Published in final edited form as:
Curr Opin Genet Dev. 2017 April ; 43: 110–119. doi:10.1016/j.gde.2017.02.007.

Transcription factor–DNA binding: beyond binding site motifs

Sachi Inukai[1,4], **Kian Hong Kock**[1,2,4], and **Martha L. Bulyk**[1,2,3,*]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

[2]Program in Biological and Biomedical Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

[3]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

Abstract

Sequence-specific transcription factors (TFs) regulate gene expression by binding to *cis*-regulatory elements in promoter and enhancer DNA. While studies of TF–DNA binding have focused on TFs' intrinsic preferences for primary nucleotide sequence motifs, recent studies have elucidated additional layers of complexity that modulate TF–DNA binding. In this review, we discuss technological developments for identifying TF binding preferences and highlight recent discoveries that elaborate how TF interactions, local DNA structure, and genomic features influence TF–DNA binding. We highlight novel approaches for characterizing functional binding site motifs that promise to inform our understanding of how TF binding controls gene expression and ultimately contributes to phenotype.

Introduction

Sequence-specific transcription factors (TFs) are key regulators of biological processes that function by binding to transcriptional regulatory regions (*e.g.*, promoters, enhancers) to control the expression of their target genes. Each TF typically recognizes a collection of similar DNA sequences, which can be represented as binding site motifs using models such as position weight matrices (PWMs) (reviewed in [1]; see Box 1.). The characterization of motifs is an important first step in understanding the regulatory functions of TFs that consequently shape gene regulatory networks.

Technological developments over the last decade have facilitated the characterization of DNA binding preferences for many TFs. Indeed, multiple large-scale studies in recent years have collectively elucidated motifs for thousands of TFs from a wide range of organisms

[*]Corresponding author. mlbulyk@genetics.med.harvard.edu.
[4]These authors contributed equally to this work.

The authors declare that they have no competing financial interests.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

[2-7]. Most of these studies have highlighted the evolutionary conservation of TF binding specificity, allowing the binding preferences of TFs lacking directly measured specificity data to be inferred from highly similar, characterized TFs [4,6,7]. Nevertheless, the current catalog of TF binding site motifs remains incomplete: binding preferences remain unknown — neither experimentally determined nor computationally inferred — for over 40% of the approximately 1,400 sequence-specific TFs encoded in the human genome [3,7-11], and several TF families (*e.g.*, those with high mobility group box or Cys2His2 zinc finger (zf) DNA binding domains (DBDs)) have disproportionately many uncharacterized TFs. Motif coverage of model organism TFs is similarly sparse [7], with the exception of *Saccharomyces cerevisiae* TFs [12]. The completion of motif catalogs remains a priority for bridging the gap between TFs and their regulatory targets.

Recent high-throughput studies have highlighted that there is more to TF–DNA binding than primary nucleotide sequence preferences. Accumulating evidence supports the widespread contributions of sequence context, including flanking sequences and DNA shape, in modulating sequence recognition. Interacting cofactors and TFs can also alter sequence preference [13]. Such additional features that impact TF–DNA recognition, together with differential TF expression and chromatin accessibility, are contributing to our understanding of what determines condition-specific TF binding [14]. In this review, we will discuss methods for identifying TF binding site motifs, emerging knowledge of additional features that influence TF–DNA recognition, and novel approaches in characterizing *in vivo*, functional consequences of TF binding. Because of space restrictions, we refer readers to recent reviews for discussions on TF binding site accessibility, mapping of regulatory elements to target genes, functional roles of low-affinity binding sites, and structural modeling of TF binding specificity [15-18].

Methods to identify TF binding site motifs

Methods to characterize TF–DNA binding preferences can be broadly categorized into *in vivo* and *in vitro* approaches. *In vivo* approaches can reveal TF binding events that occur in particular biological conditions (*e.g.*, cell type, treatment, time point), while *in vitro* methods are well suited for large-scale characterization of intrinsic TF binding sequence preferences.

A widely used *in vivo* method is chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) (reviewed in [19]). Briefly, genomic regions bound by a TF of interest are isolated via immunoprecipitation, and the bound sequences are identified through high-throughput sequencing. ChIP-seq signal ‘peaks’ are typically inferred through peak calling algorithms and then analyzed with software such as MEME-ChIP or ChIPMunk [20,21] to search for motifs enriched within the bound regions. Hundreds of ChIP-seq datasets have been generated, notably by the ENCODE Consortium [22], providing data on cell type-specific TF binding events. However, ChIP-seq presents several key challenges for determining TF binding site motifs [19]. ChIP-seq peaks can span dozens to hundreds of bases, whereas the binding site motifs for most TFs are shorter than 10 bp; together with fragment size heterogeneity and occasionally low ChIP enrichment, these factors can make precise mapping of binding sites difficult, especially when binding sites are clustered in close proximity [23]. Recent advances to the ChIP-seq protocol (*e.g.*, ChIP-exo, ChIP-nexus) address this issue by trimming excess sequences with exonucleases, allowing for nearly nucleotide-resolution mapping of binding sites [23,24]. However, these approaches are still insufficient to overcome challenges for robust *de novo* motif discovery for many TFs from ChIP-seq data. For example, the biological condition-specificity of ChIP-seq signals does not capture all possible binding sites, and the detection of indirect or cooperative binding events can obscure the profiled TF's direct binding preferences [19].

DNase-seq, ATAC-seq, and FAIRE-seq have garnered attention as alternatives to ChIP-seq to infer TF–DNA binding sites, as they allow for TF-independent investigations of transcriptional regulation [25-28]. These approaches identify regions of accessible chromatin using a nonspecific DNA nuclease, transposase, or formaldehyde crosslinking coupled with phenol-chloroform extraction, respectively [25-29]. However, their utility in identifying direct TF binding sites through nucleotide-resolution "footprints", or relatively protected DNA within accessible chromatin [27], has been debated [30,31] and appears to be TF-dependent [32]. For example, Hager and colleagues showed that DNase-seq does not adequately capture footprints of highly dynamic TFs, such as Sox2 and glucocorticoid receptor, that have short DNA residence times [33].

Given the limitations described above, current *in vivo* methods are constrained in their ability to identify motifs *de novo*. Their utility in identifying direct TF–DNA interactions relies on the quality of existing TF binding specificity models with which to scan TF-bound sequences or against which to compare the results of *de novo* motif finding analysis of the bound regions. Motifs have instead been identified largely using *in vitro* methods that can systematically assess many possible target sequences for a single TF. These methods include bacterial one-hybrid (B1H) selection [34], microfluidics-based mechanically induced trapping of molecular interactions (MITOMI) [35,36], protein binding microarrays (PBMs) [37,38], and *in vitro* selection-based (*e.g.*, SELEX) approaches [4,39-41] (reviewed in [42]). These assays have been instrumental in elucidating the intrinsic DNA binding specificities of thousands of TFs (see Table 1 for motif databases).

However, these *in vitro* approaches have been unable thus far to characterize the specificities of all TFs, as the necessity for expressing soluble, active TFs of interest precludes the characterization of TFs that have been intractable in this regard. Despite such challenges, these methods continue to be useful in elucidating features associated with TF–DNA recognition. SELEX has been adapted for characterizing the binding of TF pairs and complexes [41,43]. Spec-seq is similar to SELEX-seq but can accurately measure relative DNA binding affinities and is well suited for assessing the impacts of DNA sequence variants [44]. Genomic-context PBMs (gcPBMs) interrogate select 30- to 36-bp sequences and have elucidated the contributions of flanking genomic sequences to motif recognition [38]. The combination of *in vivo* and *in vitro* approaches has led to insights on features that influence TF–DNA association *in vivo* (Figure 1).

Complexities of TF binding determinants

Multiple specificities intrinsic to individual transcription factors

While the majority of TFs demonstrate singular binding specificities, some TFs have the intrinsic capability to recognize multiple motifs [13]. Nakagawa *et al.* identified several forkhead TFs that each bind two apparently unrelated sequence motifs (5′-RYAAAYA and 5′-GACGC); they found that this multiple binding specificity arose independently in at least two different evolutionary lineages [45]. Siggers *et al.* found that certain paralogous yeast C2H2-zf TFs in the Msn2 family recognize a core motif common to all Msn2 family members (5′-AGGGG) as well as TF-specific motifs (*e.g.*, 5′-ATAGGR, 5′-AGGNAC), and that recognition of the different motifs evolved modularly [46]. Large-scale studies of TF–DNA binding preferences have shown that recognition of multiple motifs, including those with variable spacing and orientation of motif half-sites, are not uncommon [2,4,6,12,47]. These findings highlight the utility of more complex motif representations in depicting the full sequence preferences of TFs and emphasize the need for motif catalogs to explicitly account for multiple intrinsic binding specificities.

Molecular interactions that modulate binding specificities

TF binding specificity is influenced by intra- and intermolecular TF interactions. For example, large-scale B1H assays and structural analyses have shown that in C2H2-zf proteins, residues at the finger–finger interface and the order of fingers within an array frequently influence specificity [48,49]. In particular, adjacent fingers can constrain mutual finger orientation and influence the positioning of specificity-determining residues [46,50]. Residues outside the zinc fingers (*e.g.*, within inter-zinc finger linker sequences) may also be implicated in intramolecular interactions that can contribute to alternate TF-DNA docking geometries and hence altered DNA-binding specificity [46].

Proteins that interact with TFs — whether they are TFs or non-DNA-binding cofactors forming homodimeric, heterodimeric, or multimeric complexes — can alter binding specificity [41,51] (see also review by Reiter, Wienerroither, and Stark in this issue [52]). Dimerization can elicit latent specificities of related TFs: for instance, T-box factors demonstrate similar monomeric binding specificities, but homodimeric binding preferences distinguish this family into seven distinct specificity classes [4]. The eight *Drosophila* Hox TFs show largely similar monomeric binding specificities [53]. However, when in complex with the TF Exd and the HM domain of the TF Hth, the resulting complexes demonstrate novel binding specificities, with distinct DNA-binding preferences for each HM-Exd-Hox complex [41]. Cooperative binding by multiple TFs can also give rise to new motifs. Taipale and colleagues used consecutive affinity-purification (CAP)-SELEX to identify hundreds of heterodimeric motifs for human TFs, many of which are markedly different from those expected from a simple combination of the individual TFs' motifs [43]. These heterodimeric motifs include instances where monomeric binding differences within a TF family were masked, yielding the same heterodimeric motifs; or where distinct heterodimeric specificities were revealed despite similar monomeric motifs amongst TFs of the same family. The authors estimated that cooperative TF interactions might specify ~25,000 distinct motifs [43].

In addition, non-DNA-binding cofactors can modulate TF binding specificity. The *S. cerevisiae* transcriptional cofactors Met4 and Met28 do not display intrinsic DNA-binding specificity. However, when these cofactors form a trimeric complex with the basic helix-loop-helix (bHLH)-containing TF Cbf1, the resulting complex demonstrates novel binding specificity for a composite binding site containing an E-box sequence (CACGTG) flanked by an RYAAT motif with a 2-bp spacer in between. The specific recognition of this composite binding site requires the full trimeric complex [51]. These examples highlight the importance of examining the binding specificity of TF complexes, since surprising cases of novel binding specificity may arise from such molecular interactions, distinct from monomeric binding site preferences.

Features beyond primary DNA sequence motifs that modulate binding specificity

Beyond primary nucleotide sequence motifs, genomic DNA contains contextual and epigenetic information that may affect TF binding. For example, some TFs (*e.g.*, NRF1) are unable to bind methylated forms of their canonical TF binding site motifs [54]. Intriguingly, several TFs recognize distinct methylated versus unmethylated DNA sequence motifs *in vitro* and *in vivo* [55,56] (see [57] for review).

Several TFs use DNA shape features (*e.g.*, minor groove width and rotational parameters such as helix twist, propeller twist and roll; see [58] for review) to distinguish between similar binding sites. Mann and colleagues identified specific residues in the Hox TF Scr that confer shape recognition, demonstrating that recognition of DNA shape features can be separated from that of nucleotide bases [59]. Quantitative models integrating DNA shape and sequence features outperformed sequence-only models in predicting TF–DNA binding specificity; these models further suggested that different TF families use distinct shape readout mechanisms [60,61]. DNA shape features outside the motif can also determine binding. The paralogous yeast bHLH TFs Cbf1 and Tye7 recognize distinct structural features of sequences flanking their shared E-box motif (11 bp and 5 bp flanking regions, respectively), allowing these TFs to recognize distinct sites *in vitro* and *in vivo* [38]. Sequence contexts, both those immediately flanking core binding sites and those extending farther away from the motif, have also been shown to impact binding [62,63]. In particular, different GC composition of the sequences surrounding a motif can significantly affect TF binding independently of DNA shape and nucleosome occupancy [62].

Characterizing the functional consequences of TF binding

In parallel with emerging knowledge on features that modulate TF recognition, naturally occurring genetic variants have been informative in assessing the functional role of TF binding site motifs. A handful of disease-associated variants that disrupt or introduce TF binding site motifs have been studied in detail, providing mechanistic insights into pathogenesis (reviewed in [64]). Over 70% of the thousands of noncoding variants found to be associated with common diseases or traits in genome-wide association studies overlap TF binding motifs in accessible chromatin [65]; for the vast majority of these variants, it remains to be determined if they affect gene regulation. Integration of genotype information with chromatin accessibility, ChIP-seq, and gene expression data (Figure 2) has begun to link motif-disrupting variants with altered TF binding and target gene expression [65-67]. A more accurate and complete TF binding motif catalog will be important in facilitating the identification and prioritization of damaging regulatory variants.

In contrast, large-scale genotyping and ChIP studies have highlighted that variants found in TF binding site motifs constitute only a small proportion of variants that affect TF binding *in vivo* [67-69]. This finding is consistent with the existence of additional features that affect TF binding. Furthermore, Dermitzakis and colleagues showed that distal variants (located tens of kb away from binding sites) could impact TF binding [70] through changes in chromatin state. Additional TF-binding focused quantitative trait loci studies will shed further light on the spectrum of parameters that influence *in vivo* TF binding.

Massively parallel reporter assays (MPRAs) have linked variation in TF binding sites with changes in reporter gene expression by simultaneously measuring the activity of thousands of synthetic *cis*-regulatory sequences in a particular tissue or cell type [71-74]. Combining sequence capture technology with MPRAs has allowed for the targeted characterization of naturally occurring variants in longer sequence contexts than what typically has been examined (less than 150 bp) by synthetic DNA libraries [75,76]. A major caveat of MPRAs is that regulatory elements are assayed outside of their native chromosomal contexts, and so the functional consequences of variants identified through MPRAs need to be validated using complementary approaches.

The development of programmable nucleases for genome editing has expanded the experimental toolkit for direct, functional characterization of putative TF binding sites. For example, transcription activator-like effector nuclease (TALEN)-mediated genome editing was used to show that a rare mutation in the promoter of the γ-globin gene drives transactivation of the gene through *de novo* recruitment of the TAL1 TF [77]. CRISPR/Cas9-mediated genome editing was used to characterize an obesity-associated variant that disrupted a conserved ARID binding site motif; this variant was associated with altered gene expression and cellular metabolism [78]. These advances in genome editing are beginning to allow *in situ* validation of candidate causal variants identified by MPRAs [79-81] or other approaches. Saturating mutagenesis of a BCL11A erythroid enhancer using ZF nucleases, TALENs, and CRISPR/Cas9 identified a key functional GATA1 binding site [82,83]. Such saturating mutagenesis approaches will allow for the systematic screening of TFs and *cis*-regulatory elements within native chromosomal contexts [84,85]. Further developments in high-throughput genome editing approaches are anticipated to facilitate the identification of additional *in vivo* features that modulate TF binding.

Conclusions and perspectives

Recent large-scale efforts to elucidate TF binding specificities have made great headway in linking TFs to binding sites, yet the catalog of binding specificities remains incomplete. Emerging research has highlighted the involvement of numerous features beyond sequence motifs, including DNA shape and flanking sequences, which modulate binding site recognition. These features complicate TF specificity determination but have been consistently shown to improve predictions of binding sites [61-63,86]. Saturating mutagenesis of *cis*-regulatory regions is anticipated to identify additional features that influence TF binding and will allow for a more quantitative dissection of their contributions. Because many *cis*-regulatory elements are predicted to be bound by TF complexes *in vivo*, multimeric motifs represent vast sequence specificity space that has been largely unexplored.

Additional sources of TF diversity have not been assessed systematically and require further investigation. For instance, the *Drosophila* C2H2-zf TF Lola has 23 splice isoforms with distinct sets of zinc fingers; these isoforms exhibit diverse binding specificities [87]. Over 1,000 of the approximately 1,400 human TF genes have known splice isoforms, many with the potential to regulate different sets of target genes [88]; collectively, there are nearly 8,000 human TF isoforms.

We recently reported that missense variants in TF DBDs that likely alter TF–DNA binding activity are collectively prevalent in humans [89]. Our comparison of PBM data to ChIP-seq and RNA-seq data showed that specific HOXD13 DBD mutations demonstrate changes in genomic occupancy and gene expression that are consistent with their altered *in vitro* binding specificities [89]. Furthermore, our analysis of genotype data suggests that most individuals harbor a unique repertoire of TF alleles and DNA-binding activities. Future studies are needed to determine how TF variants shape the regulatory landscape and contribute to phenotypic diversity.

Understanding how TFs recognize their DNA binding sites forms the basis for understanding transcriptional regulation and how this process goes awry in disease. This goal is complicated by the numerous layers of complexity that lie between TF activity and phenotype, starting with the difficulty in relating intrinsic TF–DNA binding preferences to gene expression. The dramatically increased throughput and accessibility of technologies including DNA sequencing and oligonucleotide synthesis have driven the development of new experimental approaches that have highlighted certain nuances of TF–DNA binding. However, we still lack both a comprehensive compendium of features that influence TF binding and an understanding of how these pieces fit together. Furthermore, we are just beginning to transition from a qualitative interpretation to a more quantitative description of how the various determinants contribute to TF–DNA binding. Advances towards these goals promise to illuminate fundamental concepts of TF-directed gene regulation and to refine predictions of *in vivo* TF–DNA binding, gene expression, and associated functional consequences for phenotypes.

Acknowledgments

This work was funded by National Institutes of Health/National Human Genome Research Institute grant # R01 HG003985 (M.L.B.) and an A*STAR National Science Scholarship (K.H.K.). We apologize to authors whose work we could not cite because of space restrictions. We thank Julia Rogers and Luca Mariani for critical reading of this manuscript.

References

1. Stormo GD. Modeling the specificity of protein-DNA interactions. Quant Biol. 2013; 1:115–130. [PubMed: 25045190]

2. Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci U S A. 2014; 111:2367–2372. [PubMed: 24477691]
3. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2015; 43:D117–122. [PubMed: 25378322]
4. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. DNA-binding specificities of human transcription factors. Cell. 2013; 152:327–339. [PubMed: 23332764]
5. Narasimhan K, Lambert SA, Yang AW, Riddell J, Mnaimneh S, Zheng H, Albu M, Najafabadi HS, Reece-Hoyes JS, Fuxman Bass JI, et al. Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. Elife. 2015; 4
6. Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EE, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife. 2015; 4

7*. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158:1431–1443. The authors used PBMs to elucidate the DNA binding specificities of more than 1,000 TFs from a wide range of organisms; their findings suggest that TF–DNA binding specificity can be inferred from TFs containing highly similar DBDs. [PubMed: 25215497]

8. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009; 10:252–263. [PubMed: 19274049]
9. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D, et al. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. 2013; 41:D171–176. [PubMed: 23203885]
10. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Soboleva AV, Kasianov AS, Ashoor H, BaAlawi W, Bajic VB, Medvedeva YA, Kolpakov FA, et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. 2016; 44:D116–125. [PubMed: 26586801]
11. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016; 44:D110–115. [PubMed: 26531826]
12. Gordân R, Murphy KF, McCord RP, Zhu C, Vedenko A, Bulyk ML. Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol. 2011; 12 R125-2011-2012-2012-r2125.
13. Siggers T, Gordân R. Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res. 2014; 42:2099–2111. [PubMed: 24243859]
14. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012; 22:1798–1812. [PubMed: 22955990]
15. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci. 2014; 39:381–399. [PubMed: 25129887]
16. Yao L, Berman BP, Farnham PJ. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. Crit Rev Biochem Mol Biol. 2015; 50:550–573. [PubMed: 26446758]
17. Crocker J, Noon EP, Stern DL. The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. Curr Top Dev Biol. 2016; 117:455–469. [PubMed: 26969995]
18. Joyce AP, Zhang C, Bradley P, Havranek JJ. Structure-based modeling of protein: DNA specificity. Brief Funct Genomics. 2015; 14:39–49. [PubMed: 25414269]

19. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet. 2012; 13:840–852. [PubMed: 23090257]

20. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ. Deep and wide digging for binding motifs in ChIP-Seq data. Bioinformatics. 2010; 26:2622–2623. [PubMed: 20736340]

21. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011; 27:1696–1697. [PubMed: 21486936]

22. Encode PC. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

23. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011; 147:1408–1419. [PubMed: 22153082]

24. 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]

25. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011; 21:456–464. [PubMed: 21106903]

26. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013; 10:1213–1218. [PubMed: 24097267]

27. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009; 6:283–289. [PubMed: 19305407]

28. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011; 21:1757–1767. [PubMed: 21750106]

29. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007; 17:877–885. [PubMed: 17179217]

30. Sung MH, Baek S, Hager GL. Genome-wide footprinting: ready for prime time? Nat Methods. 2016; 13:222–228. [PubMed: 26914206]

31. Vierstra J, Stamatoyannopoulos JA. Genomic footprinting. Nat Methods. 2016; 13:213–221. [PubMed: 26914205]

32. He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Methods. 2014; 11:73–78. [PubMed: 24317252]

33. Sung MH, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol Cell. 2014; 56:275–285. [PubMed: 25242143]

34. Meng X, Brodsky MH, Wolfe SA. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol. 2005; 23:988–994. [PubMed: 16041365]

35. Fordyce PM, Gerber D, Tran D, Zheng J, Li H, DeRisi JL, Quake SR. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol. 2010; 28:970–975. [PubMed: 20802496]

36. Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. Science. 2007; 315:233–237. [PubMed: 17218526]

37. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006; 24:1429–1435. [PubMed: 16998473]

38. Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. 2013; 3:1093–1104. [PubMed: 23562153]

39. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpaa MJ, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. 2010; 20:861–873. [PubMed: 20378718]

40. Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol. 2002; 20:831–835. [PubMed: 12101405]

41. 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]

42. Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nat Rev Genet. 2010; 11:751–760. [PubMed: 20877328]

43**. 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. CAP-SELEX analyses of TF–TF–DNA interactions revealed roles for DNA in mediating cooperative interactions between TFs and identified hundreds of unique heterodimeric TF binding site motifs. [PubMed: 26550823]

44. Zuo Z, Stormo GD. High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. Genetics. 2014; 198:1329–1343. [PubMed: 25209146]

45. Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML. DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc Natl Acad Sci U S A. 2013; 110:12349–12354. [PubMed: 23836653]

46*. Siggers T, Reddy J, Barron B, Bulyk ML. Diversification of transcription factor paralogs via noncanonical modularity in C2H2 zinc finger DNA binding. Mol Cell. 2014; 55:640–648. The authors demonstrated that paralogous C2H2 zinc finger TFs can gain novel DNA-binding specificities in a modular fashion, without altering binding to common core sequences recognized by all family members. [PubMed: 25042805]

47. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009; 324:1720–1723. [PubMed: 19443739]

48. Gupta A, Christensen RG, Bell HA, Goodwin M, Patel RY, Pandey M, Enuameh MS, Rayla AL, Zhu C, Thibodeau-Beganny S, et al. An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins. Nucleic Acids Res. 2014; 42:4800–4812. [PubMed: 24523353]

49. Persikov AV, Wetzel JL, Rowland EF, Oakes BL, Xu DJ, Singh M, Noyes MB. A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. Nucleic Acids Res. 2015; 43:1965–1984. [PubMed: 25593323]

50. Garton M, Najafabadi HS, Schmitges FW, Radovani E, Hughes TR, Kim PM. A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. Nucleic Acids Res. 2015; 43:9147–9157. [PubMed: 26384429]

51. Siggers T, Duyzend MH, Reddy J, Khan S, Bulyk ML. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Mol Syst Biol. 2011; 7:555. [PubMed: 22146299]

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. Mann RS, Lelli KM, Joshi R. Hox specificity unique roles for cofactors and collaborators. Curr Top Dev Biol. 2009; 88:63–101. [PubMed: 19651302]

54. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schubeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015; 528:575–579. [PubMed: 26675734]

55. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C, et al. DNA methylation presents distinct binding sites for human transcription factors. Elife. 2013; 2:e00726. [PubMed: 24015356]

56. Mann IK, Chatterjee R, Zhao J, He X, Weirauch MT, Hughes TR, Vinson C. CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. Genome Res. 2013; 23:988–997. [PubMed: 23590861]

57. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet. 2016

58. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein- DNA recognition. Annu Rev Biochem. 2010; 79:233–269. [PubMed: 20334529]

59*. Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, Rohs R, Mann RS. Deconvolving the recognition of DNA shape from sequence. Cell. 2015; 161:307–318. The authors used targeted mutagenesis and residue swapping experiments to show that shape readout is separable from base readout by the anterior Hox TF Scr. [PubMed: 25843630]

60. Yang L, Zhou T, Dror I, Mathelier A, Wasserman WW, Gordân R, Rohs R. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 2014; 42:D148–155. [PubMed: 24214955]

61*. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordân R, Rohs R. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A. 2015; 112:4654–4659. Combining DNA shape and base sequence features improved the modeling of *in vitro* TF–DNA binding specificities when compared to base sequence-only models. [PubMed: 25775564]

62. Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. Genome Res. 2015; 25:1268–1280. [PubMed: 26160164]

63. Levo M, Zalckvar E, Sharon E, Dantas Machado AC, Kalma Y, Lotam-Pompan M, Weinberger A, Yakhini Z, Rohs R, Segal E. Unraveling determinants of transcription factor binding outside the core binding site. Genome Res. 2015; 25:1018–1029. [PubMed: 25762553]

64. Deplancke B, Alpern D, Gardeux V. The Genetics of Transcription Factor DNA Binding Variation. Cell. 2016; 166:538–554. [PubMed: 27471964]

65. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. [PubMed: 22955828]

66. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]

67. Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. 2012; 22:860–869. [PubMed: 22300769]

68. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. Variation in transcription factor binding among humans. Science. 2010; 328:232–235. [PubMed: 20299548]

69. Tehranchi AK, Myrthil M, Martin T, Hie BL, Golan D, Fraser HB. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. Cell. 2016; 165:730–741. [PubMed: 27087447]

70. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A, et al. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell. 2015; 162:1039–1050. [PubMed: 26300124]

71. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci U S A. 2012; 109:19498–19503. [PubMed: 23129659]

72. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012; 30:271–277. [PubMed: 22371084]

73. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012; 30:265–270. [PubMed: 22371081]

74. Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. Genome Res. 2013; 23:1908–1915. [PubMed: 23921661]

75. Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC. Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Res. 2016; 26:238–255. [PubMed: 26576614]

76. Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LT, Fernandez N, Ballester B, Andrau JC, Spicuglia S. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat Commun. 2015; 6:6905. [PubMed: 25872643]

77. Wienert B, Funnell AP, Norton LJ, Pearson RC, Wilkinson-White LE, Lester K, Vadolas J, Porteus MH, Matthews JM, Quinlan KG, et al. Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. Nat Commun. 2015; 6:7085. [PubMed: 25971621]

78. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med. 2015; 373:895–907. [PubMed: 26287746]

79*. Ferreira LM, Meissner TB, Mikkelsen TS, Mallard W, O'Donnell CW, Tilburgs T, Gomes HA, Camahort R, Sherwood RI, Gifford DK, et al. A distant trophoblast-specific enhancer controls HLA-G expression at the maternal-fetal interface. Proc Natl Acad Sci U S A. 2016; 113:5364–5369. These three studies integrated MPRAs with orthogonal approaches (*e.g.*, genome editing and genomic data) to dissect the contributions of regulatory variants on TF binding. [PubMed: 27078102]

80*. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, et al. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016; 165:1519–1529. These three studies integrated MPRAs with orthogonal approaches (*e.g.*, genome editing and genomic data) to dissect the contributions of regulatory variants on TF binding. [PubMed: 27259153]

81*. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS, et al. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell. 2016; 165:1530–1545. These three studies integrated MPRAs with orthogonal approaches (*e.g.*, genome editing and genomic data) to dissect the contributions of regulatory variants on TF binding. [PubMed: 27259154]

82**. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. 2015; 527:192–197. These two studies were among the earliest large-scale applications of genome editing methods to dissect regulatory elements in enhancers. [PubMed: 26375006]

83**. Vierstra J, Reik A, Chang KH, Stehling-Sun S, Zhou Y, Hinkley SJ, Paschon DE, Zhang L, Psatha N, Bendana YR, et al. Functional footprinting of regulatory DNA. Nat Methods. 2015; 12:927–930. These two studies were among the earliest large-scale applications of genome editing methods to dissect regulatory elements in enhancers. [PubMed: 26322838]

84. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature. 2014; 513:120–123. [PubMed: 25141179]

85. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJ, Gifford DK, Sherwood RI. High-throughput mapping of regulatory DNA. Nat Biotechnol. 2016; 34:167–174. [PubMed: 26807528]

86. Isakova A, Berset Y, Hatzimanikatis V, Deplancke B. Quantification of Cooperativity in Heterodimer-DNA Binding Improves the Accuracy of Binding Specificity Models. J Biol Chem. 2016; 291:10293–10306. [PubMed: 26912662]

87. Enuameh MS, Asriyan Y, Richards A, Christensen RG, Hall VL, Kazemian M, Zhu C, Pham H, Cheng Q, Blatti C, et al. Global analysis of Drosophila Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. Genome Res. 2013; 23:928–940. [PubMed: 23471540]

88. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, Valencia A, Tress ML. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 2013; 41:D110–117. [PubMed: 23161672]

89*. Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, Woodard J, Mariani L, Kock KH, Inukai S, et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science. 2016; 351:1450–1454. The authors identified many thousands of naturally occurring human TF DBD missense variants, which were collectively prevalent in humans. They further demonstrated that numerous such variants alter *in vitro* TF–DNA binding properties. [PubMed: 27013732]

90. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990; 18:6097–6100. [PubMed: 2172928]
91. Zhao Y, Ruan S, Pandey M, Stormo GD. Improved models for transcription factor binding site identification using nonindependent interactions. Genetics. 2012; 191:781–790. [PubMed: 22505627]
92. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. PLoS Comput Biol. 2013; 9:e1003214. [PubMed: 24039567]
93. 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]
94. Schipper JL, Gordân RM. Transcription Factor-DNA Binding Motifs in Saccharomyces cerevisiae: Tools and Resources. Cold Spring Harb Protoc. 2016; 2016 pdb top080622.
95. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000; 16:16–23. [PubMed: 10812473]
96. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34:D108–110. [PubMed: 16381825]
97. Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform. 2008; 9:326–332. [PubMed: 18436575]
98. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. 2016; 7:11101. [PubMed: 27089393]
99. Jiang P, Singh M. CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. Nucleic Acids Res. 2014; 42:2833–2847. [PubMed: 24366875]
100. Kou, Y., C, EY., Clark, NR., Duan, Q., Tan, CM., Ma'ayan, A. ChEA2: Gene-set libraries from ChIP-X experiments to decode the transcription regulome. In: Springer Berlin Heidelberg. , editor. Availability, Reliability, and Security in Information Systems and HCI. 2013. p. 416-430.
101. Wang P, Qin J, Qin Y, Zhu Y, Wang LY, Li MJ, Zhang MQ, Wang J. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. Nucleic Acids Res. 2015; 43:W264–269. [PubMed: 25916854]
102. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, Wang S, Chen J, Shen L, Duan X, et al. CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics. 2012; 28:1411–1412. [PubMed: 22495751]
103. Sun H, Qin B, Liu T, Wang Q, Liu J, Wang J, Lin X, Yang Y, Taing L, Rao PK, et al. CistromeFinder for ChIP-seq and DNase-seq data reuse. Bioinformatics. 2013; 29:1352–1354. [PubMed: 23508969]
104. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015; 43:W39–49. [PubMed: 25953851]
105. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37:W202–208. [PubMed: 19458158]

Box 1

Binding site motifs and DNA-binding specificity

Binding site motifs

A single TF can recognize dozens to hundreds of DNA binding site sequences over a range of binding affinities [13]. Hence, the TF binding specificity (i.e., preferential binding of specific sequences) cannot be adequately represented using any one DNA sequence. Instead, TF binding specificities are often represented as binding site motifs, which summarize the collection of preferentially bound sequences. These motifs can be used to scan sequences of interest (e.g., genomic regions) to predict TF binding sites.

Binding specificity models

Several different binding specificity models are currently used to derive binding site motifs. The position weight matrix (PWM) is the most commonly used model. PWMs describe the probability of a given nucleotide's occurrence at each position in the DNA binding site [1]. These PWMs can be represented graphically as sequence logos [90]. A major assumption of the standard PWM model is that each position contributes independently to binding. Several TFs require more complex models to describe their binding specificities. Extensions to the PWM model that specifically address nucleotide position interdependence or that allow variable-length degenerate spacers separating half-sites have been developed [91]. Other models include k-mers, with scores assigned to individual sequences of length k [37], and those inferred from hidden Markov models or machine learning-based approaches [92,93], which are more flexible and can model variable spacing and nucleotide position interdependence in a single framework.

Multiple binding specificities

While TFs can inherently bind multiple sequences, some TFs exhibit intrinsic specificity for multiple distinct motifs. These multiple DNA-binding specificities may exist due to the differential usage of multiple DBDs within the same TF and multiple TF–DNA docking conformations [13,45,46].

For more detailed discussions on TF binding specificity, binding site motifs, different binding specificity models, and possible mechanisms for multiple binding specificities, we refer readers to other reviews [1,13,42,94,95].

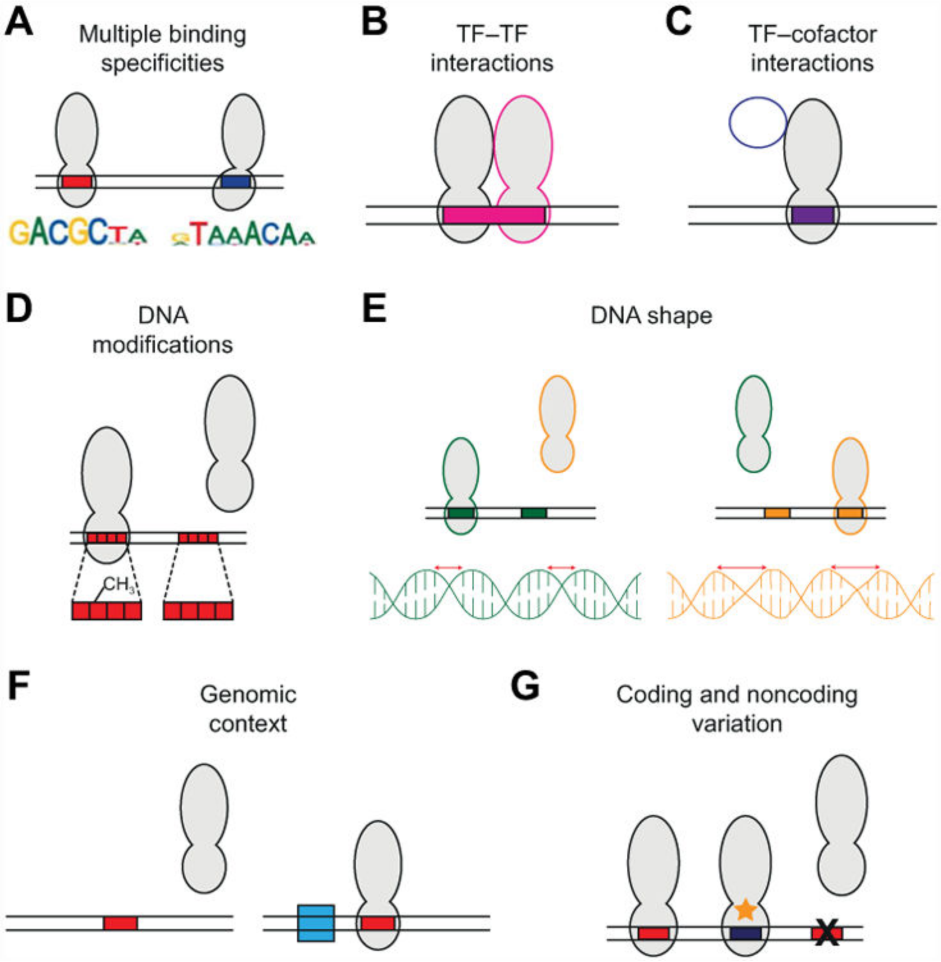

Figure 1. Numerous features of TFs or DNA binding sites beyond primary nucleotide sequence motifs can modulate transcription factor (TF)–DNA recognition

TF-level features: (A) Several TFs can display binding specificity for multiple, distinct nucleotide sequence motifs. Motifs shown are examples of two motifs bound by the bispecific human forkhead TF FOXN2: FHL (red box) and FkhP (blue box) binding site motifs; motifs were obtained from UniPROBE (Accession Number UP00521) [45]. Interactions between (B) TFs and (C) TFs and non-DNA-binding cofactors [51] can specify distinct binding site motifs from the monomeric TF motif. *DNA-level features*: (D) DNA modifications, such as 5-methylcytosine (left), can modulate TF binding. (E) Numerous TFs use DNA shape readout, such as minor groove width (depicted by red arrows), and rotational parameters such as helix twist, propeller twist, and roll, as part of TF–DNA recognition. (F) Sequences and features outside of the binding site motif (depicted by blue box), such as GC content and / or DNA shape, can modulate TF– DNA binding. These features may immediately flank the core binding site, or may extend more distally from the motif. (G) Genetic variation in either the TF protein sequence (depicted by orange star, middle) or the DNA binding site (depicted by X, right) can alter TF–DNA binding.

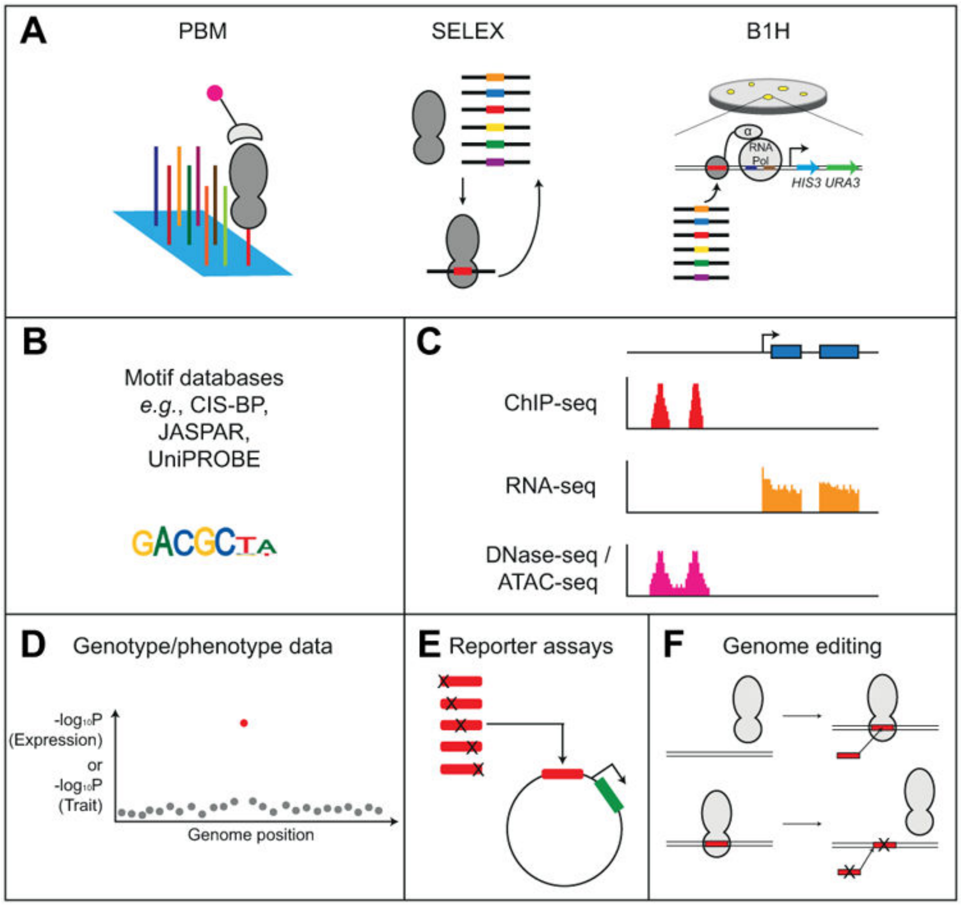

Figure 2. Combinations of *in vitro* and *in vivo* approaches allow identification and investigation of functional TF–DNA binding sites

(A) Techniques such as B1H, PBMs, and SELEX enable one to determine the intrinsic binding specificities of TFs. (B) Motif catalogs aid in identification of putative TF binding site motifs. Results from technologies in (A) can then be integrated with (C) *in vivo* genomic approaches, such as ChIP-seq, RNA-seq, and chromatin accessibility profiling methods (*e.g.*, DNase-seq, ATAC-seq) to identify or infer direct versus indirect DNA binding sites *in vivo* and regulatory roles of the TFs. Data from (D) investigations of natural genetic variation, such as in genome-wide association studies, which allow investigators to identify signals associated with TFs or TF binding sites implicated in particular traits or as expression quantitative trait loci (eQTL); (E) enhancer or promoter activity reporter assays; and (F) experimental perturbation approaches (*e.g.*, genome editing), are used in assessing the contributions of motifs to gene expression and phenotypes.

Table 1

TF binding site motif databases and related tools

Motif databases	Link	Types of motif data	Represented organisms	Data content (as of September 2016)	References
CIS-BP	http://cisbp.ccbr.utoronto.ca/	Experimentally defined (*e.g.*, through PBM, SELEX, ChIP-seq) and inferred TF binding site motifs. Also includes publicly available motifs from other databases (*e.g.*, FactorBook, JASPAR, TRANSFAC).	>300 species, spanning metazoans, plants, fungi, protists, alga, and viruses	>3,000 directly measured motifs and >55,000 inferred motifs, representing >58,000 TFs.	[7]
HOCOMOCO	http://hocomoco.autosome.ru/	Motifs from alignment of TF binding regions in ChIP-seq and HT-SELEX experiments using ChIPMunk.	Human and mouse	v10 provides motifs for 601 human TFs and 396 mouse TFs.	[10]
JASPAR	http://jaspar.genereg.net/	Curated motifs for eukaryotes based on PBM, HT-SELEX, and ChIP-seq data. Also contains web tool to infer binding profile for an input TF protein sequence.	Eukaryotes	CORE collection: 1,082 mostly non-redundant TF binding profiles.	[11]
TRANSFAC	http://www.gene-regulation.com/index2	Curated motifs for TFs and their experimentally determined (from peer-reviewed papers) and predicted binding sites and regulated genes. Also contains ChIP-seq fragment data. Requires paid subscription for full access.	>300 species, with a focus on human, mouse, rat, yeast, and plants	Professional version (paid subscription required) features 23,277 factors; 6,133 available publicly.	[96,97]
UniPROBE	http://uniprobe.org	TF binding specificity based on published PBM data. Also contains web-based sequence scanning tool for TF binding site *k*-mer matches.	>20 species, with a focus on human, mouse, yeast, nematodes, plants, fruit fly, and pathogens	566 unique TFs, and binding preferences of >100 TF variants for select human TFs spanning six major structural classes.	[3]
Other relevant databases and tools	**Link**	**Types of data**	**Represented organisms**	**Data content (as of September 2016)**	**References**
AlleleDB	http://alleledb.gersteinlab.org	*Cis*-regulatory single nucleotide variants of individuals from 1000 Genomes Project demonstrating allele-specific binding and expression.	Human	>8,000 allele-specific binding variants and >360,000 allele-specific expression variants.	[98]
CCAT	http://cat.princeton.edu/	Computational pipeline for predicting genome-wide TF co-binding and combinatorial TF interactions.	Fruit fly	Predictions for 324 TFs.	[99]
ChEA2	http://amp.pharm.mssm.edu/ChEA2/	ChIP-X (*e.g.*, ChIP-seq, ChIP-chip, ChIP-PET) and DamID data; ChIP Enrichment Analysis (ChEA) software available.	Human and mouse	ChEA2: Target gene mapping for 200 TFs from 221 publications.	[100]
ChIP-Array	http://jjwanglab.org/chip-array/	ChIP-X (ChIP-chip or ChIP-seq) and gene expression data.	*Arabidopsis*, fruit fly, human, mouse, nematode, rat, yeast	v2.0 provides 6,548 PWMs of 4,481 TFs from 7 species.	[101]

Motif databases	Link	Types of motif data	Represented organisms	Data content (as of September 2016)	References
Cistrome DB	http://cistrome.org/db	ChIP-seq and DNase-seq datasets, including those from ENCODE and Epigenome Roadmap projects.	Human and mouse	>10,000 TF ChIP-seq, >10,000 histone mark ChIP-seq, and >1,000 DNase-seq datasets.	[102,103]
FactorBook	http://www.factorbook.org	ENCODE ChIP-seq data (TF and histone modifications).	Human and mouse	167 TFs from 837 experiments.	[9]
MEME Suite	http://meme-suite.org/	Collection of motif databases and web-based tools for motif discovery, enrichment, scanning, and comparison.	Eukaryotes and prokaryotes.	Collates external motif and sequence databases for multiple species.	[104,105]
TFBSshape	http://rohslab.cmb.usc.edu/TFBSshape	DNA shape features for TF binding sites.	23 different species including fruit fly, human, mouse, nematode, yeast	739 TF datasets from 23 different species.	[60]