.02	+	2

M4572_1.02	MAFF	HCT116	0.04	0.0E+00	0.98	++	2
M6152_1.02	ATF1	HCT116	0.04	0.0E+00	1.33	+	2
CREB3_full_1	CREB3	HCT116	0.04	0.0E+00	1.01	+	2
M0300_1.02	ATF2	HCT116	0.04	4.4E-16	1.32	++	1
M6174_1.02	CEBPZ	HCT116	0.04	4.4E-16	0.10	++	-
V_NFY_01	NFY	HCT116	0.04	8.9E-16	0.06	++	2
SP4_full	SP4	HCT116	0.04	4.9E-15	1.16	++	1
Creb5_DBD	CREB5	HCT116	0.04	8.7E-15	1.32	++	2
SPDEF_DBD_1	SPDEF	HCT116	0.04	1.1E-14	0.38	++	-
KLF14_DBD	KLF14	HCT116	0.04	1.7E-14	1.08	++	2
M6240_1.02	FOXI1	HCT116	0.04	6.4E-14	-0.09	++	-
M0405_1.02	KLF7	HCT116	0.03	1.1E-13	1.55	++	1
ATF7_DBD	ATF7	HCT116	0.03	1.2E-13	1.28	+	-
SP1_DBD	SP1	HCT116	0.03	2.0E-13	1.30	++	4
M6180_1.02	CREB1	HCT116	0.03	2.4E-13	0.96	++	-
XBP1_DBD_1	XBP1	HCT116	0.03	2.5E-13	1.07	+	2
V_VJUN_01	JUN	HCT116	0.03	2.6E-13	0.97	+	2
M0443_1.02	KLF12	HCT116	0.03	1.7E-12	1.61	++	2
M6181_1.02	CREM	HCT116	0.03	1.9E-12	0.63	+	2
M4463_1.02	IRF4	HCT116	0.03	6.3E-12	1.11	-	2
M1890_1.02	NFYA	HCT116	0.03	6.8E-12	-0.16	+	2
V_TEL2_Q6	ETV7	HCT116	0.03	4.9E-11	0.68	++	5
SP3_DBD	SP3	HCT116	0.03	6.7E-11	1.11	++	4
M4473_1.02	PBX3	HCT116	0.03	7.9E-11	0.05	++	2
M6221_1.02	ETS2	HCT116	0.03	1.2E-10	0.83	+	3
BATF3_DBD	BATF3	HCT116	0.03	1.5E-10	0.78	+	-
M6324_1.02	KLF4	HCT116	0.03	2.3E-10	1.22	++	2
M2296_1.02	MAFK	HCT116	0.03	2.9E-10	0.35	+	2
KLF16_DBD	KLF16	HCT116	0.03	1.2E-09	0.69	++	1
M5292_1.02	ATF4	HCT116	0.03	1.3E-09	0.17	+++	-
M6373_1.02	NFYC	HCT116	0.03	5.2E-09	-0.06	++	1
NRF1_full	NRF1	HCT116	0.03	7.6E-09	0.55	+	4
V_ETF_Q6	TEAD2	HCT116	0.03	1.9E-08	1.31	+	1
SPIB_DBD	SPIB	HCT116	0.03	4.9E-08	0.73	-	3
M6204_1.02	ELF2	HCT116	0.03	5.4E-08	0.85	++	3
M1871_1.02	KLF2	HCT116	0.03	5.4E-08	1.27	++	2
M0422_1.02	ZIC5	HCT116	-0.03	1.4E-11	-0.93	+	1
M6422_1.02	PLAGL1	HCT116	-0.04	0.0E+00	-0.57	+	1
M2278_1.02	FOS	K562	0.18	0.0E+00	1.70	++	2
M2292_1.02	JUND	K562	0.17	0.0E+00	1.77	++	2
M4526_1.02	SMARCC1	K562	0.17	0.0E+00	1.67	++	2
M4565_1.02	FOSL2	K562	0.17	0.0E+00	1.66	+++	2
M4623_1.02	JUNB	K562	0.17	0.0E+00	1.85	++	2
V_AP1_Q4_01	AP1	K562	0.17	0.0E+00	1.49	++	2
SP1_DBD	SP1	K562	0.17	0.0E+00	1.35	++	4
M4619_1.02	FOSL1	K562	0.16	0.0E+00	1.63	+++	2

M6207_1.02	ELK1	K562	0.16	0.0E+00	0.79	++	3
M5209_1.02	SP5	K562	0.16	0.0E+00	1.55	++	2
JDP2_full_1	JDP2	K562	0.16	0.0E+00	1.77	+++	2
M2391_1.02	KLF5	K562	0.16	0.0E+00	1.25	++	2
M4522_1.02	ELK4	K562	0.16	0.0E+00	0.94	++	3
ELK3_DBD	ELK3	K562	0.16	0.0E+00	0.67	++	3
V_BACH2_01	BACH2	K562	0.16	0.0E+00	1.32	+++	2
M4462_1.02	GABPA	K562	0.15	0.0E+00	0.72	++	3
M0714_1.02	ETS1	K562	0.15	0.0E+00	0.76	++	3
M2314_1.02	SP2	K562	0.15	0.0E+00	1.27	++	1
M2275_1.02	ELF1	K562	0.15	0.0E+00	1.10	++	3
M1871_1.02	KLF2	K562	0.15	0.0E+00	1.23	++	2
M6228_1.02	FOSB	K562	0.15	0.0E+00	1.20	++	2
SP4_full	SP4	K562	0.15	0.0E+00	1.20	+	1
M6324_1.02	KLF4	K562	0.15	0.0E+00	1.23	++	2
ERG_full_1	ERG	K562	0.15	0.0E+00	0.87	++	3
FEV_DBD	FEV	K562	0.15	0.0E+00	0.86	++	3
NFE2_DBD	NFE2	K562	0.15	0.0E+00	1.47	+++	2
M0405_1.02	KLF7	K562	0.15	0.0E+00	1.55	++	1
ETV6_full_2	ETV6	K562	0.15	0.0E+00	0.92	+	3
M0443_1.02	KLF12	K562	0.14	0.0E+00	1.56	++	2
FLI1_DBD_1	FLI1	K562	0.14	0.0E+00	0.74	++	3
Elf5_DBD	ELF5	K562	0.14	0.0E+00	0.99	++	3
ELF4_full	ELF4	K562	0.14	0.0E+00	0.92	++	3
ETV4_DBD	ETV4	K562	0.14	0.0E+00	0.66	++	3
M5420_1.02	ETV1	K562	0.14	0.0E+00	0.58	++	3
EHF_full	EHF	K562	0.14	0.0E+00	0.90	++	3
SP3_DBD	SP3	K562	0.14	0.0E+00	1.25	++	4
KLF14_DBD	KLF14	K562	0.13	0.0E+00	1.08	++	2
KLF16_DBD	KLF16	K562	0.13	0.0E+00	0.93	++	1
ETV3_DBD	ETV3	K562	0.13	0.0E+00	0.54	++	3
M5398_1.02	ERF	K562	0.13	0.0E+00	0.91	++	3
V_NFY_01	NFY	K562	0.13	0.0E+00	0.07	++	2
ETV5_DBD	ETV5	K562	0.13	0.0E+00	0.41	++	-
ELF3_DBD	ELF3	K562	0.13	0.0E+00	0.93	++	3
M6221_1.02	ETS2	K562	0.12	0.0E+00	0.92	+	3
M2301_1.02	NFYB	K562	0.12	0.0E+00	0.02	+	2
SP8_DBD	SP8	K562	0.12	0.0E+00	0.92	++	1
V_MAF_Q6_01	MAF	K562	0.12	0.0E+00	0.90	++	2
M4604_1.02	ZNF263	K562	0.11	0.0E+00	1.22	+	4
M6552_1.02	ZNF148	K562	0.11	0.0E+00	1.10	+	6
V_ETF_Q6	TEAD2	K562	0.11	0.0E+00	1.41	+	1
ETV2_DBD	ETV2	K562	0.11	0.0E+00	1.03	++	2
M4473_1.02	PBX3	K562	0.11	0.0E+00	-0.12	++	2
M6123_1.02	ZNF281	K562	0.11	0.0E+00	1.17	+	4
M6174_1.02	CEBPZ	K562	0.11	0.0E+00	0.02	++	-

M6336_1.02	MAZ	K562	0.11	0.0E+00	1.24	++	4
V_CKROX_Q2	ZBTB7B	K562	0.10	0.0E+00	0.85	++	4
M6360_1.02	NFE2L2	K562	0.10	0.0E+00	1.13	+++	2
M6204_1.02	ELF2	K562	0.10	0.0E+00	0.89	+	3
M4459_1.02	EGR1	K562	0.10	0.0E+00	1.38	+	4
M6535_1.02	WT1	K562	0.10	0.0E+00	1.24	++	4
M6373_1.02	NFYC	K562	0.10	0.0E+00	0.01	++	1
M6325_1.02	KLF6	K562	0.10	0.0E+00	0.63	+	1
V_CACD_01	CACD	K562	0.10	0.0E+00	1.35	++	1
M6240_1.02	FOXI1	K562	0.09	0.0E+00	-0.16	++	-
M1890_1.02	NFYA	K562	0.09	0.0E+00	-0.11	+	2
M6119_1.02	SPI1	K562	0.09	0.0E+00	0.93	+	3
SPIB_DBD	SPIB	K562	0.09	0.0E+00	1.02	+	3
M6442_1.02	PURA	K562	0.09	0.0E+00	1.13	+	4
M6224_1.02	ETV7	K562	0.09	0.0E+00	0.58	-	5
M4452_1.02	BATF	K562	0.08	0.0E+00	0.48	-	2
M6553_1.02	ZNF219	K562	0.08	0.0E+00	0.49	-	1
GATA3_full	GATA3	K562	0.08	0.0E+00	2.22	-	3
M6201_1.02	EGR4	K562	0.08	0.0E+00	0.92	-	4
M2289_1.02	JUN	K562	0.08	0.0E+00	0.69	++	2
M6199_1.02	EGR2	K562	0.08	0.0E+00	0.90	+	1
M4572_1.02	MAFF	K562	0.08	0.0E+00	0.70	++	2
ZNF740_full	ZNF740	K562	0.08	0.0E+00	0.14	+	-
GATA4_DBD	GATA4	K562	0.08	0.0E+00	2.25	-	3
Spic_DBD	SPIC	K562	0.08	0.0E+00	0.92	+	3
KLF13_full	KLF13	K562	0.08	0.0E+00	0.42	+	2
M1868_1.02	GATA2	K562	0.08	0.0E+00	2.06	+++	3
M4600_1.02	GATA1	K562	0.08	0.0E+00	2.18	+++	3
V_GATA_Q6	GATA	K562	0.08	0.0E+00	2.19	-	3
M0300_1.02	ATF2	K562	0.08	0.0E+00	1.05	++	1
M6322_1.02	KLF1	K562	0.08	0.0E+00	0.91	++	1
Creb5_DBD	CREB5	K562	0.08	0.0E+00	1.07	++	2
SPDEF_full_1	SPDEF	K562	0.08	0.0E+00	0.51	++	-
GATA5_DBD	GATA5	K562	0.07	0.0E+00	2.11	+++	3
NRF1_full	NRF1	K562	0.07	0.0E+00	0.91	+	4
ATF7_DBD	ATF7	K562	0.07	0.0E+00	0.80	++	-
M4463_1.02	IRF4	K562	0.07	0.0E+00	0.52	-	2
M6321_1.02	KLF15	K562	0.07	0.0E+00	0.94	+	4
M4453_1.02	BCL11A	K562	0.07	0.0E+00	0.84	-	3
V_CREBATF_Q6	CREB1	K562	0.07	0.0E+00	0.81	++	-
M6547_1.02	ZFX	K562	0.07	0.0E+00	0.70	+	1
CREB3_full_1	CREB3	K562	0.07	0.0E+00	0.89	++	2
M4536_1.02	E2F1	K562	0.07	0.0E+00	1.10	+	1
V_STAT_Q6	ZNF143	K562	0.07	0.0E+00	0.40	+	2
ATF4_DBD	ATF4	K562	0.07	0.0E+00	0.21	+++	-
M1963_1.02	ZFY	K562	0.07	0.0E+00	0.72	-	1

M6152_1.02	ATF1	K562	0.07	0.0E+00	1.05	++	2
M0608_1.02	MLL	K562	0.07	0.0E+00	0.51	+	-
M1915_1.02	ZNF76	K562	0.07	0.0E+00	0.55	+	2
M2273_1.02	E2F6	K562	0.07	0.0E+00	0.90	+	4
M5932_1.02	TFEC	K562	0.07	0.0E+00	0.73	+++	2
TFEB_full	TFEB	K562	0.06	0.0E+00	0.94	+++	2
V_STAT1_01	STAT1	K562	0.06	0.0E+00	0.93	+	2
M6491_1.02	STAT5A	K562	0.06	0.0E+00	1.18	++	2
XBP1_DBD_1	XBP1	K562	0.06	0.0E+00	1.16	++	2
EGR3_DBD	EGR3	K562	0.06	0.0E+00	1.07	+	1
M1917_1.02	USF1	K562	0.06	0.0E+00	0.90	++	2
BATF3_DBD	BATF3	K562	0.06	0.0E+00	0.75	++	-
M6181_1.02	CREM	K562	0.06	0.0E+00	0.64	++	2
M4451_1.02	ATF3	K562	0.06	0.0E+00	0.91	+++	2
M4640_1.02	ZBTB7A	K562	0.06	0.0E+00	0.45	+	1
M6420_1.02	PLAG1	K562	0.06	0.0E+00	0.31	+	1
M4612_1.02	CTCFL	K562	0.06	0.0E+00	0.01	++	1
M4680_1.02	BACH1	K562	0.06	0.0E+00	0.53	+++	2
M4427_1.02	CTCF	K562	0.06	0.0E+00	0.16	++	1
M2296_1.02	MAFK	K562	0.06	0.0E+00	0.49	+	2
M4527_1.02	SMARCC2	K562	0.06	0.0E+00	0.18	+	-
M4481_1.02	USF2	K562	0.06	0.0E+00	0.93	+++	2
M6517_1.02	TFE3	K562	0.06	0.0E+00	0.51	++	2
M1919_1.02	YY1	K562	0.05	0.0E+00	0.38	+	-
M6492_1.02	STAT5B	K562	0.05	0.0E+00	0.94	++	2
M1581_1.02	CIC	K562	0.05	0.0E+00	0.50	+	2
M6537_1.02	YBX1	K562	0.05	0.0E+00	-0.24	+++	1
Srebf1_DBD	SREBF1	K562	0.05	0.0E+00	0.82	+++	2
M6333_1.02	MAFG	K562	0.05	0.0E+00	0.56	++	2
M6359_1.02	NFE2L1	K562	0.05	0.0E+00	0.56	++	2
M6548_1.02	ZIC1	K562	0.05	0.0E+00	0.33	+	1
V_IK_Q5	IK	K562	0.05	0.0E+00	0.81	++	1
M6313_1.02	IRF8	K562	0.05	0.0E+00	0.80	+	2
M0969_1.02	LHX8	K562	0.05	0.0E+00	-0.54	+	2
SREBF2_DBD	SREBF2	K562	0.05	0.0E+00	0.66	+++	2
V_PAX4_03	PAX4	K562	0.05	0.0E+00	0.82	-	1
M4478_1.02	STAT3	K562	0.05	0.0E+00	0.60	++	2
V_MINI19_B	MINI19	K562	0.05	0.0E+00	0.34	+++	-
M5632_1.02	MLX	K562	0.05	0.0E+00	0.68	++	-
M4537_1.02	E2F4	K562	0.05	0.0E+00	0.91	++	1
V_STRA13_01	STRA13	K562	0.05	0.0E+00	0.69	+++	2
M1882_1.02	IRF1	K562	0.05	0.0E+00	1.49	+	6
V_NKX25_Q5	NKX2-5	K562	0.04	0.0E+00	0.06	-	-
M0609_1.02	DNMT1	K562	0.04	0.0E+00	0.90	++	1
M6258_1.02	GATA6	K562	0.04	0.0E+00	1.91	-	-
M4692_1.02	SIX5	K562	0.04	0.0E+00	0.32	+	-

M6496_1.02	STAT4	K562	0.04	0.0E+00	0.72	+	2
M6345_1.02	MITF	K562	0.04	0.0E+00	0.53	+	2
M6162_1.02	ARNTL	K562	0.04	0.0E+00	0.85	-	2
YY2_full_1	YY2	K562	0.04	0.0E+00	0.57	+	5
V_P300_01	EP300	K562	0.04	0.0E+00	0.39	+	-
V_AP2_Q6_01	TFAP2A	K562	0.04	0.0E+00	0.37	-	1
Mafb_DBD_2	MAFB	K562	0.04	0.0E+00	0.48	++	2
BHLHE41_full	BHLHE41	K562	0.04	0.0E+00	0.94	+++	-
M6461_1.02	RXRB	K562	0.04	0.0E+00	0.57	++	2
M2323_1.02	ZBTB33	K562	0.04	0.0E+00	0.11	++	-
M6197_1.02	E4F1	K562	0.04	2.2E-16	0.75	++	-
M6456_1.02	RREB1	K562	0.04	2.2E-16	1.33	+	4
M6330_1.02	MAFA	K562	0.04	4.4E-16	0.60	-	2
V_MYCMAX_03	MYC	K562	0.04	8.9E-16	1.43	++	3
M6131_1.02	TFCP2L1	K562	0.04	1.3E-15	0.22	-	-
M6309_1.02	IRF3	K562	0.04	4.0E-15	1.00	++	4
M4454_1.02	BRCA1	K562	0.04	6.2E-15	-0.04	++	-
M6154_1.02	ATF5	K562	0.04	7.3E-15	0.55	+	5
V_VDR_Q3	VDR	K562	0.04	8.4E-15	0.45	+	1
M2307_1.02	PRDM1	K562	0.04	1.2E-14	0.67	+	6
M6381_1.02	NR0B1	K562	0.04	6.0E-14	0.48	-	-
M6443_1.02	RARA	K562	0.04	6.5E-14	0.85	+	2
M5689_1.02	NRL	K562	0.03	1.4E-13	0.40	+	2
M6523_1.02	THRB	K562	0.03	3.3E-13	0.57	+	2
M6306_1.02	INSM1	K562	0.03	5.9E-13	-0.19	-	-
M0305_1.02	CREB3L2	K562	0.03	1.1E-12	1.63	+	-
M0398_1.02	ZSCAN10	K562	0.03	2.8E-12	0.03	-	-
V_PAX5_01	PAX5	K562	0.03	3.6E-12	0.28	-	-
M6514_1.02	TFCP2	K562	0.03	4.0E-12	0.29	++	-
M6519_1.02	TGIF1	K562	0.03	4.2E-12	0.36	+	-
V_EBOX_Q6_01	TCF3	K562	0.03	9.2E-12	0.49	++	2
M1432_1.02	NR2E1	K562	0.03	1.1E-11	0.55	+	2
MLXIPL_full	MLXIPL	K562	0.03	1.1E-11	1.11	++	-
M6212_1.02	EPAS1	K562	0.03	1.6E-11	1.94	+	4
M4511_1.02	RXRA	K562	0.03	2.7E-11	0.41	+	2
NFIL3_DBD	NFIL3	K562	0.03	3.5E-11	0.35	+	-
M5435_1.02	FOXB1	K562	0.03	8.9E-11	0.83	-	1
M6356_1.02	MZF1	K562	0.03	1.3E-10	0.13	+	-
M6270_1.02	NHLH1	K562	0.03	1.4E-10	0.49	-	3
M2392_1.02	RFX2	K562	0.03	1.5E-10	0.24	+	-
BHLHB2_DBD	BHLHE40	K562	0.03	3.7E-10	1.04	+++	-
M6463_1.02	SMAD1	K562	0.03	6.2E-10	0.40	+	3
M1927_1.02	MYCL1	K562	0.03	1.1E-09	1.10	++	2
M6308_1.02	IRF2	K562	0.03	1.2E-09	0.63	+	-
M2283_1.02	FOXP1	K562	0.03	1.2E-09	2.45	+	6
M1458_1.02	RORB	K562	0.03	1.4E-09	0.59	++	2

RFX3_DBD_1	RFX3	K562	0.03	1.4E-09	0.29	+	-
M4543_1.02	MXI1	K562	0.03	1.9E-09	1.13	++	-
M4635_1.02	STAT2	K562	0.03	2.4E-09	0.78	++	2
V_ZF5_B	ZBTB14	K562	0.03	4.8E-09	-0.05	+++	-
M6311_1.02	IRF5	K562	0.03	5.2E-09	1.14	-	6
V_EVI1_03	MECOM	K562	0.03	5.7E-09	1.21	+	5
M6521_1.02	THRA	K562	0.03	5.7E-09	0.34	++	3
V_GLI_Q2	GLI1	K562	0.03	6.8E-09	0.15	-	-
MNT_DBD	MNT	K562	0.03	6.8E-09	1.62	++	-
M6337_1.02	MBD2	K562	0.03	7.7E-09	0.12	+	-
M0212_1.02	TCFL5	K562	0.03	7.8E-09	0.95	++	-
M6323_1.02	KLF3	K562	0.03	9.2E-09	0.57	+	2
M4545_1.02	ZNF683	K562	0.03	1.1E-08	0.47	+	6
M6150_1.02	ARNT2	K562	0.03	1.2E-08	0.99	-	1
M1528_1.02	RFX6	K562	0.03	1.2E-08	0.26	+	-
M0211_1.02	MLXIP	K562	0.03	1.4E-08	1.48	++	-
M6146_1.02	TFAP2D	K562	0.03	2.2E-08	0.56	-	1
V_PAX6_Q2	PAX6	K562	0.03	2.6E-08	0.30	-	2
M1968_1.02	EBF1	K562	0.03	2.9E-08	-0.06	-	-
V_ARNT_01	ARNT	K562	0.03	3.0E-08	1.44	++	-
MAX_DBD_2	MAX	K562	0.03	4.7E-08	1.58	++	3
M6159_1.02	BCL6	K562	0.03	5.5E-08	0.23	+	-
M6515_1.02	TFDP1	K562	0.03	5.5E-08	0.58	++	-
M1582_1.02	HMG20B	K562	-0.03	1.7E-08	1.31	++	6
M5512_1.02	HIC2	K562	-0.03	6.6E-09	-0.42	++	-
JDP2_full_1	JDP2	GM_2	0.09	0.0E+00	1.59	++	2
M2278_1.02	FOS	GM_2	0.09	0.0E+00	1.41	++	2
V_AP1_Q4_01	AP1	GM_2	0.09	0.0E+00	1.21	++	2
M2292_1.02	JUND	GM_2	0.09	0.0E+00	1.40	++	2
M4623_1.02	JUNB	GM_2	0.08	0.0E+00	1.53	++	2
M4526_1.02	SMARCC1	GM_2	0.08	0.0E+00	1.20	++	2
M4565_1.02	FOSL2	GM_2	0.08	0.0E+00	1.24	++	2
M4619_1.02	FOSL1	GM_2	0.08	0.0E+00	1.30	++	2
NFE2_DBD	NFE2	GM_2	0.07	0.0E+00	1.11	+	2
M6228_1.02	FOSB	GM_2	0.07	0.0E+00	0.92	++	2
V_BACH2_01	BACH2	GM_2	0.07	0.0E+00	0.78	+	2
M4452_1.02	BATF	GM_2	0.06	0.0E+00	1.23	++	2
V_MAF_Q6_01	MAF	GM_2	0.06	0.0E+00	0.74	++	2
M6360_1.02	NFE2L2	GM_2	0.05	0.0E+00	0.57	++	2
M4463_1.02	IRF4	GM_2	0.04	8.9E-16	1.28	+++	2
M6333_1.02	MAFG	GM_2	0.04	2.2E-14	0.21	++	2
M6359_1.02	NFE2L1	GM_2	0.04	2.2E-14	0.21	++	2
M4572_1.02	MAFF	GM_2	0.04	3.8E-14	0.35	+	2
M2296_1.02	MAFK	GM_2	0.03	6.6E-11	0.12	+	2
V_VJUN_01	JUN	GM_2	0.03	1.3E-08	0.97	++	2
V_CEBPA_01	CEBPA	GM_2	0.03	2.4E-08	-0.49	++	-

M1581_1.02	CIC	GM_2	0.03	2.7E-08	0.02	+	2
V_TFE_Q6	TFE3	GM_2	0.03	4.8E-08	0.54	+	2
M6552_1.02	ZNF148	GM_2	-0.03	5.1E-08	1.11	+	6
V_MINI19_B	MINI19	GM_2	-0.03	4.6E-08	0.38	+++	-
M1963_1.02	ZFY	GM_2	-0.03	3.7E-08	0.85	-	1
M0404_1.02	ZNF202	GM_2	-0.03	9.1E-09	-0.83	+	6
M0198_1.02	SOHLH2	GM_2	-0.03	5.7E-09	0.47	+	-
M0609_1.02	DNMT1	GM_2	-0.03	2.9E-10	0.89	++	1
M6337_1.02	MBD2	GM_2	-0.03	5.5E-11	0.16	++	-
M0422_1.02	ZIC5	GM_2	-0.03	2.7E-11	-0.73	-	1
M6547_1.02	ZFX	GM_2	-0.03	1.9E-13	0.81	+	1
M6146_1.02	TFAP2D	GM_2	-0.04	0.0E+00	0.72	-	1
V_SP1_Q6	SP1	HepG2	0.13	0.0E+00	1.02	++	4
M6482_1.02	SP3	HepG2	0.12	0.0E+00	1.08	+	4
M2314_1.02	SP2	HepG2	0.12	0.0E+00	0.98	++	1
M5209_1.02	SP5	HepG2	0.12	0.0E+00	0.99	++	2
M6483_1.02	SP4	HepG2	0.12	0.0E+00	0.89	+	1
M2391_1.02	KLF5	HepG2	0.11	0.0E+00	0.92	++	2
M0405_1.02	KLF7	HepG2	0.11	0.0E+00	1.11	++	1
M6539_1.02	ZBTB7B	HepG2	0.11	0.0E+00	1.08	+	4
M6324_1.02	KLF4	HepG2	0.11	0.0E+00	0.86	++	2
M0443_1.02	KLF12	HepG2	0.11	0.0E+00	0.94	+	2
M1871_1.02	KLF2	HepG2	0.11	0.0E+00	0.81	++	2
M4459_1.02	EGR1	HepG2	0.10	0.0E+00	1.10	+	4
KLF14_DBD	KLF14	HepG2	0.10	0.0E+00	0.58	++	2
M6535_1.02	WT1	HepG2	0.10	0.0E+00	0.97	+	4
KLF16_DBD	KLF16	HepG2	0.10	0.0E+00	0.62	++	1
M6547_1.02	ZFX	HepG2	0.10	0.0E+00	0.67	+	1
M6552_1.02	ZNF148	HepG2	0.10	0.0E+00	1.01	+	6
V_ETF_Q6	TEAD2	HepG2	0.10	0.0E+00	1.02	+	1
M6123_1.02	ZNF281	HepG2	0.09	0.0E+00	1.06	+	4
M4536_1.02	E2F1	HepG2	0.09	0.0E+00	0.90	++	1
M6201_1.02	EGR4	HepG2	0.09	0.0E+00	0.78	-	4
M5856_1.02	SP8	HepG2	0.09	0.0E+00	0.64	++	1
M6325_1.02	KLF6	HepG2	0.08	0.0E+00	0.61	++	1
M6336_1.02	MAZ	HepG2	0.08	0.0E+00	1.23	++	4
V_CACD_01	CACD	HepG2	0.08	0.0E+00	1.03	+	1
M4462_1.02	GABPA	HepG2	0.08	0.0E+00	0.26	+++	3
M6207_1.02	ELK1	HepG2	0.08	0.0E+00	0.49	+++	3
M4640_1.02	ZBTB7A	HepG2	0.08	0.0E+00	0.41	+	1
M4522_1.02	ELK4	HepG2	0.08	0.0E+00	0.52	+++	3
M4604_1.02	ZNF263	HepG2	0.08	0.0E+00	1.20	+	4
V_NFY_01	NFY	HepG2	0.08	0.0E+00	-0.26	+	2
M6442_1.02	PURA	HepG2	0.08	0.0E+00	1.11	+	4
ELK3_DBD	ELK3	HepG2	0.08	0.0E+00	0.28	+++	3
M6321_1.02	KLF15	HepG2	0.08	0.0E+00	0.85	+	4

M6199_1.02	EGR2	HepG2	0.07	0.0E+00	0.62	+	1
M4612_1.02	CTCFL	HepG2	0.07	0.0E+00	0.00	+	1
V_AP2_Q6_01	TFAP2A	HepG2	0.07	0.0E+00	0.31	-	1
M1963_1.02	ZFY	HepG2	0.07	0.0E+00	0.64	+	1
ETV5_DBD	ETV5	HepG2	0.07	0.0E+00	0.13	+++	-
M6553_1.02	ZNF219	HepG2	0.07	0.0E+00	0.29	+	1
M4473_1.02	PBX3	HepG2	0.07	0.0E+00	-0.31	+	2
FLI1_DBD_1	FLI1	HepG2	0.07	0.0E+00	0.42	+++	3
M2301_1.02	NFYB	HepG2	0.07	0.0E+00	-0.27	+	2
ETS1_full_1	ETS1	HepG2	0.07	0.0E+00	0.49	+++	3
M2305_1.02	NRF1	HepG2	0.07	0.0E+00	0.51	+	4
ETV4_DBD	ETV4	HepG2	0.07	0.0E+00	0.31	+++	3
ELF1_DBD	ELF1	HepG2	0.07	0.0E+00	0.48	+	3
M6146_1.02	TFAP2D	HepG2	0.07	0.0E+00	0.40	-	1
FEV_DBD	FEV	HepG2	0.07	0.0E+00	0.43	+++	3
ETV6_full_2	ETV6	HepG2	0.07	0.0E+00	0.36	++	3
ETV1_DBD	ETV1	HepG2	0.07	0.0E+00	0.28	+++	3
M0609_1.02	DNMT1	HepG2	0.07	0.0E+00	0.76	+	1
ERG_full_1	ERG	HepG2	0.07	0.0E+00	0.63	+++	3
M6322_1.02	KLF1	HepG2	0.07	0.0E+00	0.59	+	1
ZNF740_full	ZNF740	HepG2	0.07	0.0E+00	-0.06	+	-
M5422_1.02	ETV3	HepG2	0.07	0.0E+00	0.03	+++	3
M6373_1.02	NFYC	HepG2	0.07	0.0E+00	-0.27	+	1
M6420_1.02	PLAG1	HepG2	0.07	0.0E+00	0.34	+	1
EHF_full	EHF	HepG2	0.07	0.0E+00	0.55	+	3
M4427_1.02	CTCF	HepG2	0.06	0.0E+00	0.17	+	1
ELF4_full	ELF4	HepG2	0.06	0.0E+00	0.54	+	3
V_MINI19_B	MINI19	HepG2	0.06	0.0E+00	0.18	+++	-
M6174_1.02	CEBPZ	HepG2	0.06	0.0E+00	-0.20	+	-
M0608_1.02	MLL	HepG2	0.06	0.0E+00	0.42	+	-
ERF_DBD	ERF	HepG2	0.06	0.0E+00	0.61	+++	3
EGR3_DBD	EGR3	HepG2	0.06	0.0E+00	0.80	+	1
Elf5_DBD	ELF5	HepG2	0.06	0.0E+00	0.64	+++	3
M6221_1.02	ETS2	HepG2	0.06	0.0E+00	0.65	++	3
M1890_1.02	NFYA	HepG2	0.06	0.0E+00	-0.39	+	2
M6144_1.02	TFAP2B	HepG2	0.06	0.0E+00	0.09	-	-
M6337_1.02	MBD2	HepG2	0.06	0.0E+00	0.21	+	-
M4537_1.02	E2F4	HepG2	0.06	0.0E+00	0.76	++	1
ELF3_full	ELF3	HepG2	0.06	0.0E+00	0.48	+	3
M6240_1.02	FOXI1	HepG2	0.05	0.0E+00	-0.23	+	-
M5591_1.02	KLF13	HepG2	0.05	0.0E+00	-0.11	+	2
M4526_1.02	SMARCC1	HepG2	0.05	0.0E+00	0.98	++	2
M2273_1.02	E2F6	HepG2	0.05	0.0E+00	0.79	+	4
M6381_1.02	NR0B1	HepG2	0.05	0.0E+00	0.34	-	-
V_CREBATF_Q6	CREB1	HepG2	0.05	0.0E+00	0.51	++	-
HINFP1_full_3	HINFP	HepG2	0.05	0.0E+00	0.32	+	-

M4565_1.02	FOSL2	HepG2	0.05	0.0E+00	0.95	++	2
M6548_1.02	ZIC1	HepG2	0.05	0.0E+00	0.16	-	1
V_HIC1_03	HIC1	HepG2	0.05	0.0E+00	0.03	-	-
M6152_1.02	ATF1	HepG2	0.05	0.0E+00	0.89	+++	2
ZIC3_full	ZIC3	HepG2	0.05	0.0E+00	-0.37	-	1
M6204_1.02	ELF2	HepG2	0.05	0.0E+00	0.64	+	3
M6537_1.02	YBX1	HepG2	0.05	0.0E+00	-0.41	+++	1
M4619_1.02	FOSL1	HepG2	0.05	0.0E+00	0.89	++	2
V_STAF_02	ZNF143	HepG2	0.05	0.0E+00	0.32	+	2
M2278_1.02	FOS	HepG2	0.05	0.0E+00	1.00	++	2
M4525_1.02	TFAP2C	HepG2	0.04	0.0E+00	0.15	-	-
V_AP1_Q4_01	AP1	HepG2	0.04	0.0E+00	0.80	+	2
CREB3_full_1	CREB3	HepG2	0.04	0.0E+00	0.65	+++	2
M5965_1.02	ZIC4	HepG2	0.04	0.0E+00	-0.32	-	1
M5587_1.02	JDP2	HepG2	0.04	0.0E+00	1.02	++	2
M6422_1.02	PLAGL1	HepG2	0.04	0.0E+00	-0.52	-	1
M2292_1.02	JUND	HepG2	0.04	0.0E+00	1.02	++	2
Zic3_DBD	Zic3	HepG2	0.04	0.0E+00	-0.26	-	-
M0300_1.02	ATF2	HepG2	0.04	0.0E+00	0.83	++	1
V_BACH2_01	BACH2	HepG2	0.04	0.0E+00	0.63	+	2
M6273_1.02	HEY2	HepG2	0.04	0.0E+00	0.47	-	2
M4623_1.02	JUNB	HepG2	0.04	0.0E+00	1.13	++	2
M4451_1.02	ATF3	HepG2	0.04	0.0E+00	0.89	++	2
M6514_1.02	TFCP2	HepG2	0.04	0.0E+00	0.28	+	-
M1915_1.02	ZNF76	HepG2	0.04	0.0E+00	0.43	+	2
V_PAX4_01	PAX4	HepG2	0.04	0.0E+00	0.35	-	1
ETV2_DBD	ETV2	HepG2	0.04	0.0E+00	0.71	+++	2
M0432_1.02	ZFP161	HepG2	0.04	0.0E+00	0.61	+	-
M6228_1.02	FOSB	HepG2	0.04	0.0E+00	0.65	+	2
V_MTF1_Q4	MTF1	HepG2	0.04	0.0E+00	0.37	+	2
M6556_1.02	ZNF350	HepG2	0.04	0.0E+00	0.08	+	-
M0212_1.02	TCFL5	HepG2	0.04	0.0E+00	0.62	++	-
M4481_1.02	USF2	HepG2	0.04	0.0E+00	0.70	++	2
M6306_1.02	INSM1	HepG2	0.04	0.0E+00	-0.06	-	-
V_PAX5_01	PAX5	HepG2	0.04	0.0E+00	0.24	-	-
V_SREBP_Q6	SREBF1	HepG2	0.04	0.0E+00	0.59	++	2
SPDEF_full_1	SPDEF	HepG2	0.04	0.0E+00	0.32	+	-
XBP1_DBD_1	XBP1	HepG2	0.04	0.0E+00	0.99	+++	2
Creb5_DBD	CREB5	HepG2	0.04	0.0E+00	1.18	+++	2
M5293_1.02	ATF7	HepG2	0.04	0.0E+00	1.06	+++	-
V_VJUN_01	JUN	HepG2	0.04	4.0E-15	0.76	+++	2
M2323_1.02	ZBTB33	HepG2	0.04	9.3E-15	0.01	+	-
V_IK_Q5	IK	HepG2	0.04	1.4E-14	0.68	++	1
M0404_1.02	ZNF202	HepG2	0.04	2.0E-14	-1.06	+	6
M6191_1.02	E2F2	HepG2	0.04	2.9E-14	-0.35	+	-
NFE2_DBD	NFE2	HepG2	0.04	3.2E-14	0.86	+	2

V_MAF_Q6_01	MAF	HepG2	0.04	3.8E-14	0.47	++	2
M0085_1.02	TFAP2E	HepG2	0.04	5.9E-14	0.09	-	-
V_R_01	R	HepG2	0.04	7.1E-14	0.12	+++	-
M6456_1.02	RREB1	HepG2	0.03	1.5E-13	1.27	+	4
V_ZF5_B	ZBTB14	HepG2	0.03	1.5E-13	-0.08	+++	-
M6150_1.02	ARNT2	HepG2	0.03	2.0E-13	0.69	-	1
M6197_1.02	E4F1	HepG2	0.03	2.1E-13	0.49	++	-
M6192_1.02	E2F3	HepG2	0.03	2.3E-13	-0.08	++	-
M6326_1.02	KLF8	HepG2	0.03	3.7E-13	0.21	-	-
M6224_1.02	ETV7	HepG2	0.03	5.1E-13	0.41	-	5
YY2_full_1	YY2	HepG2	0.03	6.6E-13	0.37	+	5
M6271_1.02	HES1	HepG2	0.03	1.3E-12	0.19	+	-
V_STAT1_01	STAT1	HepG2	0.03	1.7E-12	0.37	+	2
V_SPZ1_01	SPZ1	HepG2	0.03	1.7E-12	0.19	-	-
M5491_1.02	GLIS2	HepG2	0.03	2.3E-12	-0.70	-	1
M6161_1.02	BHLHE41	HepG2	0.03	3.0E-12	0.17	-	-
M1934_1.02	ESR1	HepG2	0.03	3.6E-12	0.32	+	2
M6181_1.02	CREM	HepG2	0.03	4.8E-12	0.53	+	2
M6549_1.02	ZIC2	HepG2	0.03	5.0E-12	-0.08	-	-
M6339_1.02	MECP2	HepG2	0.03	5.7E-12	0.04	+	-
M2303_1.02	NR2C2	HepG2	0.03	7.0E-12	0.89	++	2
M4527_1.02	SMARCC2	HepG2	0.03	1.8E-11	0.05	+	-
BATF3_DBD	BATF3	HepG2	0.03	1.9E-11	0.60	+++	-
M6323_1.02	KLF3	HepG2	0.03	2.0E-11	0.28	++	2
M6270_1.02	NHLH1	HepG2	0.03	3.0E-11	0.45	-	3
M6212_1.02	EPAS1	HepG2	0.03	3.3E-11	1.79	++	4
M0422_1.02	ZIC5	HepG2	0.03	5.0E-11	-0.83	-	1
YY1_full	YY1	HepG2	0.03	8.9E-11	0.28	+	-
M0969_1.02	LHX8	HepG2	0.03	1.3E-10	-0.67	+	2
M0603_1.02	CXXC1	HepG2	0.03	1.8E-10	0.23	+	-
TFEC_DBD	TFEC	HepG2	0.03	1.8E-10	0.90	++	2
V_VDR_Q3	VDR	HepG2	0.03	1.9E-10	0.54	-	1
M1917_1.02	USF1	HepG2	0.03	2.8E-10	0.72	++	2
M6330_1.02	MAFA	HepG2	0.03	4.7E-10	0.47	-	2
M1968_1.02	EBF1	HepG2	0.03	5.7E-10	-0.04	-	-
V_DEAF1_02	DEAF1	HepG2	0.03	6.8E-10	0.15	+	-
M6518_1.02	TFEB	HepG2	0.03	1.9E-09	0.74	+++	2
M6523_1.02	THRB	HepG2	0.03	2.0E-09	0.54	+	2
M6360_1.02	NFE2L2	HepG2	0.03	2.6E-09	0.32	++	2
M6155_1.02	ATF6	HepG2	0.03	6.2E-09	0.35	+	-
Creb3l2_DBD_2	CREB3L2	HepG2	0.03	6.5E-09	0.71	++	-
V_EBOX_Q6_01	TCF3	HepG2	0.03	7.6E-09	0.57	+++	2
ARNTL_DBD	ARNTL	HepG2	0.03	9.9E-09	0.73	++	2
M6131_1.02	TFCP2L1	HepG2	0.03	2.9E-08	0.19	-	-
TFE3_DBD	TFE3	HepG2	0.03	4.1E-08	1.07	+++	2
M6207_1.02	ELK1	Karpas	0.14	0.0E+00	0.79	++	3

M0714_1.02	ETS1	Karpas	0.13	0.0E+00	0.76	++	3
M4522_1.02	ELK4	Karpas	0.13	0.0E+00	0.94	++	3
M2275_1.02	ELF1	Karpas	0.13	0.0E+00	1.10	++	3
ELK3_DBD	ELK3	Karpas	0.13	0.0E+00	0.67	++	3
M4462_1.02	GABPA	Karpas	0.13	0.0E+00	0.72	++	3
Elf5_DBD	ELF5	Karpas	0.13	0.0E+00	0.99	++	3
ERG_full_1	ERG	Karpas	0.13	0.0E+00	0.87	++	3
ETV6_full_2	ETV6	Karpas	0.13	0.0E+00	0.92	+	3
FEV_DBD	FEV	Karpas	0.12	0.0E+00	0.86	++	3
SP1_DBD	SP1	Karpas	0.12	0.0E+00	1.35	++	4
FLI1_full_1	FLI1	Karpas	0.12	0.0E+00	0.78	++	3
EHF_full	EHF	Karpas	0.12	0.0E+00	0.90	++	3
ELF4_full	ELF4	Karpas	0.12	0.0E+00	0.92	++	3
ETV4_DBD	ETV4	Karpas	0.12	0.0E+00	0.66	++	3
M5420_1.02	ETV1	Karpas	0.12	0.0E+00	0.58	++	3
M5209_1.02	SP5	Karpas	0.11	0.0E+00	1.55	++	2
SP4_full	SP4	Karpas	0.11	0.0E+00	1.20	+	1
M5398_1.02	ERF	Karpas	0.11	0.0E+00	0.91	++	3
ELF3_DBD	ELF3	Karpas	0.11	0.0E+00	0.93	++	3
M5422_1.02	ETV3	Karpas	0.11	0.0E+00	0.54	++	3
M2278_1.02	FOS	Karpas	0.11	0.0E+00	1.70	++	2
V_AP1_Q4_01	AP1	Karpas	0.11	0.0E+00	1.49	++	2
M2292_1.02	JUND	Karpas	0.11	0.0E+00	1.77	++	2
M4526_1.02	SMARCC1	Karpas	0.11	0.0E+00	1.67	++	2
M4623_1.02	JUNB	Karpas	0.11	0.0E+00	1.85	++	2
M2391_1.02	KLF5	Karpas	0.11	0.0E+00	1.25	++	2
M6221_1.02	ETS2	Karpas	0.10	0.0E+00	0.92	+	3
M4619_1.02	FOSL1	Karpas	0.10	0.0E+00	1.63	+++	2
M4565_1.02	FOSL2	Karpas	0.10	0.0E+00	1.66	+++	2
ETV5_DBD	ETV5	Karpas	0.10	0.0E+00	0.41	++	-
M0443_1.02	KLF12	Karpas	0.10	0.0E+00	1.56	++	2
M6324_1.02	KLF4	Karpas	0.10	0.0E+00	1.23	++	2
JDP2_full_1	JDP2	Karpas	0.10	0.0E+00	1.77	+++	2
V_NFY_01	NFY	Karpas	0.10	0.0E+00	0.07	++	2
M0405_1.02	KLF7	Karpas	0.10	0.0E+00	1.55	++	1
ETV2_DBD	ETV2	Karpas	0.10	0.0E+00	1.03	++	2
M1871_1.02	KLF2	Karpas	0.10	0.0E+00	1.23	++	2
KLF14_DBD	KLF14	Karpas	0.10	0.0E+00	1.08	++	2
SP3_DBD	SP3	Karpas	0.10	0.0E+00	1.25	++	4
M2314_1.02	SP2	Karpas	0.10	0.0E+00	1.27	++	1
KLF16_DBD	KLF16	Karpas	0.10	0.0E+00	0.93	++	1
V_BACH2_01	BACH2	Karpas	0.09	0.0E+00	1.32	+++	2
M6228_1.02	FOSB	Karpas	0.09	0.0E+00	1.20	++	2
M6204_1.02	ELF2	Karpas	0.09	0.0E+00	0.89	+	3
M2301_1.02	NFYB	Karpas	0.09	0.0E+00	0.02	+	2
SPIB_DBD	SPIB	Karpas	0.09	0.0E+00	1.02	+	3

NFE2_DBD	NFE2	Karpas	0.09	0.0E+00	1.47	+++	2
Spic_DBD	SPIC	Karpas	0.08	0.0E+00	0.92	+	3
M4473_1.02	PBX3	Karpas	0.08	0.0E+00	-0.12	++	2
V_MAF_Q6_01	MAF	Karpas	0.08	0.0E+00	0.90	++	2
SPI1_full	SPI1	Karpas	0.08	0.0E+00	0.93	+	3
SP8_DBD	SP8	Karpas	0.08	0.0E+00	0.92	++	1
GATA3_full	GATA3	Karpas	0.08	0.0E+00	2.22	-	3
M6174_1.02	CEBPZ	Karpas	0.08	0.0E+00	0.02	++	-
M6240_1.02	FOXI1	Karpas	0.08	0.0E+00	-0.16	++	-
V_GATA_Q6	GATA	Karpas	0.08	0.0E+00	2.19	-	3
GATA4_DBD	GATA4	Karpas	0.08	0.0E+00	2.25	-	3
M1868_1.02	GATA2	Karpas	0.07	0.0E+00	2.06	+++	3
M6360_1.02	NFE2L2	Karpas	0.07	0.0E+00	1.13	+++	2
M6373_1.02	NFYC	Karpas	0.07	0.0E+00	0.01	++	1
M6123_1.02	ZNF281	Karpas	0.07	0.0E+00	1.17	+	4
M4600_1.02	GATA1	Karpas	0.07	0.0E+00	2.18	+++	3
V_TEL2_Q6	ETV7	Karpas	0.07	0.0E+00	0.67	++	5
M6552_1.02	ZNF148	Karpas	0.07	0.0E+00	1.10	+	6
V_ETF_Q6	TEAD2	Karpas	0.07	0.0E+00	1.41	+	1
M6336_1.02	MAZ	Karpas	0.07	0.0E+00	1.24	++	4
V_CACD_01	CACD	Karpas	0.07	0.0E+00	1.35	++	1
GATA5_DBD	GATA5	Karpas	0.07	0.0E+00	2.11	+++	3
M4604_1.02	ZNF263	Karpas	0.07	0.0E+00	1.22	+	4
M1890_1.02	NFYA	Karpas	0.06	0.0E+00	-0.11	+	2
V_CKROX_Q2	ZBTB7B	Karpas	0.06	0.0E+00	0.85	++	4
SPDEF_full_1	SPDEF	Karpas	0.06	0.0E+00	0.51	++	-
M6325_1.02	KLF6	Karpas	0.06	0.0E+00	0.63	+	1
M6535_1.02	WT1	Karpas	0.06	0.0E+00	1.24	++	4
M4453_1.02	BCL11A	Karpas	0.06	0.0E+00	0.84	-	3
M0300_1.02	ATF2	Karpas	0.06	0.0E+00	1.05	++	1
V_STAT_Q6	ZNF143	Karpas	0.06	0.0E+00	0.40	+	2
NRF1_full	NRF1	Karpas	0.06	0.0E+00	0.91	+	4
V_KROX_Q6	EGR1	Karpas	0.06	0.0E+00	1.01	++	4
M1915_1.02	ZNF76	Karpas	0.06	0.0E+00	0.55	+	2
KLF13_full	KLF13	Karpas	0.05	0.0E+00	0.42	+	2
M6199_1.02	EGR2	Karpas	0.05	0.0E+00	0.90	+	1
M2289_1.02	JUN	Karpas	0.05	0.0E+00	0.69	++	2
M4572_1.02	MAFF	Karpas	0.05	0.0E+00	0.70	++	2
ZNF740_full	ZNF740	Karpas	0.05	0.0E+00	0.14	+	-
V_CREBATF_Q6	CREB1	Karpas	0.05	0.0E+00	0.81	++	-
M4463_1.02	IRF4	Karpas	0.05	0.0E+00	0.52	-	2
Creb5_DBD	CREB5	Karpas	0.05	0.0E+00	1.07	++	2
M6442_1.02	PURA	Karpas	0.05	0.0E+00	1.13	+	4
M1955_1.02	STAT1	Karpas	0.05	0.0E+00	0.86	++	2
M4452_1.02	BATF	Karpas	0.05	0.0E+00	0.48	-	2
M4527_1.02	SMARCC2	Karpas	0.05	0.0E+00	0.18	+	-

M5293_1.02	ATF7	Karpas	0.05	0.0E+00	0.82	++	-
M6258_1.02	GATA6	Karpas	0.05	0.0E+00	1.91	-	-
M6152_1.02	ATF1	Karpas	0.05	0.0E+00	1.05	++	2
M6553_1.02	ZNF219	Karpas	0.05	0.0E+00	0.49	-	1
M6201_1.02	EGR4	Karpas	0.05	0.0E+00	0.92	-	4
CREB3_full_1	CREB3	Karpas	0.05	0.0E+00	0.89	++	2
M6491_1.02	STAT5A	Karpas	0.05	0.0E+00	1.18	++	2
M2296_1.02	MAFK	Karpas	0.05	0.0E+00	0.49	+	2
M6492_1.02	STAT5B	Karpas	0.05	0.0E+00	0.94	++	2
XBP1_DBD_1	XBP1	Karpas	0.05	0.0E+00	1.16	++	2
M6322_1.02	KLF1	Karpas	0.05	0.0E+00	0.91	++	1
M0608_1.02	MLL	Karpas	0.05	0.0E+00	0.51	+	-
M6333_1.02	MAFG	Karpas	0.05	0.0E+00	0.56	++	2
M6359_1.02	NFE2L1	Karpas	0.05	0.0E+00	0.56	++	2
V_NKX25_Q5	NKX2-5	Karpas	0.04	0.0E+00	0.06	-	-
M4478_1.02	STAT3	Karpas	0.04	0.0E+00	0.60	++	2
M2273_1.02	E2F6	Karpas	0.04	0.0E+00	0.90	+	4
M1919_1.02	YY1	Karpas	0.04	0.0E+00	0.38	+	-
BATF3_DBD	BATF3	Karpas	0.04	0.0E+00	0.75	++	-
ATF4_DBD	ATF4	Karpas	0.04	0.0E+00	0.21	+++	-
EGR3_DBD	EGR3	Karpas	0.04	0.0E+00	1.07	+	1
M6181_1.02	CREM	Karpas	0.04	0.0E+00	0.64	++	2
M5932_1.02	TFEC	Karpas	0.04	0.0E+00	0.73	+++	2
USF1_DBD	USF1	Karpas	0.04	0.0E+00	1.19	+++	2
M2307_1.02	PRDM1	Karpas	0.04	6.7E-16	0.67	+	6
M6321_1.02	KLF15	Karpas	0.04	2.0E-15	0.94	+	4
M6154_1.02	ATF5	Karpas	0.04	3.6E-15	0.55	+	5
Srebf1_DBD	SREBF1	Karpas	0.04	1.3E-14	0.82	+++	2
M6537_1.02	YBX1	Karpas	0.04	2.1E-14	-0.24	+++	1
M4692_1.02	SIX5	Karpas	0.04	2.7E-14	0.32	+	-
M6496_1.02	STAT4	Karpas	0.04	3.3E-14	0.72	+	2
M6313_1.02	IRF8	Karpas	0.04	4.5E-14	0.80	+	2
YY2_full_1	YY2	Karpas	0.04	5.7E-14	0.57	+	5
TFEB_full	TFEB	Karpas	0.03	1.4E-13	0.94	+++	2
M0969_1.02	LHX8	Karpas	0.03	2.4E-13	-0.54	+	2
M4536_1.02	E2F1	Karpas	0.03	9.1E-13	1.10	+	1
M1581_1.02	CIC	Karpas	0.03	1.4E-12	0.50	+	2
SREBF2_DBD	SREBF2	Karpas	0.03	7.0E-12	0.66	+++	2
V_EVI1_03	MECOM	Karpas	0.03	1.4E-11	1.21	+	5
M1882_1.02	IRF1	Karpas	0.03	2.0E-11	1.49	+	6
M4680_1.02	BACH1	Karpas	0.03	2.3E-11	0.53	+++	2
M6517_1.02	TFE3	Karpas	0.03	6.2E-11	0.51	++	2
M6309_1.02	IRF3	Karpas	0.03	1.1E-10	1.00	++	4
V_P300_01	EP300	Karpas	0.03	1.5E-10	0.39	+	-
M4545_1.02	ZNF683	Karpas	0.03	3.7E-10	0.47	+	6
Mafb_DBD_1	MAFB	Karpas	0.03	5.4E-10	0.41	+	2

M4640_1.02	ZBTB7A	Karpas	0.03	6.1E-10	0.45	+	1
M4451_1.02	ATF3	Karpas	0.03	1.5E-09	0.91	+++	2
M5689_1.02	NRL	Karpas	0.03	1.7E-09	0.40	+	2
M6197_1.02	E4F1	Karpas	0.03	4.4E-09	0.75	++	-
M6159_1.02	BCL6	Karpas	0.03	6.7E-09	0.23	+	-
M4537_1.02	E2F4	Karpas	0.03	1.3E-08	0.91	++	1
V_PAX4_03	PAX4	Karpas	0.03	2.6E-08	0.82	-	1
M6311_1.02	IRF5	Karpas	0.03	3.6E-08	1.14	-	6
M6548_1.02	ZIC1	Karpas	0.03	5.0E-08	0.33	+	1
M0422_1.02	ZIC5	Karpas	-0.03	3.2E-08	-0.86	-	1
M6422_1.02	PLAGL1	Karpas	-0.03	9.7E-09	-0.49	+	1
M6468_1.02	SNAI1	Karpas	-0.03	1.8E-10	0.10	++	-
M5571_1.02	ID4	Karpas	-0.03	9.4E-12	0.03	++	3
MESP1_DBD	MESP1	Karpas	-0.03	9.4E-12	0.05	++	-
M5512_1.02	HIC2	Karpas	-0.03	3.3E-12	-0.42	++	-
M5209_1.02	SP5	Jurkat_2	0.07	0.0E+00	1.24	++	2
M6207_1.02	ELK1	Jurkat_2	0.07	0.0E+00	1.58	+++	3
V_SP1_Q6	SP1	Jurkat_2	0.07	0.0E+00	1.12	++	4
M4462_1.02	GABPA	Jurkat_2	0.07	0.0E+00	1.18	+++	3
SP4_full	SP4	Jurkat_2	0.07	0.0E+00	0.98	++	1
ELK3_DBD	ELK3	Jurkat_2	0.07	0.0E+00	1.34	+++	3
ELK4_DBD	ELK4	Jurkat_2	0.07	0.0E+00	1.42	+++	3
V_NFY_01	NFY	Jurkat_2	0.07	0.0E+00	-0.03	++	2
M0714_1.02	ETS1	Jurkat_2	0.06	0.0E+00	1.44	+++	3
KLF14_DBD	KLF14	Jurkat_2	0.06	0.0E+00	0.87	++	2
ELF4_full	ELF4	Jurkat_2	0.06	0.0E+00	1.17	++	3
ELF1_DBD	ELF1	Jurkat_2	0.06	0.0E+00	1.23	+++	3
Elf5_DBD	ELF5	Jurkat_2	0.06	0.0E+00	1.37	+++	3
M5422_1.02	ETV3	Jurkat_2	0.06	0.0E+00	1.17	+++	3
FLI1_full_1	FLI1	Jurkat_2	0.06	0.0E+00	1.54	+++	3
FEV_DBD	FEV	Jurkat_2	0.06	0.0E+00	1.56	+++	3
ETV4_DBD	ETV4	Jurkat_2	0.06	0.0E+00	1.27	+++	3
ERG_full_1	ERG	Jurkat_2	0.06	0.0E+00	1.75	+++	3
M0443_1.02	KLF12	Jurkat_2	0.06	0.0E+00	1.21	++	2
M5420_1.02	ETV1	Jurkat_2	0.06	0.0E+00	1.25	+++	3
EHF_full	EHF	Jurkat_2	0.06	0.0E+00	1.16	+++	3
ETV6_full_2	ETV6	Jurkat_2	0.06	0.0E+00	1.27	+++	3
M0405_1.02	KLF7	Jurkat_2	0.06	0.0E+00	1.27	++	1
ETV5_DBD	ETV5	Jurkat_2	0.06	0.0E+00	0.93	+++	-
M2301_1.02	NFYB	Jurkat_2	0.06	0.0E+00	-0.11	+	2
M2314_1.02	SP2	Jurkat_2	0.06	0.0E+00	1.07	++	1
V_ETF_Q6	TEAD2	Jurkat_2	0.06	0.0E+00	1.17	+	1
KLF16_DBD	KLF16	Jurkat_2	0.06	0.0E+00	0.84	++	1
M6174_1.02	CEBPZ	Jurkat_2	0.06	0.0E+00	0.01	++	-
M1871_1.02	KLF2	Jurkat_2	0.06	0.0E+00	1.00	+++	2
M2391_1.02	KLF5	Jurkat_2	0.06	0.0E+00	1.15	+++	2

M5398_1.02	ERF	Jurkat_2	0.06	0.0E+00	1.62	+++	3
M6482_1.02	SP3	Jurkat_2	0.06	0.0E+00	1.12	++	4
M4473_1.02	PBX3	Jurkat_2	0.05	0.0E+00	-0.11	++	2
M6324_1.02	KLF4	Jurkat_2	0.05	0.0E+00	1.05	+++	2
ELF3_full	ELF3	Jurkat_2	0.05	0.0E+00	1.20	++	3
M6373_1.02	NFYC	Jurkat_2	0.05	0.0E+00	-0.15	++	1
M5856_1.02	SP8	Jurkat_2	0.05	0.0E+00	0.86	++	1
M6221_1.02	ETS2	Jurkat_2	0.05	0.0E+00	1.37	+	3
M4459_1.02	EGR1	Jurkat_2	0.05	0.0E+00	1.11	+	4
M6240_1.02	FOXI1	Jurkat_2	0.05	0.0E+00	-0.20	+	-
M1890_1.02	NFYA	Jurkat_2	0.05	0.0E+00	-0.12	+	2
M6325_1.02	KLF6	Jurkat_2	0.05	0.0E+00	0.76	+	1
M6552_1.02	ZNF148	Jurkat_2	0.04	0.0E+00	1.08	+	6
M6535_1.02	WT1	Jurkat_2	0.04	0.0E+00	1.06	++	4
M4536_1.02	E2F1	Jurkat_2	0.04	0.0E+00	0.82	+	1
M6539_1.02	ZBTB7B	Jurkat_2	0.04	0.0E+00	1.10	++	4
M2305_1.02	NRF1	Jurkat_2	0.04	0.0E+00	0.51	+	4
ETV2_DBD	ETV2	Jurkat_2	0.04	0.0E+00	1.77	+++	2
V_MAZ_Q6	MAZ	Jurkat_2	0.04	0.0E+00	0.86	+++	4
M0608_1.02	MLL	Jurkat_2	0.04	0.0E+00	0.60	++	-
M4604_1.02	ZNF263	Jurkat_2	0.04	0.0E+00	1.27	+	4
M0300_1.02	ATF2	Jurkat_2	0.04	0.0E+00	0.95	++	1
M6204_1.02	ELF2	Jurkat_2	0.04	0.0E+00	1.25	+	3
M6201_1.02	EGR4	Jurkat_2	0.04	2.2E-16	0.95	-	4
KLF13_full	KLF13	Jurkat_2	0.04	4.4E-16	0.19	+	2
M6123_1.02	ZNF281	Jurkat_2	0.04	4.4E-16	1.19	+	4
V_CACD_01	CACD	Jurkat_2	0.04	1.1E-15	1.22	++	1
M4640_1.02	ZBTB7A	Jurkat_2	0.04	1.8E-15	0.48	+	1
M6199_1.02	EGR2	Jurkat_2	0.04	2.7E-15	0.80	+	1
ZNF740_full	ZNF740	Jurkat_2	0.04	3.6E-15	0.08	+	-
SPDEF_full_1	SPDEF	Jurkat_2	0.04	6.2E-15	1.23	++	-
SPIB_DBD	SPIB	Jurkat_2	0.04	6.4E-15	0.94	+	3
M6553_1.02	ZNF219	Jurkat_2	0.04	1.5E-14	0.39	+	1
V_CREBATF_Q6	CREB1	Jurkat_2	0.04	1.6E-14	0.92	++	-
Spic_DBD	SPIC	Jurkat_2	0.04	1.8E-14	0.98	+	3
M6224_1.02	ETV7	Jurkat_2	0.03	6.9E-13	1.06	-	5
SPI1_full	SPI1	Jurkat_2	0.03	2.2E-12	0.81	+	3
YY2_full_1	YY2	Jurkat_2	0.03	7.1E-12	0.57	++	5
M6442_1.02	PURA	Jurkat_2	0.03	1.5E-11	1.17	+	4
Creb5_DBD	CREB5	Jurkat_2	0.03	4.1E-11	1.04	++	2
V_STAF_02	ZNF143	Jurkat_2	0.03	7.3E-11	0.21	+	2
M0609_1.02	DNMT1	Jurkat_2	0.03	9.9E-11	0.86	+++	1
V_YY1_Q6	YY1	Jurkat_2	0.03	3.1E-10	0.15	++	-
Jdp2_DBD_2	JDP2	Jurkat_2	0.03	5.3E-10	0.98	++	2
EGR3_DBD	EGR3	Jurkat_2	0.03	1.0E-09	0.85	+	1
XBP1_DBD_1	XBP1	Jurkat_2	0.03	1.0E-09	0.90	++	2

M6321_1.02	KLF15	Jurkat_2	0.03	1.0E-09	0.93	+	4
M6547_1.02	ZFX	Jurkat_2	0.03	1.7E-09	0.43	+	1
CREB3_full_1	CREB3	Jurkat_2	0.03	2.0E-09	0.57	++	2
M5932_1.02	TFEC	Jurkat_2	0.03	2.8E-09	0.69	+++	2
M6152_1.02	ATF1	Jurkat_2	0.03	5.2E-09	0.74	++	2
TFEB_full	TFEB	Jurkat_2	0.03	5.6E-09	0.86	+++	2
M2323_1.02	ZBTB33	Jurkat_2	0.03	7.8E-09	-0.17	++	-
M4527_1.02	SMARCC2	Jurkat_2	0.03	8.3E-09	0.16	++	-
USF1_DBD	USF1	Jurkat_2	0.03	1.1E-08	1.07	+++	2
M4680_1.02	BACH1	Jurkat_2	0.03	2.3E-08	0.48	+++	2
ATF7_DBD	ATF7	Jurkat_2	0.03	3.7E-08	1.03	++	-
M6197_1.02	E4F1	Jurkat_2	0.03	3.9E-08	0.62	++	-
M6146_1.02	TFAP2D	Jurkat_2	0.03	4.0E-08	0.28	-	1
M6537_1.02	YBX1	Jurkat_2	0.03	4.1E-08	-0.36	+++	1
M6337_1.02	MBD2	Jurkat_2	0.03	4.9E-08	-0.01	+	-

REFERENCES

1. Grossman SR, *et al.* (2018) Positional specificity of different transcription factor classes within enhancers. *Proc Natl Acad Sci U S A* 115(30):E7222-E7230.

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

3. Levo M & Segal E (2014) In pursuit of design principles of regulatory sequences. *Nature Publishing Group* 15(7):453-468.

4. Struhl K (1991) Mechanisms for diversity in gene expression patterns. *Neuron* 7(2):177-181.

5. Small S, Blair A, & Levine M (1992) Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J* 11(11):4047-4057.

6. Swanson CI, Evans NC, & Barolo S (2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell* 18(3):359-370.

7. Farley EK, *et al.* (2015) Suboptimization of developmental enhancers. *Science* 350(6258):325-328.

8. Halfon MS, *et al.* (2000) Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell* 103(1):63-74.

9. Thanos D & Maniatis T (1995) Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* 83(7):1091-1100.

10. Weingarten-Gabbay S & Segal E (2014) The grammar of transcriptional regulation. *Hum Genet* 133(6):701-711.

11. Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10(9):605-616.

12. Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. *Dev Cell* 21(4):611-626.

13. Lambert SA, *et al.* (2018) The Human Transcription Factors. *Cell* 175(2):598-599.

14. White MA, Myers CA, Corbo JC, & Cohen BA (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America* 110(29):11952-11957.

15. Melnikov A, *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*:1-9.

16. Patwardhan RP, *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology* 30(3):265-270.

17. Grossman SR, *et al.* (2017) Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences of the United States of America* 114(7):E1291-E1300.

18. Roadmap Epigenomics C, *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317-330.

19. Kheradpour P, *et al.* (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research* 23(5):800-811.

20. Melnikov A, *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30(3):271-277.

21. Ulirsch JC, *et al.* (2014) Altered chromatin occupancy of master regulators underlies evolutionary divergence in the transcriptional landscape of erythroid differentiation. *PLoS Genetics* 10(12):e1004890.

22. Stampfel G, *et al.* (2015) Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* 528(7580):147-151.

23. Deaton AM & Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev* 25(10):1010-1022.

24. Curina A, *et al.* (2017) High constitutive activity of a broad panel of housekeeping and tissue-specific cis-regulatory elements depends on a subset of ETS proteins. *Genes Dev* 31(4):399-412.

25. Kaczynski J, Cook T, & Urrutia R (2003) Sp1- and Kruppel-like transcription factors. *Genome Biol* 4(2):206.

26. Scarpulla RC (2002) Transcriptional activators and coactivators in the nuclear control of mitochondrial function in mammalian cells. *Gene* 286(1):81-89.

27. Garber M, *et al.* (2012) A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular Cell* 47(5):810-822.

28. Sieweke MH, Tekotte H, Frampton J, & Graf T (1996) MafB is an interaction partner and repressor of Ets-1 that inhibits erythroid differentiation. *Cell* 85(1):49-60.

29. Hakimi MA, *et al.* (2002) A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes. *Proc Natl Acad Sci U S A* 99(11):7420-7425.

30. You A, Tong JK, Grozinger CM, & Schreiber SL (2001) CoREST is an integral component of the CoREST- human histone deacetylase complex. *Proc Natl Acad Sci U S A* 98(4):1454-1458.

31. Shi YJ, *et al.* (2005) Regulation of LSD1 histone demethylase activity by its associated factors. *Mol Cell* 19(6):857-864.

32. Gyory I, Wu J, Fejer G, Seto E, & Wright KL (2004) PRDI-BF1 recruits the histone H3 methyltransferase G9a in transcriptional silencing. *Nat Immunol* 5(3):299-308.

33. Su ST, *et al.* (2009) Involvement of histone demethylase LSD1 in Blimp-1-mediated gene repression during plasma cell differentiation. *Mol Cell Biol* 29(6):1421-1431.

34. Bikoff EK, Morgan MA, & Robertson EJ (2009) An expanding job description for Blimp-1/PRDM1. *Curr Opin Genet Dev* 19(4):379-385.

35. Zarkower D (2013) DMRT genes in vertebrate gametogenesis. *Curr Top Dev Biol* 102:327-356.

36. Cusanovich DA, Pavlovic B, Pritchard JK, & Gilad Y (2014) The functional consequences of variation in transcription factor binding. *PLoS Genet* 10(3):e1004226.

37. Spitz F & Furlong EEM (2012) Transcription factors: from enhancer binding to developmental control. *Nature Publishing Group* 13(9):613-626.

38. Biddie SC, *et al.* (2011) Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. *Molecular Cell* 43(1):145-155.

39. Vierbuchen T, *et al.* (2017) AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Mol Cell* 68(6):1067-1082 e1012.

40. Ferreira R, Ohneda K, Yamamoto M, & Philipsen S (2005) GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol Cell Biol* 25(4):1215-1227.

41. Zaret KS (2002) Regulatory phases of early liver development: paradigms of organogenesis. *Nat Rev Genet* 3(7):499-512.

42. Zaret KS & Carroll JS (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes & Development* 25(21):2227-2241.

43. Panne D (2008) The enhanceosome. *Curr Opin Struct Biol* 18(2):236-242.

44. Arnosti DN & Kulkarni MM (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry* 94(5):890-898.

45. Hare EE, Peterson BK, Iyer VN, Meier R, & Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genetics* 4(6):e1000106.

46. Kulkarni MM & Arnosti DN (2003) Information display by transcriptional enhancers. *Development* 130(26):6569-6575.

47. Evans NC, Swanson CI, & Barolo S (2012) Sparkling insights into enhancer structure, function, and evolution. *Current topics in developmental biology* 98:97-120.

48. Ludwig MZ, Bergman C, Patel NH, & Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403(6769):564-567.

49. Rada-Iglesias A, *et al.* (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333):279-283.

50. May D, *et al.* (2011) Large-scale discovery of enhancers from human heart tissue. *Nat Genet* 44(1):89-93.

51. Corces MR, *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics* 48(10):1193-1203.

52. Buenrostro JD, *et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523(7561):486-490.

53. Buenrostro JD, Wu B, Chang HY, & Greenleaf WJ (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* 109:21 29 21-29.

54. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.

55. Zhang Y, *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9(9):R137.

56. Tewhey R, *et al.* (2016) Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* 165(6):1519-1529.

57. Matys V, *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34(Database issue):D108-110.

58. Mathelier A, *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42(Database issue):D142-147.

59. Weirauch MT, *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158(6):1431-1443.

60. McLeay RC & Bailey TL (2010) Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* 11:165.

61. Grant CE, Bailey TL, & Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017-1018.

Chapter 5

Conclusion and Future Directions

ABSTRACT

Here we synthesize the collective work of this thesis and propose future avenues of research towards deciphering the transcriptional regulatory code and understanding combinatorial enhancer control by TFs. We review a model for enhancer architecture that integrates elements of both the enhanceosome and billboard models, and extends the current models to include new type of positional constraint. We discuss various specialized functions that might distinguish different types of TFs, and some approaches to detecting these TF classes. Finally, we propose a general regulatory code at the level of TF function, whereby different classes of TFs contribute different regulatory activities to produce functional enhancers. Together, these insights motivate further studies to identify regulatory activities associated with different TF classes, dissect the role of TF positions within enhancers, and understand the connection between binding site positioning and TF function.

SECTION I. Synthesis: Insights into regulatory sequence design and function

Combinatorial TF regulation is encoded in regulatory sequences in the form of arrays of specific recognition sites for distinct combinations of transcription factors (TFs). Very few TFs can activate transcription on their own, instead functioning cooperatively with partner TFs bound to the same enhancers to specifically activate transcription. This combinatorial property of enhancers allows a relatively small set of TFs to implement varied expression outputs for 20,000 genes across thousands of cell types and conditions by acting in different groupings (1-3). Furthermore, it ensures specific activity for enhancers activated by the convergence of multiple signaling pathways, such as developmental and immune enhancers (4-8).

When I began this thesis, most of the knowledge about regulatory logic and the mechanisms of transcriptional activation came from functional case studies of various regulatory elements. These revealed that both the binding site composition (identity and number of different TF binding sites) and organization (position, order and orientation of TF binding sites) can affect the expression output (9). However, not every potential TF binding site in an enhancer is bound *in vivo* (10, 11), and for all but a few enhancers the functional TF binding sites have not been defined. Furthermore, the relationship between input composition and arrangement of TF binding sites and output gene expression has proven to be complex and many different relationships (or "computations") exist (1-3). Finally, the functions of most TFs in transcriptional control are unknown (12), making it hard to generalize across TFs or infer universal principles of regulatory grammar. In short, many questions remained about how combinatorial gene regulation is encoded in enhancers and promoters and how that information is integrated by TFs to determine transcriptional output.

The work presented in this thesis provides insight into the general rules and mechanisms of regulatory sequence function. At the outset of this thesis, new high-throughput and quantitative technologies developed by our lab and others enabled the synthesis and simultaneous measurement of the activities of thousands of regulatory sequences (13-17). Using these methods and computational analysis, we studied TF binding sites in enhancers from numerous different cell types and regulatory systems, shedding light general principles of motif composition and organization in typical cellular regulatory elements. We find extensive synergy between TF binding sites, some with organizational constraints and some with flexible positioning. We demonstrate that different TFs bind at distinct positions within regulatory elements, suggesting a new type of architectural constraint in enhancers. Importantly, our analysis of both TF organization and cooperativity revealed distinctive patterns that separates TFs into potential functional classes.

Together, our results suggest a structure of the regulatory code at the level of TF function and generate new hypotheses about regiospecific binding patterns and functions of TF classes within enhancers. Here, we discuss these insights and propose future directions of research to test these emerging models.

Properties of functional TF binding sites

To begin our investigation of the regulatory grammar of enhancers, we focused on the smallest functional unit: TF binding sites. Before this work began, the proliferation of genome-wide maps of TF occupancies had recently revealed that nearly all mammalian TFs bind to only a minority of their potential genomic binding sites *in*

vivo, and a large fraction of these binding events do not appear to change the expression of any nearby gene (10, 11). To determine the sequence and chromatin features that determine whether a potential binding site is bound *in vivo* and its contribution to transcription, we systematically varied the core motif, the surrounding sequence and the chromatin context of potential genomic binding sites. Our studies of a prototypical TF (PPARγ in mouse adipocytes) revealed several notable features that distinguish functional motif sites that are bound *in vivo* and contribute to transcriptional regulation (Chapter 2), forming the basis for our subsequent work.

Genomic motif sites that are bound in vivo

Genomic binding specificity could arise at several levels: the core motif site (e.g., from latent specificity brought out by interactions with cofactors (18, 19)), the sequence immediately flanking the site (e.g. from cooperative binding with partner TFs (20)), or the larger chromatin context (e.g. from previously established differences in DNA accessibility (21-24)). To distinguish these three possibilities, we systematically varied the core motif, flanking sequence and chromatin context of genomic motif sites for a model TF (PPARγ in mouse adipocytes; Chapter 2). We found that virtually all genomic motif sites can be bound by PPARγ when placed in an episomal context, regardless of their sequence context, suggesting PPARγ binding in the genome is largely determined by the preexisting DNA accessibility (as opposed to latent specificity or cooperative binding interactions with partner TFs). Genome-wide PPARγ occupancy correlates strongly with both the qualitative and quantitative DNA accessibility, corroborating this theory.

More generally, our analysis shows that TF binding and quantitative DNA accessibility are closely linked for many of 61 other human and mouse TFs profiled in the ENCODE project, with the notable exception of known pioneer factors. This observation is consistent with a proposed hierarchical model of TF binding (25), whereby a set of pioneer factors initially establish the accessible chromatin landscape, which then guides the binding specificity of "settler" TFs (such as PPARγ) that cannot bind sites occluded by nucleosomes. Direct evidence for this model was recently provided by a SELEX-based study, which found that the majority of TFs only bind weakly to motif sites within nucleosomes, while a subset can recognize and robustly bind occluded motif sites (26).

TF binding sites that contribute to enhancer activity

In contrast to TF binding, the transcriptional output driven by a motif site is largely determined by its sequence context. In particular, the enhancer activity of a bound genomic site depends on a set of co-regulatory TFs that bind nearby and cooperatively regulate transcription. In sequences without these neighboring TF binding sites, PPARγ binding does not result in transcription.

Because enhancer-bound TFs act so collaboratively to control transcription, identifying the network of active TFs in any enhancers of interest is essential to predict their expression output. We developed a simple paradigm to identify and validate active TFs and binding sites in sets of co-regulated enhancers by (1) measuring the activity of thousands of genomic regulatory elements using high-throughput reporter assays, (2) identifying all known motifs whose counts significantly correlate with enhancer activity, and (3) characterizing regulatory contributions of correlated motifs using perturbations.

We used this approach to identify the active TFs in PPARγ response elements (PPREs) in mouse adipocytes (Chapter 2) and cell type-restricted and ubiquitous enhancers from 6 human cell lines (Chapter 4). The size of these TF regulatory networks depends on how narrowly the system of enhancers is defined, ranging from ~30 TFs in the focused PPARγ response elements to ~200 in the more broadly defined cell type-restricted regulatory elements.

Importantly, we found that TF motifs associated with enhancer activity in functional assays can be approximated by highly enriched motifs in the active regulatory elements. Since chromatin states can be readily mapped, this approximation enables the prediction of active TFs and functional binding sites in a wide range of cellular contexts and conditions. We demonstrate the power of this prediction strategy in four ENCODE cell lines, where virtually all of the TF binding sites we predicted to be functional based on their enrichment in cell type-restricted enhancers directly contribute to enhancer activity (as assayed by mutational perturbation).

Architecture of motif sites in regulatory elements

The prevalence and nature of organizational constraints on motifs in regulatory sequences is an area of active investigation and great interest. Examples from the regulatory sequences that have extensively characterized span a wide range. At one extreme are enhanceosomes such as the IFNβ enhancer, which requires the binding of multiple TFs to overlapping sites. Enhanceosome activity depends on extensive protein-protein interactions between TFs, making them exquisitely sensitive to shifts or rearrangements of motifs (27). At the other extreme are so-called "billboard" enhancers, which contain independent motif sites that contribute additively to expression and can

be freely rearranged (28-30). Other examples lie between these two extremes, allowing rearrangement of some elements but not others (6, 31). Evolutionary analyses indicate that enhancer sequences diverge rapidly between species but often retain the ancestral function through compensatory TF binding site turnover (32-34), which would seem to support more flexible organizational principles.

Based on our studies, we propose a novel architectural model combining elements of enhanceosomes and billboard enhancers within an overarching organizational structure based on the position of binding sites within the NDR. The model provides a theoretical framework that can be translated to any cellular context, and makes concrete predictions that can be tested through comparison of enhancer conservation between species and functional studies.

Our proposed model was suggested by the following observations:

(i) **Both motif composition and arrangement are key determinants of enhancer activity.** We find that ~25% of the variance in expression levels driven by cellular enhancers can be explained using an additive model based solely on the identity and number of motif sites and nucleotide composition, and ~50% by a model that includes pairwise interaction terms to capture TF cooperativity (in PPREs and preliminary analysis of ENCODE regulatory elements). While motif predictions are not always accurate, these results imply that the arrangement and spacing of binding sites and higher-order interactions account for close to half the variance in enhancer activity.

(ii) **Some but not all synergistic interactions between pairs of TFs display a preferred spacing and orientation.** About a third of significant TF interactions we detected in cellular enhancers showed significant enrichment of *adjacent* co-occurrences, suggesting they require specific spacing, and half of the linked binding sites also had a preferred orientation. Several of the linked pairs are known to physically interact and/or cooperatively regulate gene expression through and orientation, suggesting the enriched configurations reflect functional protein-protein interactions between TFs.

(iii) **Different classes of TFs display distinct regiospecific binding patterns within NDR.** Empirical and inferred functional binding sites for 285 TFs across 47 cell types show six distinct positional preferences, some concentrated in the center, and others at specific positions or towards the edge of the NDR. TFs with similar binding patterns also share various structural and functional properties, such as binding stability, pioneering ability, and interactions with other TFs and cofactors. Furthermore, most of the interacting pairs we identified above involve two TFs that belong to the same positional class.

(iv) **Motif sites in preferred positions show increased binding and regulatory effects.** Across all classes, these motif sites were about twice as likely to be bound by TFs, and had 3-fold higher activity relative to motif sites in non-optimal positions. These results demonstrate that regiospecific binding constraints play an important role in determining enhancer activity.

Two approaches to test proposed architectural model

Comparisons of enhancers across species. Despite rapid turnover of TF binding sites across evolution, our model would predict that binding sites will tend to remain within their preferred subregions of NDR. This prediction could be easily tested across mammalian enhancers, in particular focusing on enhancers with low sequence conservation but conserved function, such as murine and human heart enhancers (34, 35).

Functional testing with dCas9. Selected TFs representing each regiospecific binding class could be fused to dCas9 and targeted to different positions within genomic enhancers. For each TF at each position, various regulatory activities could be measured, including DNA accessibility (ATAC-seq), chromatin modifications (ChIP-seq), 3D localization (3C/HiC), and gene regulation (RNA-seq). Ideally, these experiments would be performed in bulk samples for a small number of TFs and enhancers to detect subtle effects, as well as single cells for a larger set of TFs and enhancers to identify general trends.

Distinguishing functional classes of TFs elements

One of the key themes that emerged from this work is the idea of functionally distinct groups of TFs that contribute different regulatory activities in enhancers. The most concrete example of a specialized class of TFs is pioneer factors, which possess unique biochemical properties that allow them to bind motif sites within closed chromatin and facilitate the subsequent binding of additional TFs (36). However, transcription in metazoans is comprised of numerous distinct steps, each of which could be targeted by specialized groups of TFs (Chapter 1, Fig. X). Such classes could potentially serve as a key to decrypt regulatory sequence, allowing us to identify general formulas at the level of TF functions resulting in various expression outputs. Three approaches to functionally classify TFs emerged from this thesis and are discussed below.

TF classes based on activity patterns

Regulatory elements can be generally separated into those that are ubiquitously active across the majority of cell types and those whose activity is developmentally regulated and restricted to certain cell types. Similarly, we observed three groups of TFs that show different patterns of activity and motif enrichment across cell types and in the two kinds of regulatory elements. The first group comprises TFs that primarily activate ubiquitous regulatory elements and tend to bind in promoter regions. The second group includes TFs that mostly bind to distal enhancers with cell type-specific activity, but are expressed in most or all cell types, and the last group also bind to cell type-specific regulatory elements but shows cell type-restricted expression and activity. These functional TF classes determine the cell type specificity and kinetics of enhancers.

TF classes based on interactions with other TFs

Another approach to distinguishing TF functional classes emerged from our analysis of synergistic and antagonistic interactions between TFs in PPARγ enhancers. Although the set of TFs was too small to reach significance, we noticed that the TFs could be grouped by their interaction patterns into "equivalence" classes that had similar patterns of interactions with other TF classes and no intra-class interactions. This observation is consistent with the hypothesis that enhancer function requires members of different functional classes of TFs that play distinct roles in activating transcription, with different equivalence classes representing each functional role. Thus, a larger graph of TFs and interactions could be used to identify hypothesized TF classes for further functional characterization. This approach is similar to a recent study in *Drosophila* that classified TFs based on their ability to substitute for each other in diverse regulatory contexts (37).

TF classes based on position in enhancers

Finally, the regiospecific binding classes we identified bring together factors that have a number of similar functional properties, such as binding stability, interactions with other TFs and cofactors, cell-type specificity, and pioneering ability, suggesting they may represent functional classes of TFs. Consistent with this hypothesis, the position of motif sites appears to be related to their known functions—for example, localizing pioneer factors to the optimal positions to displace nucleosomes and targeting chromatin remodelers in close proximity to flanking nucleosome.

Testing TF functional classes

Generalizing interaction classes.

Our observations in PPREs suggested that identifying TF-TF interactions could shed light on functionally equivalent classes of TFs. To further investigate this idea, we designed two MPRA experiments to experimentally characterize interactions between active TFs in the six ENCODE studied in Chapter 4. We designed a library of synthetic enhancers consisting of binding sites from 20 TFs with the most highly enriched motifs in active regulatory elements patterned on to 20 different neutral templates. First, we placed 1-6 copies of each individual motif on to each neutral template in different configurations (Fig. 1A). This set allows us to model the independent effect of each TF and identify self-synergy or redundancy. Second, we placed pairs of TFs with 1-5 copies of each TF in the pair (Fig. 1B). Using this set and the estimates of the individual effects of the TFs, we will model quantitative interaction effects.

The second MPRA pool that will be used to characterize TF interactions is the cell type-specific mutagenesis pool described in Chapter 4. This library contains sequences from 2,500 active regulatory elements in each of the six cell types. For each of the 20 enriched TFs from the cognate cell type, we disrupted motif sites for each TF individually as well as in pairs. This approach complements the synthetic pool, allowing us to measure individual and interaction effects of each pair of TFs in native genomic enhancers. Using these two MPRA experiments, we will build a network of interactions between all the tested TFs.

From this network of TF-TF interactions, we will examine whether groups of TFs show interaction patterns that suggest functional equivalence. We will also compare the TF interactions across cell types. In the simplest case of TF functional classes, each TF has the same functional properties in each cell type so their interactions should be

consistent across different cell types. Finally, we will compare identified TF interaction classes and the positional classes. If both of these independent classification strategies identify similar classes of TFs, it would suggest the classes reflect true functional differences between TFs.

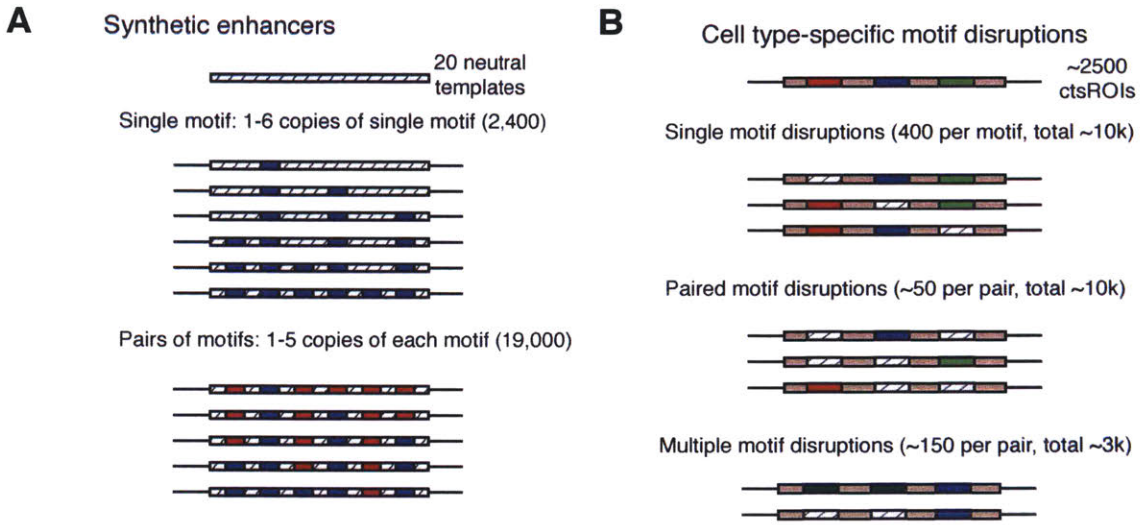


Figure 1. Schematic of MPRA pools to characterize TF-TF interactions. (A) Synthetic enhancers contain copies of each motif individually and each pair of motifs from the cognate cell type patterned on to 20 neutral templates. **(B)** Motif disruption pool contains 2,500 active regulatory element sequences with disruptions of each motif individually and each pair of motifs.

Conclusion

Recent advances in genomic technologies have sparked a renaissance in genomics and medicine. Having mapped the protein-coding genes in the genome, we are now beginning to decipher the mechanisms and encoding of gene regulation. These efforts have led to a growing understanding of how regulatory elements in the genome function, and revealed a complex transcriptional regulatory code based on combinatorial TF activity. While TFs are often classified as simply activators or

repressors, recent work, including this thesis, highlights the unique properties of different classes of TFs and suggests that they might contribute different functional roles to enhancer activity. This work also highlights how enhancer architecture involves both constraints in the positioning of TFs relative to each other as well as their overall positioning within the regulatory element. As we continue our exploration of TF functions, we will undoubtedly uncover more mechanisms guiding the positioning and synergy of TFs in transcriptional regulation. Ultimately, understanding the cis-regulatory code will enable us to not only understand the intricate biochemistry that drives the cellular basis of life, but also to manipulate the processes that interpret our genome sequence, yielding new understanding of and treatments for human disease.

REFERENCES

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

2. Levo M & Segal E (2014) In pursuit of design principles of regulatory sequences. *Nature Publishing Group* 15(7):453-468.

3. Struhl K (1991) Mechanisms for diversity in gene expression patterns. *Neuron* 7(2):177-181.

4. Small S, Blair A, & Levine M (1992) Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J* 11(11):4047-4057.

5. Swanson CI, Evans NC, & Barolo S (2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell* 18(3):359-370.

6. Farley EK, *et al.* (2015) Suboptimization of developmental enhancers. *Science* 350(6258):325-328.

7. Halfon MS, *et al.* (2000) Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell* 103(1):63-74.

8. Thanos D & Maniatis T (1995) Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* 83(7):1091-1100.

9. Weingarten-Gabbay S & Segal E (2014) The grammar of transcriptional regulation. *Hum Genet* 133(6):701-711.

10. Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10(9):605-616.

11. Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. *Dev Cell* 21(4):611-626.

12. Lambert SA, *et al.* (2018) The Human Transcription Factors. *Cell* 175(2):598-599.

13. Melnikov A, *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30(3):271-277.

14. Patwardhan RP, *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30(3):265-270.

15. Arnold CD, *et al.* (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339(6123):1074-1077.

16. Murtha M, *et al.* (2014) FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat Methods* 11(5):559-565.

17. Dickel DE, *et al.* (2014) Function-based identification of mammalian enhancers using site-specific integration. *Nat Methods* 11(5):566-571.

18. Siggers T, Duyzend MH, Reddy J, Khan S, & Bulyk ML (2011) Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol Syst Biol* 7:555.

19. Slattery M, *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147(6):1270-1282.

20. Ptashne M, *et al.* (1980) How the lambda repressor and cro work. *Cell* 19(1):1-11.

21. Liu X, Lee CK, Granek JA, Clarke ND, & Lieb JD (2006) Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res* 16(12):1517-1528.

22. Li XY, *et al.* (2011) The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol* 12(4):R34.

23. John S, *et al.* (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* 43(3):264-268.

24. Degner JF, *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482(7385):390-394.

25. Garber M, *et al.* (2012) A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular Cell* 47(5):810-822.

26. Zhu F, *et al.* (2018) The interaction landscape between transcription factors and the nucleosome. *Nature* 562(7725):76-81.

27. Panne D (2008) The enhanceosome. *Curr Opin Struct Biol* 18(2):236-242.

28. Arnosti DN & Kulkarni MM (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry* 94(5):890-898.

29. Hare EE, Peterson BK, Iyer VN, Meier R, & Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genetics* 4(6):e1000106.

30. Kulkarni MM & Arnosti DN (2003) Information display by transcriptional enhancers. *Development* 130(26):6569-6575.

31. Evans NC, Swanson CI, & Barolo S (2012) Sparkling insights into enhancer structure, function, and evolution. *Current topics in developmental biology* 98:97-120.

32. Ludwig MZ, Bergman C, Patel NH, & Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403(6769):564-567.

33. Rada-Iglesias A, *et al.* (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333):279-283.

34. May D, *et al.* (2011) Large-scale discovery of enhancers from human heart tissue. *Nat Genet* 44(1):89-93.

35. Blow MJ, *et al.* (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* 42(9):806-810.

36. Zaret KS & Carroll JS (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes & Development* 25(21):2227-2241.

37. Stampfel G, *et al.* (2015) Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* 528(7580):147-151.