HHS Public Access

Author manuscript

Curr Opin Syst Biol. Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

Curr Opin Syst Biol. 2020 October ; 23: 22–31. doi:10.1016/j.coisb.2020.08.002.

Seven myths of how transcription factors read the cis-regulatory code

Julia Zeitlinger[1,2]

[1]Stowers Institute for Medical Research, Kansas City, MO, USA

[2]The University of Kansas Medical Center, Kansas City, KS, USA

Abstract

Genomics data are now being generated at large quantities, of exquisite high resolution and from single cells. They offer a unique opportunity to develop powerful machine learning algorithms, including neural networks, to uncover the rules of the cis-regulatory code. However, current modeling assumptions are often not based on state-of-the-art knowledge of the cis-regulatory code from transcription, developmental genetics, imaging and structural studies. Here I aim to fill this gap by giving a brief historical overview of the field, describing common misconceptions and providing knowledge that might help to guide computational approaches. I will describe the principles and mechanisms involved in the combinatorial requirement of transcription factor binding motifs for enhancer activity, including the role of chromatin accessibility, repressors and low-affinity motifs in the cis-regulatory code. Deciphering the cis-regulatory code would unlock an enormous amount of regulatory information in the genome and would allow us to locate cis-regulatory genetic variants involved in development and disease.

Keywords

Transcription factors; cis-regulatory code; motif syntax; chromatin accessibility; cooperative binding; low-affinity binding motif; enhancer repression; transcriptional regulatory networks

Introduction

A fundamentally unresolved problem in biology is the cis-regulatory code, also known as the genome's "second code", which provides the means to read regulatory information in the genome. The most abundant cis-regulatory sequences are enhancers, which become active under very specific conditions and "enhance" the transcription of nearby genes. Since the activity of enhancers is determined by their sequence and can be reproduced outside their genomic context (e.g. in reporter assays), deciphering the cis-regulatory code of their

correspondence: jbz@stowers.org.

Conflict of interest statement

J.Z. owns a patent on ChIP-nexus (Patent No. 10287628).

activation should be a tractable problem. It is also a problem of great significance as over 80% of genetic variants associated with complex traits and diseases in humans are estimated to be found in cis-regulatory regions [1]. If we could decipher the cis-regulatory code, it would unlock an enormous amount of regulatory information and would allow us to locate cis-regulatory mutations and predict their effect on the organism.

The cis-regulatory code has been a long-standing problem and the focus of much research. Using genetics and biochemistry, Jacob and Monod discovered in the 1960s that genes in bacteria are regulated by DNA sequences found nearby on the same DNA molecule (in *cis*) [2,3]. With the advent of molecular biology in the 1970s and 1980s, it became possible to cut and paste sequences into different genomic contexts. This showed that an enhancer can function autonomously outside its normal genomic environment, independently of its orientation and exact distance to the gene [4]. Furthermore, it was discovered that enhancers contain short sequence motifs (~6–12 bases) that are recognized by transcription factors (TFs) [5]. Since TFs are often responsive to extracellular signals or are transcriptionally regulated in a spatial and temporal fashion during embryonic development, they provide the means for the cell to regulate when enhancers and the associated genes become active [6] (Figure 1A).

How exactly TFs regulate the activity of specific enhancers remains elusive. Developmental enhancers typically contain motifs for multiple TFs [7,8] and it is the specific combination of motifs that gives them their unique properties [9,10]. Experimental dissections of individual enhancer sequences suggest that enhancer activity may depend on the motif arrangement, also known as motif syntax [11,12]. Syntax includes the overall motif composition, motif order, motif orientation, and the spacing between the motifs. Early studies on the interferon-beta enhancer suggested very strict syntax rules, where multiple TFs assemble as an ‘enhanceosome’ [13]. However, studies on other enhancers suggested a very flexible syntax (‘billboard model’) [14]. Thus, individual enhancer studies did not reveal clear rules that could be applied genome-wide. Until today, we cannot predict the regulatory activity of enhancers from sequence alone.

With the advent of genomics technology, finding the rules of the cis-regulatory code seemed to be within reach [9,15]. Co-regulated genes pointed to the existence of enhancers with similar activity [16], chromatin immunoprecipitation (ChIP) technology enabled the identification of genomic regions bound by a TF *in vivo* [17–19], and the eventual development of large-scale reporter assays allowed the identification of large numbers of sequences with similar enhancer activity in a specific cell type [20]. Given large numbers of enhancers, it was reasonable to assume that the rules under which specific combinations and arrangements of motifs leading to enhancer activation could be identified. However, despite extensive experimental and computational efforts in the 2000s, no clear rules of the cis-regulatory code emerged [21,22]. The available data likely lacked the depth and resolution required to map with certainty the exact sequence motifs bound by TFs *in vivo.*

Ironically, with the development of increasingly powerful genomics technologies and computational methods in the last decade, efforts into deciphering the cis-regulatory code have declined rather than increased. Rather than focusing on the relationship between

sequence and gene regulation, research efforts have increasingly focused on chromatin states, RNA and 3D organization of the nucleus. Thus, the scientific questions evolved with the new opportunities that genomics technology offered and diverted from the fundamental problem of the cis-regulatory code, which became to be seen as either solved in principle or intractable.

Now we find ourselves in an era with genomics data of large quantities [23], of exquisite high resolution [24,25] and from single cells [26,27], all of which substantially improve the analysis of cis-regulatory sequences. Furthermore, powerful machine learning algorithms, including neural networks, have been developed for analyzing DNA sequences and predicting many associated genomic measurements [28–32]. This allowed the discovery of genome-wide syntax for the first time and revealed that syntax is often soft: two motifs may enhance TF binding when found in a preferred distance and do not have to be spaced at an exact base distance [33].

Despite the breakthrough tools that are now becoming available, the cis-regulatory code is not the primary focus of most current studies. Due to the diversion of the genomics field into other aspects of gene regulation, there is no longer a clear consensus on what we know about the cis-regulatory code from transcription, developmental genetics, imaging, structural studies and computational biology. This is problematic since computational analyses of cis-regulatory sequences, e.g. as part of transcriptional regulatory networks, often use overly simplified or unrealistic assumptions for how cis-regulatory information is encoded in the DNA. Here, I will outline some of these common misconceptions, point to the evidence that argues against them and describe a path towards more realistic assumptions.

MYTH 1: If we understand the cis-regulatory code in one cell type, we can predict enhancer activity in all cell types

A current emphasis in computational genomics is to learn patterns in some cell types and then impute these patterns in other cell types where experimental data are limited or not available [34,35]. While this can work well, there is often an unquestioned expectation that this is a solvable problem. Similarly, it is sometimes assumed that a complete map of all TF binding motifs and their target genes can be experimentally determined. In both cases, the underlying assumption is that the cis-regulatory code is universally applicable and that by studying a few cell types, other cell types can be inferred. After all, the biochemical and biophysical principles underlying TF binding and gene activation are shared among all cell types. However, unlike the genomic code for proteins, the cis-regulatory code is not a universal code. Cell types use vastly different TF motifs and syntax rules; even a few deviating details can make it impossible to accurately predict which cis-regulatory sequences are read out by a cell type. At the current state of knowledge, it might be more useful to think of each cell type having its own cis-regulatory code. Once we understand the code for a number of cell types in great detail, we might be able to derive general principles and determine the minimal set of experiments required to impute cis-regulatory code for new cell types.

REALITY: The cis-regulatory code is highly complex and specific for each cellular state

Cell types read out very different sets of cis-regulatory sequences in order to have a unique gene expression program and respond appropriately to the environment. Likewise, in the developing embryo, cells use different cis-regulatory sequences across time and space in order to develop along specified developmental trajectories and acquire specific cell identities [15,22]. To accomplish this, each cellular state has a unique combination of TFs, each with their specific rules of interactions and response to extracellular signals [6,36] (Figure 1A). Mammalian genomes encode over 1000 TFs, and TFs may bind with different sequence specificities and follow different syntax rules depending on their partner TF [37–40]. This large combinatorial complexity allows a vast number of possibilities by which cis-regulatory sequences are accessed in the genome and lead to enhancer activity. Some regions might only be active under very specific conditions, e.g. in a particular cell type at a certain time point during development, and their genomic coordinates may overlap with other cis-regulatory regions [41]. Learning this cis-regulatory repertoire will require a large number of high-quality experimental data *in vitro* and *in vivo,* as well as sophisticated computational tools that can learn the interaction rules that underlie the cis-regulatory code.

MYTH 2: Enhancers are regulated promiscuously by many TFs

Based on ChIP-seq and imaging data, TF binding occurs very promiscuously at hundreds of thousands of regions in the genome [42,43]. Thus, enhancers, especially when active, appear to be bound by a large number of TFs. This can be interpreted as enhancers having a large number of TF inputs, each with small contributions to gene activation [42]. However, ChIP-seq signal can be unspecific or non-functional. Furthermore, genetics studies show that deletion of a single TF or mutating an individual DNA binding motif often has very large effects on gene expression (Figure 2A). Thus, enhancers are regulated cooperatively by a small number of TFs, rather than additively by a large number of TFs.

REALITY: TFs are required in a combinatorial manner for enhancer activation

Genetics has shown that mutations in TF genes produce specific and striking phenotypes, with drastically altered gene expression patterns [44]. Likewise, mutating individual binding motifs within an enhancer may abolish the enhancer's activity [45–47]. This is often true for multiple motifs within enhancers, suggesting that their function is coupled via AND logic (Figure 2A) [48]. To achieve such coupling, the best understood mechanism is a composite motif, i.e. two motifs to which two TFs bind cooperatively through protein-protein interactions (Figure 2B) [21,37,49,50]. This often requires a strict spacing between the two motifs or may involve preferred soft spacings at helical distances (Figure 2C) [33]. TF cooperativity at preferred distances may also occur with weak or no protein-protein interactions between TFs [37,39]. For example, a TF motif may not be able to access a motif in chromatin unless a so-called pioneer TF binds nearby (< ~150 bp) and opens chromatin through nucleosome remodelling (Figure 2D) [33,51]. Finally, two TFs may also act

synergistically downstream of TF binding, e.g. by recruiting different co-factors that synergize in target gene activation (Figure 2E) [52]. Consistent with these mechanisms, the activation of developmental enhancers typically follows a sigmoidal curve and may show ultra-sensitivity in response to increasing concentrations of TFs (Figure 2F) [53–55]. This allows genes to be expressed in relatively sharp ON-OFF patterns and makes the cis-regulatory code more specific.

MYTH 3: Understanding the cis-regulatory code is a matter of mapping the direct TF binding sites

The recognition of DNA sequence motifs by TFs is the basis for the cis-regulatory code. These interactions have been studied by a variety of experimental techniques *in vitro* and are increasingly performed at high-throughput [37]. From such *in vitro* experiments, simple computational models such as a position weight matrix (PWM) can be derived and used to predict sequence matches in the genome. Due to the strong biophysical basis, identifying *bona fide* TF binding sites is typically the first step when analyzing enhancers. For example, Eric Davidson, who pioneered the study of transcriptional regulatory networks during sea urchin development, saw three steps in the identification of the cis-regulatory code: (1) identify TF binding sites, (2) experimentally determine their individual function (e.g. activation, repression, signal-induced) and (3) identify the rules by which the TF binding sites function together as Boolean input-output devices [56]. However, identifying TF binding sites based on their *in vitro* properties or their statistical significance relies on arbitrary thresholds that do not reflect how TFs bind *in vivo.* In order to fully understand the cis-regulatory code, binding sites should not be modeled separately from the cooperative interactions or downstream functions they mediate.

REALITY: TF binding and function are inherently combinatorial

TF binding *in vivo* depends on other TFs [36,38,57,58]. For example, TF may cooperate in binding with other TFs through physical interactions, or pioneer TFs may help the binding of other TFs by making the binding site accessible in chromatin [33,51]. Furthermore, TFs may function either as an activator or repressor dependent on nearby motifs [46,59]. Therefore, if we want to systematically decipher the cis-regulatory code from sequence, potential TF binding sites should be modeled directly in their cis-regulatory context, and not selected based on a fixed *in vitro* binding threshold prior to modeling. Convolutional neural networks are ideally suited for this since they model entire cis-regulatory sequences, including their higher-order motif combinations and syntax, without defining any features *a priori* [28–33]. They have therefore emerged as powerful tools for discovering elements of the cis-regulatory code.

MYTH 4: TF binding is secondary to chromatin regulation

It has been known since the first genome-wide ChIP experiments that TF binding *in vivo* does not correlate well with the presence of consensus binding motifs. However, TF binding is vastly improved when taking chromatin accessibility into account [35,60]. This leads to the impression that chromatin accessibility is regulated prior to the binding of most TFs.

Although this view acknowledges that pioneer TFs are important in creating the chromatin accessibility in a sequence-dependent manner, it prioritizes studying the regulation of chromatin states (histone modifications, 3D organization, etc.) and how they are established and maintained over time. Long-range chromatin repression mechanisms, such as those establishing different types of heterochromatin, indeed play an important role in keeping certain regions in the genome mostly inaccessible. However, the dynamic chromatin accessibility of enhancers during development is, for the most part, determined by TFs binding to cis-regulatory sequences and not the other way around [61–63].

REALITY: Chromatin accessibility is determined by cis-regulatory sequences

While pioneer TFs play an important role in making enhancer regions first accessible, chromatin accessibility is a result of the combined action of TFs (Figure 3) [64]. Pioneer TFs often work together with other TFs to increase chromatin accessibility [39,65,66] and are themselves required for the enhancer's activity [67,68]. When the enhancer is active, the central nucleosome is evicted [69–71] and chromatin accessibility is further increased [72]. Thus, chromatin accessibility appears to be the result of the interplay of multiple TFs and is likely an important mechanism by which these TFs function combinatorially as part of the cis-regulatory code. By using chromatin accessibility as prior probability for TF binding, we miss the opportunity to discover some of the pioneer TFs that mediate this accessibility.

MYTH 5: ChIP-seq binding data can be classified as binary events

To simplify models of gene regulation, ChIP-seq data are often classified as binary binding events. However, identifying a set of bound regions based on a chosen threshold has implications. It not only determines the level of unspecific binding or noise that is included in the data set, but also affects the functional contents of these regions. TF binding is higher at functional enhancers [73] and even higher at active enhancers [72] where the chromatin accessibility is highest (Figure 3). Therefore, dependent on the chosen threshold, the cis-regulatory context, including the presence of other TF binding motifs, is likely to be different.

REALITY: ChIP-seq binding at enhancers is a quantitative readout

ChIP-seq data show a continuum of binding levels. To understand the various components, ChIP-seq data can be compared to high-resolution ChIP-exo/nexus data, in which the TF binding signal has distinct footprints over motifs. Notably, ChIP-seq data contain higher levels of experimental background noise compared to ChIP-exo/nexus data, suggesting that some signal in ChIP-seq data is not specific for the measured TF [24,25]. Even in ChIP-exo/nexus data, small portions of signal are randomly distributed, most often across regions of highly accessible chromatin. This suggests that TFs may also bind nonspecifically to DNA. This interpretation is consistent with imaging studies showing that TFs may search and bind to many genomic regions very briefly (<1 s) before binding to a region with prolonged dwell time (~10 s), presumably because of a high-affinity binding motif [74]. The TF's dwell time may however not only depend on the motif's binding affinity, but also on the presence/state

of the nucleosome or the presence of partner TFs [74,75]. Even with short dwell times, a TF might have a high fractional occupancy in ChIP-seq data if the local TF concentration is high and the TF can quickly re-bind without long search times [76]. Such locally high TF concentrations have been observed at enhancers by imaging and described as condensates or hubs [77,78]. Lastly, it is important to keep in mind that ChIP-seq data represent cell population averages. For example, if certain TF binding events only occur in a fraction of cells, they will have a reduced ChIP-seq signal overall.

MYTH 6: Transcription factors mainly function as activators in mammalian cells

Although the lac and lambda repressors were the first sequence-specific TFs that were identified and extensively characterized [79], the role of repressors in enhancer activation is poorly studied in mammalian systems. This may be because the first mammalian enhancer, derived from SV40 and characterized by Walter Schaffner [4], did not involve relief of repression. Instead, it was proposed that nucleosomes repress enhancers in the absence of activation [79]. However, in model organisms such as Drosophila, sea urchin or yeast, genetics has shown that repressors are essential for gene regulation [6,44,80]. Mammalian systems have long lacked such extensive genetic characterization, but when in-depth analyses were performed for mouse development the importance of repressors has been clearly documented [81]. Recent genomics analysis have also confirmed that cis-regulatory elements frequently result in repressive activity [82]. This suggests that repressors are common throughout the animal kingdom and should be incorporated into models of gene expression in mammalian systems.

REALITY: Transcription factors frequently repress enhancers

The most detailed mechanistic understanding of repressors comes from pioneering work in Drosophila, where precise spatiotemporal gene expression patterns during development require a combination of activating and repressing TFs. A large number of TFs act as dedicated repressors (Figure 4A), and thus are generally repressive when bound to an enhancer [6,8,46]. Other TFs are dual TFs that can act as both an activator or repressor (Figure 4B,C). For example, binding sites for Drosophila NFkB are essential for either activating or repressing an enhancer [6]. It does so by acting intrinsically as a weak activator, but strongly promotes repression by helping the binding of a repressor to specific sequences nearby [83]. A repressor typically serves to repress and fine-tune the activity of enhancers by counteracting the effect of activating TFs that are bound nearby, e.g. through histone deacetylation [53,59,72]. Repressed enhancers are accessible in chromatin and show a poised/weakly active histone modification signature, a signature that is very common during mammalian development [72,82].

MYTH 7: Low-affinity binding motifs do not have a strong effect on enhancer function

Binding sequences that deviate from the consensus binding motif and are bound *in vitro* at low affinity are often omitted from analyses. They occur with high frequency by chance in genomic sequences and are therefore hard to identify as functional motifs *in vivo.* However, experimental evidence suggests that we miss crucial cis-regulatory information if we ignore low-affinity motifs [12,83,84]. It is therefore important to increase efforts into identifying and characterizing the effects of low-affinity motifs in cis-regulatory regions, especially with the emergence of neural networks which can detect low-affinity motifs [33]. Similarly, sequence information beyond motifs, such as DNA shape, subtle base preferences in motif flanks, and dinucleotide repeats may contribute to TF binding specificity [21].

REALITY: Low-affinity binding motifs are critical for the specificity of enhancers *in vivo*

Systematic analysis of synthetic enhancer constructs in Ciona has shown that low-affinity motifs are critical for producing *in vivo* expression patterns that are highly tissue-specific [12]. Several mechanisms could explain the requirement of low-affinity motifs for enhancer specificity. First, low-affinity motifs resulting in shorter TF dwell times may nevertheless be bound when the local TF concentration is high (Figure 4D). The shorter dwell times may even be advantageous by making the enhancer more tunable [76]. Second, some TF families, such as homeodomain TFs, bind very similar binding motifs, thus a low-affinity motif can render an enhancer more specific for a particular TF (Figure 4E) [84,85]. Finally, low-affinity motifs may make the binding of a TF dependent on its partner TF if high-affinity motifs are constitutively bound (Figure 4F) [83]. Thus, low-affinity motifs might be a common mechanism by which combinatorial TF requirements are embedded into the cis-regulatory code.

Conclusions

There is still much to be learned about the cis-regulatory code. We are only beginning to understand the mechanisms of how TFs function combinatorially in enhancer activation and how subtle motif syntax and low-affinity motifs influence this process. So far, there are too few examples to derive general principles. However, we likely have sufficient information to make reasonable assumptions when developing computational models. The goal is go beyond the identification of relevant motifs and to learn the rules of syntax and combinatorial interactions that predict enhancer activity from raw sequence. Neural networks are ideally suited for this since they can learn highly complex sequence patterns with unprecedented predictive power, allowing motifs to be directly modeled in their cis-regulatory context. Moreover, interpretation tools have recently been developed to extract the relevant sequence information, including motifs and their rules of syntax [28,29,33]. Combined with cutting-edge genomics technology and large-scale datasets, these approaches promise to revolutionize our ability to predict the function of cis-regulatory sequences in any

genome and provide us with unprecedented opportunities to study genetic cis-regulatory variation during development and disease.

Acknowledgments

I would like to thank Anshul Kundaje, Ziga Avsec, Melanie Weilert, Sabrina Krueger and Kaelan Brennan for discussions and comments on the manuscript. J.Z. is funded by the Stowers Institute for Medical Research and the NIH grant 1R01HG010211 to J.Z.

References

Papers with original research of particular interest for understanding the cis-regulatory code have been marked: * older paper, ** more recent paper

1. Gallagher MD, Chen-Plotkin AS: The Post-GWAS Era: From Association to Function. Am. J. Hum. Genet 2018, 102:717–730. [PubMed: 29727686]
2. Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol 1961, 3:318–356. [PubMed: 13718526]
3. Lewis M: The lac repressor. C. R. Biol 2005, 328:521–548. [PubMed: 15950160]
4. Banerji J, Rusconi S, Schaffner W: Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 1981,27:299–308. [PubMed: 6277502]
5. Mitchell PJ, Tjian R: Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science 1989, 245:371–378. [PubMed: 2667136]
6. Levine M, Davidson EH: Gene regulatory networks for development. Proc Natl Acad Sci USA 2005, 102:4936–4942. [PubMed: 15788537]
7. Kirchhamer CV, Davidson EH: Spatial and temporal information processing in the sea urchin embryo: modular and intramodular organization of the CyIIIa gene cis-regulatory system. Development 1996, 122:333–348. [PubMed: 8565846]
8. Small S, Kraut R, Hoey T, Warrior R, Levine M: Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev. 1991,5:827–839. [PubMed: 2026328] * This paper discovers and dissects the cis-regulatory sequence information for the expression of the second stripe of the Drosophila evenskipped gene (eve stripe 2), including the generation of a sharp ON-OFF pattern and the extensive involvement of repressors that inactivate (‘quench’) activators at short distances within the enhancer.
9. Halfon MS, Michelson AM: Exploring genetic regulatory networks in metazoan development: methods and models. Physiol. Genomics 2002, 10:131–143. [PubMed: 12209016]
10. Istrail S, Davidson EH: Logic functions of the genomic cis-regulatory code. Proc Natl Acad Sci USA 2005, 102:4954–4959. [PubMed: 15788531]
11. Zinzen RP, Senger K, Levine M, Papatsenko D: Computational models for neurogenic gene expression in the Drosophila embryo. Curr. Biol 2006, 16:1358–1365. [PubMed: 16750631]
12. Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS: Suboptimization of developmental enhancers. Science 2015, 350:325–328. [PubMed: 26472909] ** Large-scale mutagenesis of a Ciona enhancer and reporter gene expression in the embryo show that enhancer specificity is lost when the TF binding motifs and their syntax is optimal. This suggests that a combinatorial requirement for multiple enhancer inputs is often accomplished by suboptimal individual inputs.
13. Panne D: The enhanceosome. Curr. Opin. Struct. Biol 2008, 18:236–242. [PubMed: 18206362]
14. Kulkarni MM, Arnosti DN: Information display by transcriptional enhancers. Development 2003, 130:6569–6575. [PubMed: 14660545]
15. Levine M, Tjian R: Transcription regulation and animal diversity. Nature 2003, 424:147–151. [PubMed: 12853946]
16. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998, 95:14863–14868. [PubMed: 9843981]

17. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al.: Genome-wide location and function of DNA binding proteins. Science 2000, 290:2306–2309. [PubMed: 11125145]

18. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 2001,409:533–538. [PubMed: 11206552]

19. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J, et al.: Transcriptional regulatory code of a eukaryotic genome. Nature 2004, 431:99–104. [PubMed: 15343339]

20. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 2013, 339:1074–1077. [PubMed: 23328393] * This is the first large-scale enhancer reporter assay. It confirms that enhancer activity can be reproduced and discovered outside its genomic context. It also shows that some genomic fragments are active in the assay but inactive in the genome due to long-range chromatin repression.

21. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R: Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci 2014, 39:381–399. [PubMed: 25129887]

22. Spitz F, Furlong EEM: Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet 2012, 13:613–626. [PubMed: 22868264]

23. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al.: Architecture of the human regulatory network derived from ENCODE data. Nature 2012, 489:91–100. [PubMed: 22955619]

24. Rhee HS, Pugh BF: Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 2011, 147:1408–1419. [PubMed: 22153082] * The authors developed ChIP-exo, which introduces an exonuclease step during ChIP to generate TF footprints at base-resolution.

25. He Q, Johnston J, Zeitlinger J: ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat. Biotechnol 2015, 33:395–401. [PubMed: 25751057] ** The paper introduces ChIP-nexus, a robust ChIP-exo protocol that produces base-resolution TF binding footprints in mammalian cells. As shown later, the data are well suited for the training of neural networks, allowing them to predict these data from raw DNA sequence.

26. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ: Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015, 523:486–490. [PubMed: 26083756] ** This single-cell version of ATAC-seq is able to identify cis-regulatory regions and the relevant TF binding motifs across heterogenous cell populations. The data can be used for the training of neural networks in order to extract cis-regulatory sequence information from them.

27. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, Pliner HA, Christiansen L, Qiu X, Steemers FJ, et al.: The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature 2018, 555:538–542. [PubMed: 29539636] ** This is the first example applying single-cell ATAC-seq to a metazoan embryo during development and provides guiding principles for identifying tissue-specific cis-regulatory regions and understanding their dynamic regulation during development.

28. Eraslan G, Avsec Ž, Gagneur J, Theis FJ: Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet 2019, 20:389–403. [PubMed: 30971806]

29. Crawford J, Greene CS: Incorporating biological structure into machine learning models in biomedicine. Curr. Opin. Biotechnol 2020, 63:126–134. [PubMed: 31962244]

30. Alipanahi B, Delong A, Weirauch MT, Frey BJ: Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol 2015, 33:831–838. [PubMed: 26213851]

31. Zhou J, Troyanskaya OG: Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods 2015, 12:931–934. [PubMed: 26301843]

32. Kelley DR, Snoek J, Rinn JL: Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016, 26:990–999. [PubMed: 27197224]

33. Avsec Z, Weilert M, Shrikumar A, Alexandari A, Krueger S, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al.: Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. BioRxiv 2019, doi:10.1101/737981.** A convolutional neural network called BPNet is trained to predict ChIP-nexus data for pluripotency TFs at base-resolution. By using interpretation tools, TF binding motifs and rules of syntax are extracted, revealing a ~10 bp periodicity for Nanog binding and directional cooperativity.
34. Ernst J, Kellis M: Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol 2015, 33:364–376. [PubMed: 25690853] ** The principle of data imputation is demonstrated, showing that missing epigenomic datasets can be predicted by machine learning.
35. Schreiber J, Bilmes J, Noble WS: Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. Genome Biol. 2020, 21:82. [PubMed: 32228713]
36. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 2010, 38:576–589. [PubMed: 20513432]
37. Morgunova E, Taipale J: Structural perspective of cooperative transcription factor binding. Curr. Opin. Struct. Biol 2017, 47:1–8. [PubMed: 28349863]
38. Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA: Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell 2003, 113:395–404. [PubMed: 12732146] * It is shown that the cis-regulatory code manifests itself at the level of TF binding. Using yeast as a model, it is shown that the genome-wide binding of Ste12 in vivo changes with the developmental condition and is specified by a combinatorial requirement for partner TFs and signaling.
39. Deplancke B, Alpern D, Gardeux V: The genetics of transcription factor DNA binding variation. Cell 2016, 166:538–554. [PubMed: 27471964]
40. Halfon MS, Carmena A, Gisselbrecht S, Sackerson CM, Jiménez F, Baylies MK, Michelson AM: Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. Cell 2000, 103:63–74. [PubMed: 11051548]
41. Rice G, Rebeiz M: Evolution: how many phenotypes do regulatory mutations affect? Curr. Biol 2019, 29:R21–R23. [PubMed: 30620910]
42. Biggin MD: Animal transcription networks as highly connected, quantitative continua. Dev. Cell 2011, 21:611–626. [PubMed: 22014521]
43. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu X-J, White KP, Bussemaker HJ, et al.: Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proc Natl Acad Sci USA 2006, 103:12027–12032. [PubMed: 16880385]
44. Wieschaus E, Nusslein-Volhard C: The heidelberg screen for pattern mutants of drosophila: A personal account. Annu. Rev. Cell Dev. Biol 2016, 32:1–46. [PubMed: 27501451]
45. Robertson LM, Kerppola TK, Vendrell M, Luk D, Smeyne RJ, Bocchiaro C, Morgan JI, Curran T: Regulation of c-fos expression in transgenic mice requires multiple interdependent transcription control elements. Neuron 1995, 14:241–252. [PubMed: 7857636]
46. Stampfel G, Kazmar T, Frank O, Wienerroither S, Reiter F, Stark A: Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature 2015, 528:147–151. [PubMed: 26550828] ** Using a plasmid reporter assay, TFs and co-factors are tethered to various inactive enhancers to test their ability to restore enhancer activity. This shows that a wide variety of factors are activating or repressing. They sometimes function in a context-dependent manner, showing that TFs and co-factors may function synergistically in gene activation.
47. Kvon EZ, Kazmar T, Stampfel G, Yanez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A: Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature 2014, 512:91–95. [PubMed: 24896182]
48. Peter IS, Davidson EH: Assessing regulatory information in developmental gene regulatory networks. Proc Natl Acad Sci USA 2017, 114:5862–5869. [PubMed: 28584110]

49. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J: DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature 2015, 527:384–388. [PubMed: 26550823]
50. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM: Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 2009, 462:65–70. [PubMed: 19890324]
51. Zaret KS, Mango SE: Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr. Opin. Genet. Dev 2016, 37:76–81. [PubMed: 26826681]
52. Reiter F, Wienerroither S, Stark A: Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev 2017, 43:73–81. [PubMed: 28110180]
53. Crocker J, Ilsley GR, Stern DL: Quantitatively predictable control of Drosophila transcriptional enhancers in vivo with engineered transcription factors. Nat. Genet 2016, 48:292–298. [PubMed: 26854918] ** Enhancer expression patterns in the Drosophila embryo are manipulated using TALEN-engineered TFs that are either activating or repressing, avoiding uncertain effects of sequence manipulations. This engineered TF input is found to be additive across a sigmoidal curve of enhancer activation.
54. Melen GJ, Levy S, Barkai N, Shilo B-Z: Threshold responses to morphogen gradients by zero-order ultrasensitivity. Mol. Syst. Biol 2005, 1:2005.0028.
55. Burz DS, Rivera-Pomar R, Jackle H, Hanes SD: Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. EMBO J. 1998, 17:5998–6009. [PubMed: 9774343] * It is shown that Bicoid in Drosophila regulates enhancer activity by binding DNA cooperatively, allowing low-affinity binding sites to be bound. The resulting activation curve is sigmoidal.
56. Istrail S: Eric Davidson's Regulatory Genome for Computer Science: Causality, Logic, and Proof Principles of the Genomic cis-Regulatory Code. J. Comput. Biol 2019, 26:653–684. [PubMed: 31356126]
57. Yáñez-Cuna JO, Arnold CD, Stampfel G, Boryń LM, Gerlach D, Rath M, Stark A: Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res. 2014, 24:1147–1156. [PubMed: 24714811]
58. He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J: High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. Nat. Genet 2011, 43:414–420. [PubMed: 21478888]
59. Kulkarni MM, Arnosti DN: cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. Mol. Cell. Biol 2005, 25:3411–3420. [PubMed: 15831448]
60. Kaplan T, Li X-Y, Sabo PJ, Thomas S, Stamatoyannopoulos JA, Biggin MD, Eisen MB: Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet. 2011, 7:e1001290. [PubMed: 21304941] * This is one of the first systematic analyses of how TF binding levels measured by ChIP-seq might depend on TF motifs, cooperative TF binding and chromatin accessibility. This showed that incorporating experimentally measured chromatin accessibility data strongly improved the predictions of TF binding.
61. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavare S, Odom DT: Species-specific transcription in mice carrying human chromosome 21. Science 2008, 322:434–438. [PubMed: 18787134]
62. Sun Y, Nien C-Y, Chen K, Liu H-Y, Johnston J, Zeitlinger J, Rushlow C: Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation. Genome Res. 2015, 25:1703–1714. [PubMed: 26335633]
63. Foo SM, Sun Y, Lim B, Ziukaite R, O'Brien K, Nien C-Y, Kirov N, Shvartsman SY, Rushlow CA: Zelda potentiates morphogen activity by increasing chromatin accessibility. Curr. Biol 2014, 24:1341–1346. [PubMed: 24909324]
64. Zaret KS: Pioneering the chromatin landscape. Nat. Genet 2018, 50:167–169. [PubMed: 29374252]
65. Swinstead EE, Paakinaho V, Presman DM, Hager GL: Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription

factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. Bioessays 2016, 38:1150–1157. [PubMed: 27633730]

66. Long HK, Prescott SL, Wysocka J: Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 2016, 167:1170–1187. [PubMed: 27863239]
67. McDaniel SL, Gibson TJ, Schulz KN, Fernandez Garcia M, Nevil M, Jain SU, Lewis PW, Zaret KS, Harrison MM: Continued activity of the pioneer factor zelda is required to drive zygotic genome activation. Mol. Cell 2019, 74:185–195.e4. [PubMed: 30797686] ** It is demonstrated that the function of the pioneer TF Zelda is not just to open enhancers before activation, but is required throughout the stage where the enhancers are active.
68. Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, Hulselmans G, Potier D, Wouters J, Taskiran II, et al.: The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. Nat. Genet 2018, 50:1011–1020. [PubMed: 29867222]
69. He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, et al.: Nucleosome dynamics define transcriptional enhancers. Nat. Genet 2010, 42:343–347. [PubMed: 20208536] * This is the first demonstration that mammalian enhancers tend to have a central nucleosome that gets evicted during enhancer activation.
70. Gjidoda A, Tagore M, McAndrew MJ, Woods A, Floer M: Nucleosomes are stably evicted from enhancers but not promoters upon induction of certain pro-inflammatory genes in mouse macrophages. PLoS ONE 2014, 9:e93971. [PubMed: 24705533]
71. Brown CR, Mao C, Falkovskaia E, Jurica MS, Boeger H: Linking stochastic fluctuations in chromatin structure and gene expression. PLoS Biol. 2013, 11:e1001621. [PubMed: 23940458] ** Single-molecule maps of nucleosomes are generated for the yeast PHO5 gene and modeled, suggesting that stochastic fluctuations in nucleosome occupancy underlie gene regulation and that removal of the nucleosome where the activator binds is critical for gene activation.
72. Koenecke N, Johnston J, He Q, Meier S, Zeitlinger J: Drosophila poised enhancers are generated during tissue patterning with the help of repression. Genome Res. 2017, 27:64–74. [PubMed: 27979994] ** Genomics analysis of enhancer states in the Drosophila embryo shows that repressed enhancers are accessible for TFs, albeit less than active enhancers. They show histone modifications of ‘poised’ enhancers lacking histone acetylation, which is common in mammalian development and consistent with sequence-specific repressors mediating deacetylation.
73. Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, et al.: DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. Proc Natl Acad Sci USA 2012, 109:21330–21335. [PubMed: 23236164]
74. Chen J, Zhang Z, Li L, Chen B-C, Revyakin A, Hajj B, Legant W, Dahan M, Lionnet T, Betzig E, et al.: Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell 2014, 156:1274–1285. [PubMed: 24630727] ** One of the first studies tracking individual TFs inside cells using super-resolution microscopy. It shows that Sox2 and Oct4 frequently bind briefly (~1 s) and only occasionally longer (>10 s), presumably after finding a matching binding motif by trial-and-error. It also suggests that cooperative binding between Oct4 and Sox2 increases the dwell time and reduces the search time of Oct4.
75. Swinstead EE, Miranda TB, Paakinaho V, Baek S, Goldstein I, Hawkins M, Karpova TS, Ball D, Mazza D, Lavis LD, et al.: Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. Cell 2016, 165:593–605. [PubMed: 27062924] ** It is shown that the previously known pioneer TF FoxA1 binds to chromatin in a dynamic fashion (~10 s for long binding), suggesting that TFs make chromatin accessible through the recruitment of chromatin remodelers, rather than stable nucleosome binding. The pioneering role of FoxA1 is shared with other pioneer TFs (ER, GR), showing that chromatin accessibility can be regulated in a complex, combinatorial manner.
76. Liu Z, Tjian R: Visualizing transcription factor dynamics in living cells. J. Cell Biol 2018, 217:1181–1191. [PubMed: 29378780]
77. Tsai A, Alves MR, Crocker J: Multi-enhancer transcriptional hubs confer phenotypic robustness. elife 2019, 8.

78. Shrinivas K, Sabari BR, Coffey EL, Klein IA, Boija A, Zamudio AV, Schuijers J, Hannett NM, Sharp PA, Young RA, et al.: Enhancer Features that Drive Formation of Transcriptional Condensates. Mol. Cell 2019, 75:549–561.e7. [PubMed: 31398323]

79. Ptashne M: The chemistry of regulation of genes and other things. J. Biol. Chem 2014, 289:5417–5435. [PubMed: 24385432]

80. Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW, et al.: Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell 2014, 157:740–752. [PubMed: 24766815]

81. Frum T, Watts JL, Ralston A: TEAD4, YAP1 and WWTR1 prevent the premature onset of pluripotency prior to the 16-cell stage. Development 2019, 146.

82. Pang B, Snyder MP: Systematic identification of silencers in human cells. Nat. Genet 2020, 52:254–263. [PubMed: 32094911]

83. Papagianni A, Forés M, Shao W, He S, Koenecke N, Andreu MJ, Samper N, Paroush Z, González-Crespo S, Zeitlinger J, et al.: Capicua controls Toll/IL-1 signaling targets independently of RTK regulation. Proc Natl Acad Sci USA 2018, 115:1807–1812. [PubMed: 29432195] ** It is shown that Dorsal (NFkB) functions as repressor in Drosophila embryos by recruiting the repressor Capicua to low-affinity binding sites nearby. In the absence of Dorsal activation, Capicua still binds to high-affinity binding sites, but shows reduced binding to low-affinity binding sites. This demonstrates a combinatorial requirement for Dorsal and Capicua in enhancer repression.

84. Crocker J, Abe N, Rinaldi L, McGregor AP, Frankel N, Wang S, Alsawadi A, Valenti P, Plaza S, Payre F, et al.: Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell 2015, 160:191–203. [PubMed: 25557079] ** It is shown that clusters of multiple low-affinity Ubx binding sites are required for enhancer specificity in the late Drosophila embryo in vivo, demonstrating the importance of identifying and studying low-affinity motifs in enhancers.

85. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, et al.: Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell 2011, 147:1270–1282. [PubMed: 22153072]

Highlights

- To apply cutting-edge machine learning algorithms to genomics data, we need an ongoing discussion on how cis-regulatory information is encoded in DNA.
- The cis-regulatory code is inherently combinatorial and cell-type-specific.
- Chromatin accessibility and enhancer repression are encoded in cis-regulatory sequences and may involve low-affinity transcription factor binding sites.

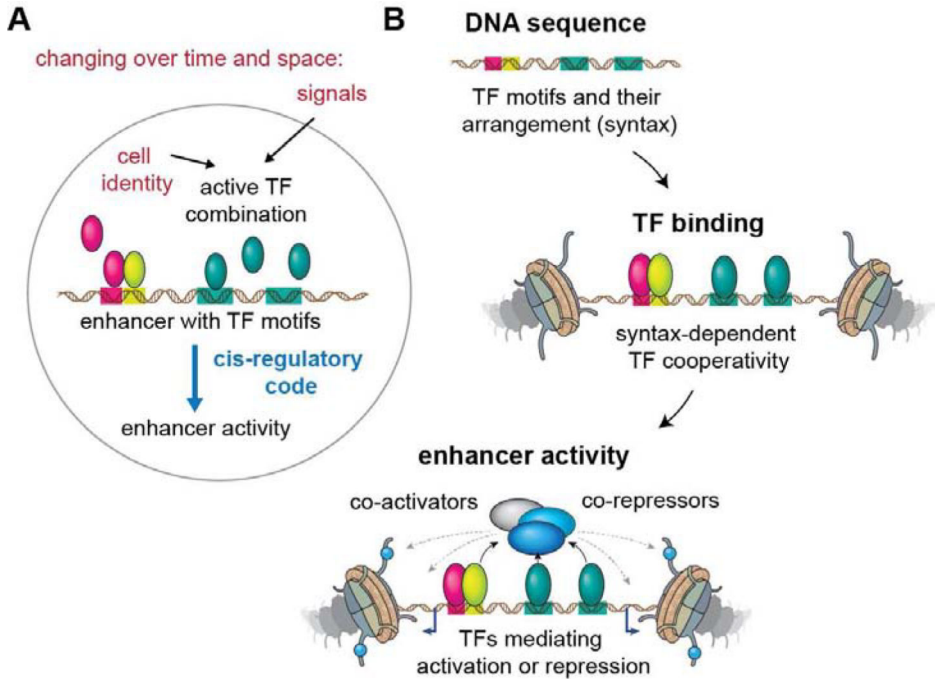

Figure 1: The cis-regulatory code defines how DNA sequence regulates enhancer activity.

(A) TFs are regulated transcriptionally and by extracellular signals such that each cell type contains a unique set of active TFs. Dependent on the specific TF combination, different sets of enhancers become active in each cell type. (B) The cis-regulatory DNA sequence contains TF motifs in specific arrangements (syntax). Dependent on syntax, the motifs are bound by TFs cooperatively. TFs then recruit co-activators or co-repressors, which regulate the activity of the enhancer.

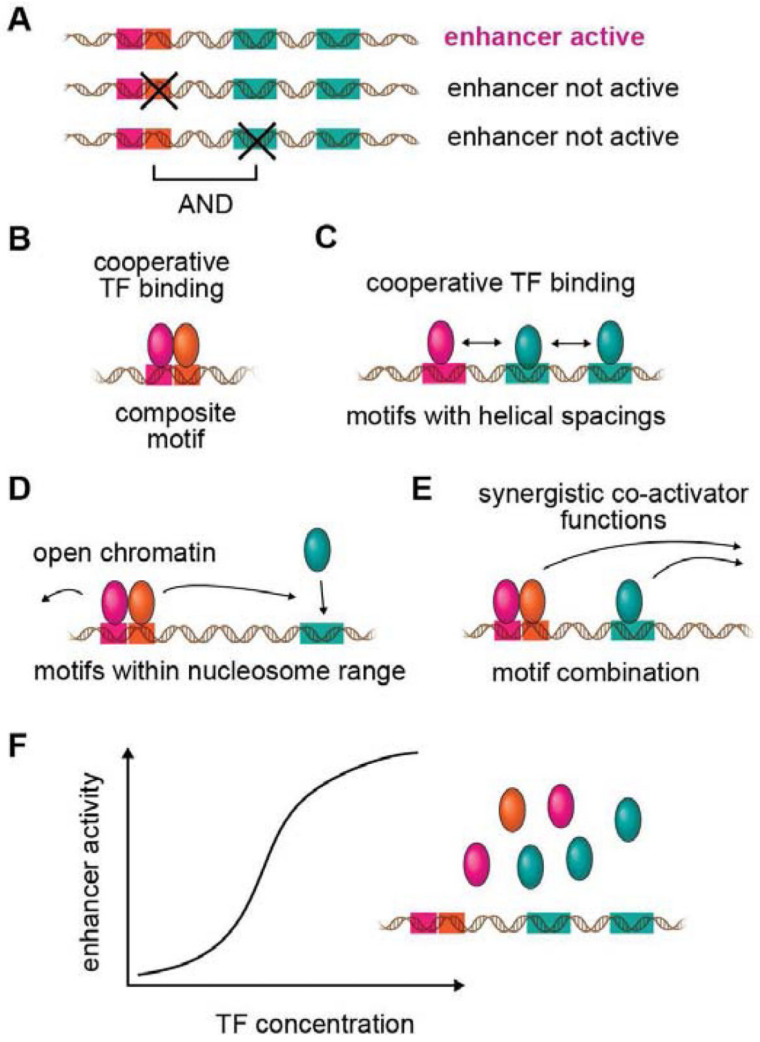

Figure 2: TF motifs often function together in an AND logic.

(A) Mutating different motifs in an enhancer can each lead to a loss of enhancer activity. Such AND logic between motifs can occur through (B) cooperative TF binding to composite motifs, (C) cooperative binding to motifs spaced with helical periodicity (~10 bp x N), (D) one TF opening chromatin such that another TF can bind (assisted loading), or (E) synergistic co-activator function. (F) The resulting enhancer activity follows a sigmoidal curve with increasing concentrations of a TF.

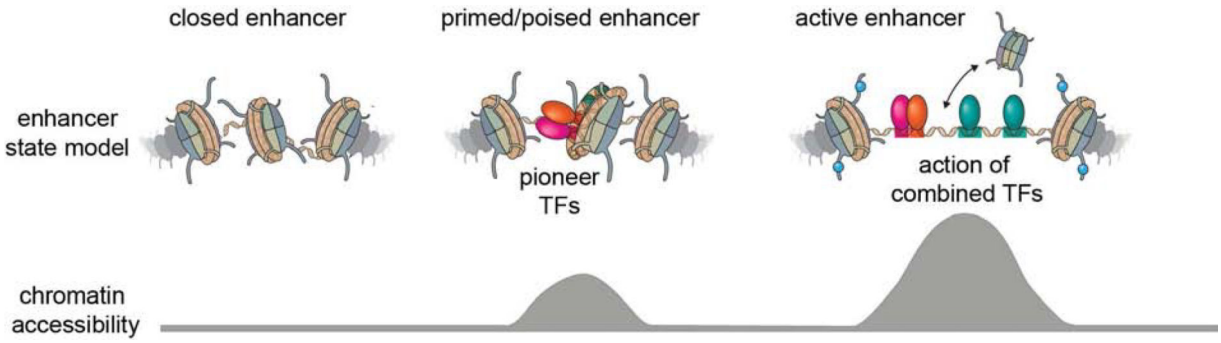

Figure 3: Chromatin accessibility is a readout of multiple TFs.

In the absence of appropriate TFs, nucleosomes maintain DNA in an inaccessible state (left). Pioneer TFs can bind their motifs in the presence of chromatin and make the region accessible (primed or poised enhancer, middle). The chromatin accessibility may be further increased by TFs both during the pioneering phase and during enhancer activation.

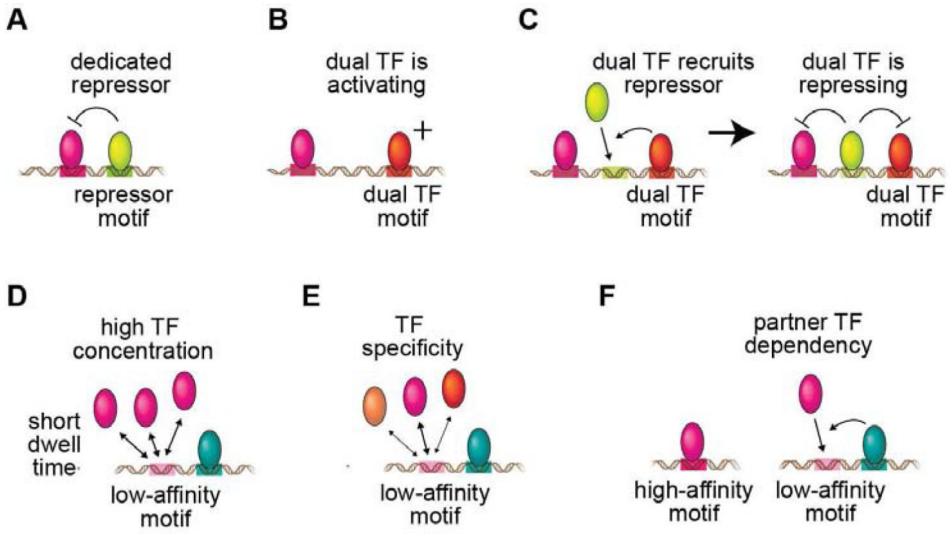

Figure 4: Mechanisms by which repressors (A-C) or low-affinity TF binding motifs (D-F) regulate enhancer activity and specificity.

(A) When dedicated repressors bind to their motifs, they counteract the activity of TFs bound nearby. (B) Dual TFs may be weakly activating by themselves, but (C) have a repressing effect when they recruit a repressor to a nearby repressor motif. Low-affinity motifs (D) are likely bound with shorter dwell times and require higher TF concentration to mediate enhancer activation, (E) may discriminate between closely related TF family members, or (F) may be dependent on a partner TF for binding.