Transcription factor–DNA binding: beyond binding site motifs
Sachi Inukai1,*, Kian Hong Kock1,2,* and Martha L Bulyk1,2,3

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.

Addresses
1 Division of Genetics, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA
2 Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA 02138, USA
3 Department of Pathology, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA

Corresponding author: Bulyk, Martha L (mlbulyk@genetics.med.harvard.edu)
*These authors contributed equally to this work.

Current Opinion in Genetics & Development 2017, 43:110-119
This review comes from a themed issue on Genome architecture and expression
Edited by Bart Deplancke and Charles Sagerstrom
For a complete overview see the Issue and the Editorial
Available online 27th March 2017
http://dx.doi.org/10.1016/j.gde.2017.02.007
0959-437X/© 2017 Elsevier Ltd. All rights reserved.

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 Ref. [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 [2–6,7*]. Most of these studies have highlighted the evolutionary conservation of TF binding specificities, 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 1400 sequence-specific TFs encoded in the human genome [3,7*,8–11], and several TF families (e.g., those with high mobility group box or Cys2His2 zinc finger (C2H2-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
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].

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 Ref. [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 non-specific DNA nuclease, transpose, 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, Sung et al. 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 (PBM) [37,38], and in vitro selection-based (e.g., SELEX) approaches [4,39–41] (reviewed in Ref. [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
### Table 1

**TF binding site motif databases and related tools**

<table>
<thead>
<tr>
<th>Motif databases</th>
<th>Link</th>
<th>Types of data</th>
<th>Represented organisms</th>
<th>Data content (as of September 2016)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CIS-BP</strong></td>
<td><a href="http://cisbp.ccb.r.utoronto.ca/">http://cisbp.ccb.r.utoronto.ca/</a></td>
<td>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)</td>
<td>&gt;300 species, spanning metazoans, plants, fungi, protists, algae, and viruses</td>
<td>&gt;3000 directly measured motifs and &gt;55 000 inferred motifs, representing &gt;58 000 TFs</td>
<td>[7]</td>
</tr>
<tr>
<td><strong>HOCOMOCO</strong></td>
<td><a href="http://hocomoco.autosome.ru/">http://hocomoco.autosome.ru/</a></td>
<td>Motifs from alignment of TF binding regions in ChIP-seq and HT-SELEX experiments using ChIPMunk</td>
<td>Human and mouse</td>
<td>v10 provides motifs for 601 human TFs and 396 mouse TFs</td>
<td>[10]</td>
</tr>
<tr>
<td><strong>TRANSFAC</strong></td>
<td><a href="http://www.gene-regulation.com/index2">http://www.gene-regulation.com/index2</a></td>
<td>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</td>
<td>&gt;300 species, with a focus on human, mouse, rat, yeast, and plants</td>
<td>Professional version (paid subscription required) features 23 277 factors; 6133 available publicly</td>
<td>[96,97]</td>
</tr>
<tr>
<td><strong>UniPROBE</strong></td>
<td><a href="http://uniprobe.org">http://uniprobe.org</a></td>
<td>TF binding specificity based on published PBM data. Also contains web-based sequence scanning tool for TF binding site k-mer matches</td>
<td>&gt;20 species, with a focus on human, mouse, yeast, nematodes, plants, fruit fly, and pathogens</td>
<td>566 unique TFs, and binding preferences of &gt;100 TF variants for select human TFs spanning six major structural classes</td>
<td>[3]</td>
</tr>
</tbody>
</table>

**Other relevant databases and tools**

<table>
<thead>
<tr>
<th>Motif databases</th>
<th>Link</th>
<th>Types of data</th>
<th>Represented organisms</th>
<th>Data content (as of September 2016)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AlleleDB</strong></td>
<td><a href="http://alleledb.gersteinlab.org">http://alleledb.gersteinlab.org</a></td>
<td>Cis-regulatory single nucleotide variants of individuals from 1000 Genomes Project demonstrating allele-specific binding and expression</td>
<td>Human</td>
<td>&gt;8000 allele-specific binding variants and &gt;360 000 allele-specific expression variants</td>
<td>[98]</td>
</tr>
<tr>
<td><strong>CCAT</strong></td>
<td><a href="http://cat.princeton.edu/">http://cat.princeton.edu/</a></td>
<td>Computational pipeline for predicting genome-wide TF co-binding and combinatorial TF interactions</td>
<td>Fruit fly</td>
<td>Predictions for 324 TFs</td>
<td>[99]</td>
</tr>
<tr>
<td><strong>ChEA2</strong></td>
<td><a href="http://amp.pharm.mssm.edu/ChEA2/">http://amp.pharm.mssm.edu/ChEA2/</a></td>
<td>Chip-X (Chip-chip or ChIP-seq) and gene expression data</td>
<td>Human and mouse</td>
<td>ChEA2: Target gene mapping for 200 TFs from 221 publications</td>
<td>[100]</td>
</tr>
<tr>
<td><strong>ChiP-Array</strong></td>
<td><a href="http://ijwanglab.org/chip-array/">http://ijwanglab.org/chip-array/</a></td>
<td>Arabidopsis, fruit fly, human, mouse, nematode, rat, yeast</td>
<td>Human and mouse</td>
<td>v2.0 provides 6548 PWMs of 4481 TFs from seven species</td>
<td>[101]</td>
</tr>
<tr>
<td><strong>Cistrome DB</strong></td>
<td><a href="http://cistrome.org/db">http://cistrome.org/db</a></td>
<td>ChIP-seq and DNase-seq datasets, including those from ENCODE and Epigenome Roadmap projects</td>
<td>Human and mouse</td>
<td>&gt;10 000 TF ChIP-seq, &gt;10 000 histone mark ChIP-seq, and &gt;1000 DNase-seq datasets</td>
<td>[102,103]</td>
</tr>
<tr>
<td><strong>FactorBook</strong></td>
<td><a href="http://www.factorbook.org">http://www.factorbook.org</a></td>
<td>ENCODE ChIP-seq data (TF and histone modifications)</td>
<td>Human and mouse</td>
<td>167 TFs from 837 experiments</td>
<td>[9]</td>
</tr>
<tr>
<td><strong>TFBSshape</strong></td>
<td><a href="http://rohslab.cmb.usc.edu/TFBSshape">http://rohslab.cmb.usc.edu/TFBSshape</a></td>
<td>DNA shape features for TF binding sites</td>
<td>23 different species including fruit fly, human, mouse, nematode, yeast</td>
<td>739 TF datasets from 23 different species</td>
<td>[60]</td>
</tr>
</tbody>
</table>
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

---

**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) [48]. 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 major 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.
paralogous yeast C2H2-zf TFs in the Msn2 family recognize a core motif common to all Msn2 family members (5′-AGGGG-3′) as well as TF-specific motifs (e.g., 5′-ATAGGR-3′, 5′-AGGNAC-3′), 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 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 intramolecular 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 may 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 et al. [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. Jolma et al. 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 monomorphic binding differences within a TF family were masked, yielding the same heterodimeric motif; or where distinct heterodimeric specificities were revealed despite similar monomorphic 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 (CAGTG) flanked by an RYAA 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 Ref. [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 Ref. [58] for review) to distinguish between similar binding sites. Abe et al. identified specific residues in the Hox TFScr 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 roles 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 Ref. [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, Waszak et al. 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

![Figure 2](https://www.sciencedirect.com)
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 [79,80,81] or other approaches. Saturating mutagenesis of a BCL11A erythroid enhancer using ZF nucleases, TALENs, and CRISPR/Cas9 identified a putative 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 [61,62,63,86]. 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 1000 of the approximately 1400 human TF genes have known splice isoforms, many with the potential to regulate different sets of target genes [88]; collectively, there are nearly 8000 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 toward 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.

Conflict of interest statement
Nothing declared.

Acknowledgements
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 and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest


The authors used PBMs to elucidate the DNA binding specificities of more than 1000 TFs from a wide range of organisms; their findings suggest that TF–DNA binding specificity can be inferred from TFs containing highly similar DBDs.


118 Genome architecture and expression


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.


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.


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.


Combining DNA shape and base sequence features improved the modeling of in vitro TF–DNA binding specificities when compared to base sequence-only models.


82. This study integrated MPRAs with orthogonal approaches (e.g., genome editing and genomic data) to dissect the contributions of regulatory variants on TF binding.
84. This study was among the earliest large-scale applications of genome editing methods to dissect regulatory elements in enhancers.
86. This study was among the earliest large-scale applications of genome editing methods to dissect regulatory elements in enhancers.
93. The authors identified millions of naturally occurring human TF-BEDD missense variants, which were collectively prevalent in humans. They further demonstrated that numerous such variants alter in vitro TF-DNA binding properties.