Cell

Leading Edge
Review

The Human Transcription Factors

Samuel A. Lambert,[1,9] Arttu Jolma,[2,9] Laura F. Campitelli,[1,9] Pratyush K. Das,[3] Yimeng Yin,[4] Mihai Albu,[2] Xiaoting Chen,[5] Jussi Taipale,[3,4,6,*] Timothy R. Hughes,[1,2,*] and Matthew T. Weirauch[5,7,8,*]

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[2]Donnelly Centre, University of Toronto, Toronto, ON, Canada
[3]Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland
[4]Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna, Sweden
[5]Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[6]Department of Biochemistry, Cambridge University, Cambridge CB2 1GA, United Kingdom
[7]Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[8]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA
[9]These authors contributed equally
*Correspondence: ajt208@cam.ac.uk (J.T.), t.hughes@utoronto.ca (T.R.H.), Matthew.Weirauch@cchmc.org (M.T.W.)
https://doi.org/10.1016/j.cell.2018.01.029

Transcription factors (TFs) recognize specific DNA sequences to control chromatin and transcription, forming a complex system that guides expression of the genome. Despite keen interest in understanding how TFs control gene expression, it remains challenging to determine how the precise genomic binding sites of TFs are specified and how TF binding ultimately relates to regulation of transcription. This review considers how TFs are identified and functionally characterized, principally through the lens of a catalog of over 1,600 likely human TFs and binding motifs for two-thirds of them. Major classes of human TFs differ markedly in their evolutionary trajectories and expression patterns, underscoring distinct functions. TFs likewise underlie many different aspects of human physiology, disease, and variation, highlighting the importance of continued effort to understand TF-mediated gene regulation.

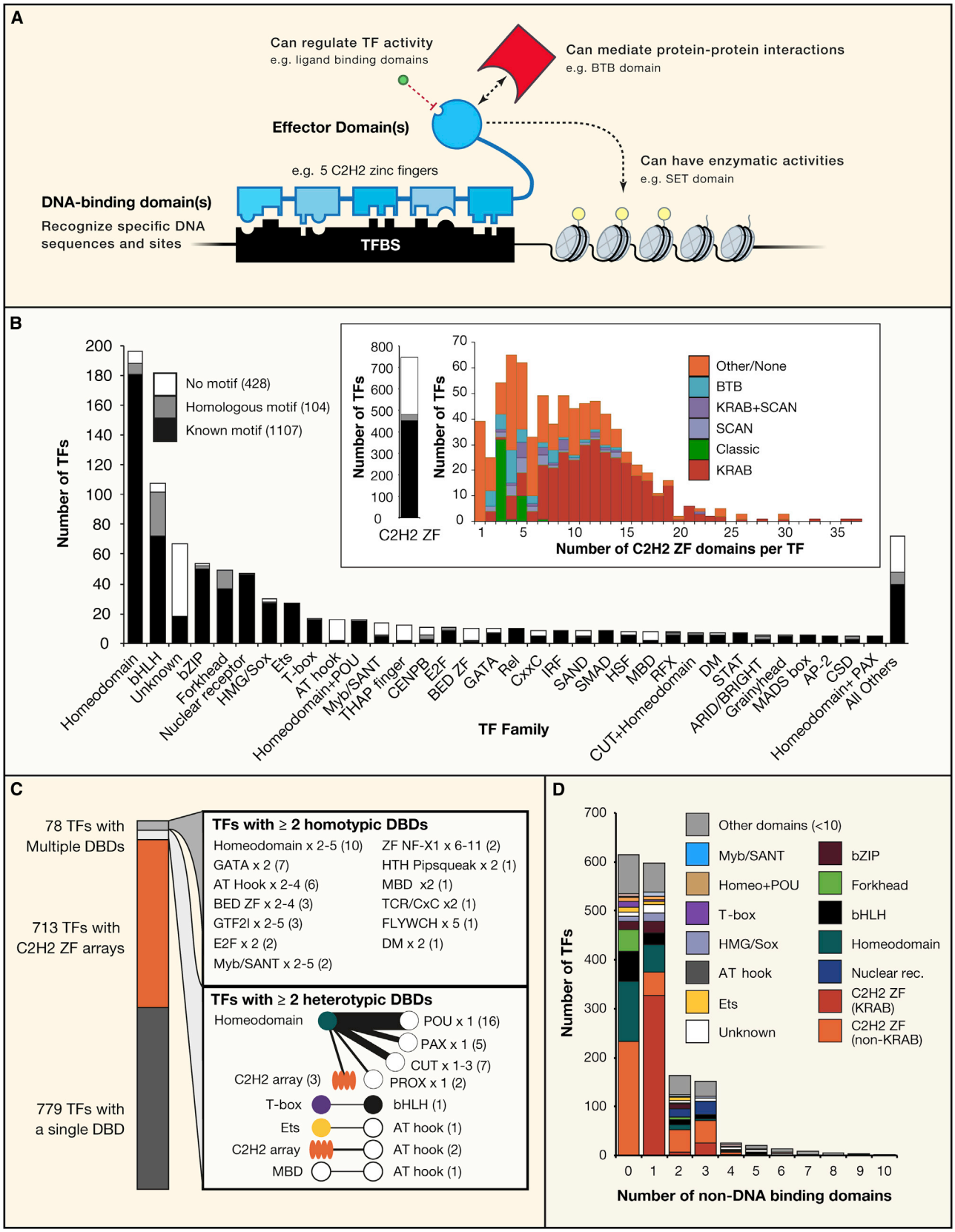
Introduction

Transcription factors (TFs) directly interpret the genome, performing the first step in decoding the DNA sequence. Many function as "master regulators" and "selector genes", exerting control over processes that specify cell types and developmental patterning (Lee and Young, 2013) and controlling specific pathways such as immune responses (Singh et al., 2014). In the laboratory, TFs can drive cell differentiation (Fong and Tapscott, 2013) and even de-differentiation and trans-differentiation (Takahashi and Yamanaka, 2016). Mutations in TFs and TF-binding sites underlie many human diseases. Their protein sequences, regulatory regions, and physiological roles are often deeply conserved among metazoans (Bejerano et al., 2004; Carroll, 2008), suggesting that global gene regulatory "networks" may be similarly conserved. And yet, there is high turnover in individual regulatory sequences (Weirauch and Hughes, 2010), and over longer timescales, TFs duplicate and diverge. The same TF can regulate different genes in different cell types (e.g., ESR1 in breast and endometrial cell lines [Gertz et al., 2012]), indicating that regulatory networks are dynamic even within the same organism. Determining how TFs are assembled in different ways to recognize binding sites and control transcription is daunting yet paramount to understanding their physiological roles, decoding specific functional properties of genomes, and mapping how highly specific expression programs are orchestrated in complex organisms.

This review considers our current understanding of TFs and their global functions to provide context for thinking about how TFs work individually and as an ensemble. We also provide a catalog of the human TF complement and a comprehensive assessment of whether a DNA-binding motif is known for each TF. We use this catalog to survey human TF function, expression, and evolution, highlighting the roles played by TFs in human disease, including the effect of variation within TF proteins and TF-binding sites. A comprehensive review of ~1,600 proteins is impossible; instead, we attempt to exemplify emerging trends and techniques, as well as shortcomings in existing data.

Historically, the term transcription factor has been applied to describe any protein involved in transcription and/or capable of altering gene-expression levels. In the current vernacular, however, the term is reserved for proteins capable of (1) binding DNA in a sequence-specific manner and (2) regulating transcription (Figure 1A) (Fulton et al., 2009; Vaquerizas et al., 2009). TFs can have 1,000-fold or greater preference for specific binding sequences relative to other sequences (Damante et al., 1994; Geertz et al., 2012). Because TFs can act by occluding the DNA-binding site of other proteins (e.g., the classic lambda, lac, and trp repressors [Ptashne, 2011]), the ability to bind to specific DNA sequences alone is often taken as an indicator of ability to regulate transcription.

These proteins cannot be understood functionally without accompanying detailed knowledge of the DNA sequences they bind. TF DNA-binding specificities are frequently summarized as "motifs"—models representing the set of related, short



(legend on next page)

sequences preferred by a given TF, which can be used to scan longer sequences (e.g., promoters) to identify potential binding sites. Determining a DNA-binding motif is often the first step toward detailed examination of the function of a TF because identification of potential binding sites provides a gateway to further analyses. Our ability to generate both motifs and genomic binding sites has improved dramatically over the last decade, leading to an unprecedented wealth of data on TF-DNA interactions. To develop the current TF catalog, we have drawn heavily upon motif collections such as TRANSFAC (Matys et al., 2006), JASPAR (Mathelier et al., 2016), HT-SELEX (Jolma et al., 2013; Jolma et al., 2015; Yin et al., 2017), UniPROBE (Hume et al., 2015), and CisBP (Weirauch et al., 2014), along with previous catalogs of human TFs (Fulton et al., 2009; Vaquerizas et al., 2009; Wingender et al., 2015).

There is typically only a partial overlap between experimentally determined binding sites in the genome and sequences matching the motif; moreover, even experimentally determined binding sites are relatively poor predictors of genes that the TFs actually regulate (Cusanovich et al., 2014). At the same time, motif matches are often among the most enriched sequences in a ChIP-seq (chromatin immunoprecipitation sequencing) dataset, indicating that intrinsic DNA-binding specificity is important for TF binding *in vivo*. In retrospect, this outcome should have been expected: most TF-binding sites are small (usually 6–12 bases) and flexible, so a typical human gene (>20 kb) will contain multiple potential binding sites for most TFs (Wunderlich and Mirny, 2009). Well-established concepts such as cooperativity and synergy between TFs provide a ready solution to this deficit in specificity—most human TFs have to work together to get anything done—but the details of their interactions and relationships are generally lacking. The biochemical effects of TFs subsequent to binding DNA are also largely unmapped and known to be context dependent. As a result, decoding how gene regulation relates to TF-binding motifs and gene sequences remains a major practical challenge; the resulting frustration has been embodied in the term "futility theorem" (Wasserman and Sandelin, 2004).

How Transcription Factors Are Identified

The major TF families in eukaryotes, such as C2H2-zinc finger (ZF), Homeodomain, basic helix-loop-helix (bHLH), basic leucine zipper (bZIP), and nuclear hormone receptor (NHR), were initially described in the 1980s (reviewed in Johnson and McKnight [1989]). Knowledge of binding sites, often identified by methods such as DNase footprinting or mobility shift, led to identification of the particular binding proteins using N-terminal peptide sequencing, phage libraries, or one-hybrid screening. Similarities in amino acid composition and structure were then noted among different DNA-binding proteins. New DNA-binding proteins continue to be identified by experimental methods (e.g., one-hybrid assays [see Reece-Hoyes and Marian Walhout (2012)], DNA affinity purification-mass spectrometry [reviewed in Tacheny et al. (2013)], and protein microarrays [Hu et al., 2009] can screen for new DNA-binding proteins), but today, most known and putative TFs have instead been identified by sequence homology to a previously characterized DNA-binding domain (DBD), which is also used to classify the TF (see Weirauch and Hughes [2011] for review). With the possible exception of the very simple AT-hook (Aravind and Landsman, 1998), all extant examples of DBDs are assumed to be derived from a small set of common ancestors representing the major DBD folds, with the families arising by duplication. There are ~100 known eukaryotic DBD types, which are cataloged in Pfam (Finn et al., 2016), SMART (Letunic et al., 2015) or Interpro (Finn et al., 2017) as hidden Markov models (HMMs), which are used to scan protein sequences for these domains. DBD structures in complex with DNA are currently available in the Protein Data Bank (PDB) (Berman et al., 2000) for most families of human TFs, with AP2, BED-ZF, CP2, SAND, and NRF being notable exceptions. To date, all but a handful of well-characterized mammalian TFs contain a known DBD (Fulton et al., 2009). It is likely that additional DBDs remain to be discovered; for example, extended homologous regions in polycomb-like proteins were recently found to bind motifs containing CG dinucleotides (Li et al., 2017).

Care must be taken when inferring function based only on a homology match to a DBD because not all instances of these domains will necessarily bind specific DNA sequences. The CERS/Lass-type Homeodomains, for example, are not likely to be DNA-binding proteins at all; they instead appear to have been co-opted to function in sphingolipid synthesis (Mesika et al., 2007). Likewise, only a subset of Myb/SANT, HMG, and ARID domain-containing proteins bind specific DNA sequences. In addition, domains with similar names should not be confused. For example, C2H2-ZFs and CCCH-ZFs are structurally and evolutionarily distinct, and while C2H2-ZFs generally bind double-stranded DNA, CCCH-ZFs typically bind single-stranded RNA (reviewed in Font and Mackay [2010]).

Determining TF DNA-Binding Motifs

Motifs are typically displayed as a sequence logo (Schneider and Stephens, 1990), which in turn represents an underlying table or "position weight matrix" (PWM) of relative preference of the TF for each base in the binding site (Stormo and Zhao, 2010). At each base position, each of the four bases has a score, and multiplying these scores for each base of a sequence yields a predicted relative affinity of the TF to that sequence. In many cases, these logos reflect strong preference to one or a small number of related sequences, although they can also represent weak base preferences that nonetheless contribute to binding. In addition, complications can arise that are not captured by a

Figure 1. The Human Transcription Factor Repertoire

(A) Schematic of a prototypical TF.

(B) Number of TFs and motif status for each DBD family. Inset displays the distribution of the number of C2H2-ZF domains for classes of effector domains (KRAB, SCAN, or BTB domains); "Classic" indicates the related and highly conserved SP, KLF, EGR, GLI GLIS, ZIC, and WT proteins.

(C) DBD configurations of human TFs. In the network diagram, edge width reflects the number of TFs with each combination of DBDs.

(D) Number of auxiliary (non-DNA-binding) domains (from Interpro) present in TFs, broken down by DBD family.

PWM: there may be dependencies among base positions (Bulyk et al., 2002; Jolma et al., 2013), for example, due to DNA shape or deformability (Rohs et al., 2009); the TF may have multiple binding modes (e.g., different physical configurations of the protein leading to separate, distinct motifs) (Badis et al., 2009); cooperative interactions may influence the sites bound by a TF (Jolma et al., 2015); or DNA methylation can impact binding, positively or negatively (Yin et al., 2017). To account for these complexities, more complicated models have been developed, e.g., that incorporate preferences to dinucleotides and higher-order k-mers (reviewed in Slattery et al. [2014]), with improvement in accuracy depending on the TF and its family. In many cases, however, the improvement is minor or even undetectable, especially when comparing across different datasets (Weirauch et al., 2013), and the PWM remains the most commonly used model for analysis of TF binding. Hereafter, we use the term "motif" to signify PWM.

The sequence preferences and binding sites of TFs can be assessed by a wide variety of techniques both *in vitro* and *in vivo* (reviewed in Jolma and Taipale [2011]); Table 1 outlines the most prevalent methods and their attributes. As a predictor of relative binding affinity, motifs are most accurately obtained from quantitative affinity measurements for a large number of sequences, preferably using purified proteins and DNA (Stormo and Zhao, 2010). Nonetheless, motifs for many well-studied proteins were initially obtained from very few sequences (e.g., dozens of Sanger reads) and used in thousands of subsequent studies (Mathelier et al., 2016; Matys et al., 2006), illustrating the utility of even approximate descriptions of binding ability.

ChIP-seq (Johnson et al., 2007) has revolutionized the study of TF-binding sites *in vivo* by enabling the genome-wide identification of region occupied by a TF of interest. The semiquantitative measurements obtained have several limitations with regard to motif derivation, however. First, binding is influenced by chromatin state—many TFs bind almost exclusively in open chromatin—as well as biases in the sequence content of the genome. Second, ChIP-seq can clearly detect indirect binding, which can lead to identification of motifs for proteins other than the one ChIPped (Wang et al., 2013; Worsley Hunt and Wasserman, 2014). Third, due to the use of cross-linkers, ChIP does not measure equilibrium binding. Finally, ChIP data is highly dependent on antibody quality—many antibodies cross-react, and ChIP-grade antibodies are not available for many TFs. It is thus often helpful to use prior knowledge regarding the motif expected—for example, the C2H2-ZF "recognition code" (which relates DNA-contacting residues to preferred base positions in the binding site [Najafabadi et al., 2015]) can be used to restrict the analysis to those motifs that resemble computational-based specificity predictions. Some of these issues are in theory addressed by higher resolution approaches such as ChIP-exo (ChIP with exonuclease digestion) (Rhee and Pugh, 2011), but relatively few examples are currently available.

In summary, we now appear to possess the tools needed to identify TF motifs globally. Having these motifs, however, is only a first step in decoding the functions of these proteins in gene regulation; we outline additional complexities in the following sections.

TF Cooperativity and Interactions with Nucleosomes

Both theoretical arguments and practical observations indicate that metazoan TFs must, in general, work together to achieve needed specificity in both DNA binding and effector function—hence the "futility theorem" (Reiter et al., 2017; Wasserman and Sandelin, 2004; Wunderlich and Mirny, 2009). In human, it appears that very few proteins occupy most of their motif matches under physiological conditions; the only clear example out of hundreds that have been examined by ChIP-seq is CTCF, which occupies most of the ~14,000 matches to its ~14-base motif in the human genome with most of the sites occupied across the tested cell types (Fu et al., 2008; Kim et al., 2007). There are myriad ways that TFs are known to collaborate, including aiding each other in binding DNA (cooperative binding) or by impacting chromatin state or transcription through different mechanisms (synergistic regulation). TFs can also bind cooperatively as homodimers (e.g., bZIPs and bHLHs), trimers (e.g., heat shock factors), or higher-order structures (see below). TF interplay is intrinsically related to enhancer function and "logic" (reviewed in Reiter et al. [2017] and Spitz and Furlong [2012]). Here, we mainly consider how cooperative binding is achieved, as it is germane to TF function.

Cooperative binding can occur by several means (reviewed in Morgunova and Taipale [2017]). It is most easily understood when it is mediated by protein-protein interactions, which confer additional stability when two (or more) interacting proteins bind DNA in a compatible spacing and orientation. High-throughput *in vitro* studies indicate that cooperative binding often impacts the sequence preferences of TFs in a complex and can also introduce constraints on intervening sequence between the two binding sites, presumably due to stereochemical requirements (Jolma et al., 2015; Slattery et al., 2011). Results from single-molecule imaging studies confirm that binding sites are occupied longer when multiple TFs bind together (Chen et al., 2014; Gebhardt et al., 2013).

Recent evidence suggests that DNA-mediated cooperative binding also plays an important role in TF function. A test of 9,400 human TF pairs using consecutive affinity purification (CAP)-SELEX identified 315 pairs with clear spacing and orientation preferences between their binding sites (Jolma et al., 2015). Molecular modeling and structural analyses indicated that in some cases cooperativity was due to DNA facilitating contacts between the proteins. In other cases, the proteins bound on the opposite sides of the DNA or relatively far from each other, suggesting that DNA directly mediated the cooperativity. That is, binding of one TF influenced the shape of the DNA in a manner that promoted the binding of the second TF. Indeed, one of the best-studied enhancers, the highly ordered IFNβ enhanceosome, appears to exemplify this mechanism. At this ~50 bp locus, constrained spacing and orientation of binding sites for eight TFs facilitates interactions, allowing for the recruitment of three non-DNA binding cofactors. Structural analysis, however, reveals relatively few contacts among the TFs (Panne, 2008), with stability conferred instead by induced changes in DNA structure and interactions with cofactors. DNA-mediated cooperative binding for TFs bound within ~10 bases of each other can also be mediated by DNA vibrational modes, which is predicted to occur to

Table 1. Experimental Methods for Determining and Validating TF-Binding Specificities

		Method	Description	Features			
				Capability of **motif discovery** (approx. length in base pairs)	Identifies **genomic binding locations** of a TF	Can measure effect of **CpG Methylation**	Can measure **cooperative binding and/or multimers**
High-throughput	*In Vitro Methods*	Protein Binding Microarray (PBM)	A GST-tagged TF is bound to a glass slide that has ~41,000 spots of short immobilized DNA sequences. Fluorescence-based detection of bound spots and k-mer enrichment analysis yields motifs.	✔ (< 12 bp)	✘	✔ Methyl-PBM	✔
		Bacterial one-hybrid	TF binding sites are selected in bacterial cells from a randomized library that is cloned in front of selectable marker genes. Can be reversed to select proteins able to bind a constant DNA sequence using a library of variant protein sequences.	✔ (< 14 bp)	✘	✘	✘
		SELEX-based methods	Systematic evolution of ligands through exponential enrichment (SELEX) involves adding TFs to a DNA pool containing many randomized sequences and selecting for binding in multiple rounds. Related methods include HT-SELEX, SELEX-seq, and Bind-n-Seq. Selection can be performed using affinity tags, or molecular trapping on a microfluidic platform (SMiLE-seq).	✔ (< 25 bp)	✘	✔ Methyl-HT-SELEX	✔ CAP-SELEX ✔ SMiLE-seq
Mid-throughput		DAP-seq	Single step SELEX using a library of fragmented genomic sequences. Sequence diversity is less than HT-SELEX, but genomic sequences that have co-evolved with the TF are included.	✔ Limited by skewed distribution of genomic sequences	✔ Peaks are not necessarily indicative of *in vivo* binding	✔ AmpDAP-seq	✘
		HiTS-FLIP	Uses an Illumina sequencer's flowcell as a PBM chip to measure binding to orders of magnitude more DNA sequences.	✔ (< 17 bp)	✘	✘	✘
		Spec-seq	Single step SELEX with a microarray synthesized library. The lower complexity library is useful for quantitatively measuring effects of binding site mutations using sequencing.	✔ Limited by number of sequences assayed	✘	✔ Methyl-Spec-seq	✘
		MITOMI	A microfluidic device is used to isolate DNA-protein complexes from free DNA instantaneously to accurately measure the relative binding affinities of TFs to ~10,000 individual sites.		✘	✘	✘
Low-throughput		EMSA	Tests if a DNA sequence is bound by a protein by observing a shift in the electrophoretic migration of DNA.	✘ Useful for validating known binding sites	✘	✔	✔ EMSA-FRET
		DNA footprinting	DNA is incubated with a TF and then degraded using DNase-I, resulting in cuts in all positions except those that were protected by the bound TF.		✘	✔	✔
		ITC, SPR, MSTP	Isothermal titration calorimetry (ITC), Surface plasmon resonance (SPR) and Microscale thermophoresis (MSTP) measure the binding affinity of TF-DNA interactions.		✘	✔	✔
	In Vivo Methods	ChIP-based assays	Proteins are crosslinked to DNA using formaldehyde and precipitated with an antibody. Bound DNA is detected with qPCR, microarray (ChIP-chip), or sequencing (ChIP-seq). The ChIP-Exo variant incorporates exonuclease treatment to enhance resolution.	✔ Limited by skewed distribution of genomic sequences, and inability to distinguish direct from indirect binding	✔	✔ ChIP + Bisulfite-sequencing	✔ Re-ChIP
		DamID-seq	A TF is expressed in mammalian cells as a fusion to bacterial Dam-methylase. The enzyme methylates a consensus sequence in close proximity to the TF's binding sites, which can be mapped using restriction enzymes and high-throughput sequencing.		✔	✘	✔ Split DamID-seq

In vitro and *in vivo* methods currently used to experimentally derive and confirm TF-binding sites and motifs.

some extent between all possible pairs of TFs (Jolma et al., 2015; Kim et al., 2013).

In order to bind to nucleosomal DNA, TFs must either compete with nucleosomes or interact with nucleosomes or nucleosomal DNA in some way to access their sites. TFs can inherently cooperate with each other to compete with nucleosomes (Adams and Workman, 1995; Polach and Widom, 1996), and indeed, binding sites identified in ChIP-seq are often biased toward homotypic clusters, especially when low-affinity motifs are considered (Gotea et al., 2010). In addition, some TFs can initiate the displacement of nucleosomes or at least change their conformations (e.g., Foxa1 [Iwafuchi-Doi et al., 2016; Swinstead et al.,

2016a]), most likely by recruiting ATP-dependent chromatin remodelers and other TFs (reviewed in Swinstead et al. [2016b]). The activity of these TFs may also be dependent on their ability to bind nucleosomal DNA, which can be influenced by the rotational positioning of the binding site on the nucleosome (e.g., the Yamanaka factors POU5F1, SOX2, KLF4, and MYC [Soufi et al., 2015]). An additional intriguing observation is that different chromatin remodelers possess preferences for specific DNA sequences and/or nucleosome conformations (Rippe et al., 2007), suggesting that both nucleosomes and nucleosome-positioning mechanisms impart additional DNA-sequence specificity to TF action.

TF Effector Functions

TFs vary dramatically in how they impact transcription upon DNA binding. Some human TFs (e.g., TBP) can directly recruit RNA polymerase, while others recruit accessory factors that promote specific phases of transcription (reviewed in Frietze and Farnham [2011]). As in bacteria, human TFs can lack a specific effector function and instead act by steric mechanisms, which can be as simple as blocking other proteins from binding to the same site (Akerblom et al., 1988). Most eukaryotic TFs, however, are thought to act by recruiting cofactors (Reiter et al., 2017). Such "coactivators" and "corepressors," initially identified as mediators of TF effector activity, are frequently large multi-subunit protein complexes or multi-domain proteins that regulate transcription via several mechanisms. They commonly contain domains involved in chromatin binding, nucleosome remodeling, and/or covalent modification of histones or other proteins, including TFs and RNA polymerase (Frietze and Farnham, 2011). The IFNβ enhanceosome is a classic illustration of coactivator recruitment, with the binding of multiple TFs resulting in the recruitment of GCN5/KAT2A and CBP/p300 histone acetyltransferases (reviewed in Panne [2008]). The resulting changes to the local chromatin environment recruit nucleosome remodelers such as the SWI/SNF complex to create room for RNA polymerase to bind and initiate transcription. Some coactivators and corepressors appear to be more widely used than others. p300 is often used as a marker of enhancers (Visel et al., 2009), associating with dozens of TFs (Frietze and Farnham, 2011). The Mediator complex, which bridges TFs and RNA polymerase II, is similarly associated with thousands of loci—possibly the majority of transcribed genes (Kagey et al., 2010)—and is recruited by dozens of TFs (Malik and Roeder, 2010).

Dedicated effector domains often mediate the recruitment of specific cofactors by TFs. The KRAB domain, for instance, is found in ~350 human C2H2-ZF proteins. It recruits TRIM28/KAP1, which in turn recruits HP1/CBX5 and SETDB1, catalyzing deposition of the repressive H3K9me3 histone mark (reviewed in Ecco et al. [2017]). Likewise, ligand-binding domains of nuclear hormone receptors facilitate interactions with coactivators, corepressors, and other TFs in a ligand- and context-dependent manner (reviewed in Rosenfeld et al. [2006]). Many TFs do not contain well-defined effector domains, however. Some are comprised almost entirely of a single DBD and are thus unlikely to contain separable activation domains, especially in the bZIP (e.g., BATF, CREBL2, and MAFK) and bHLH (e.g., MAX, NHLH1, and ATOH7) families. Classical transcriptional activator sequences present in well-studied proteins (e.g., the acidic sequences found in TP53, E2F, and SP1) are often unstructured low-complexity sequences with small functional regions dubbed short linear motifs (Garza et al., 2009). The LxxLL motif, for instance, was originally identified as a protein-protein interaction interface of nuclear hormone receptors with their cofactors (NCoA, CBP, Mediator, etc.) but is also present in unrelated TF families (e.g., Myb/SANT and STAT) (Plevin et al., 2005). Many of the best-characterized C2H2-ZF TFs are also known to exploit unstructured regions and/or DBDs to interact with cofactors (Brayer and Segal, 2008).

TFs have traditionally been classified as either "activators" or "repressors"; however, this notion has been repeatedly questioned. Many TFs can recruit multiple cofactors that have opposite effects (Frietze and Farnham, 2011; Rosenfeld et al., 2006; Schmitges et al., 2016), dependent on the local sequence context and availability of cofactors (Meijsing et al., 2009; Wong and Struhl, 2011). MAX, for example, functions as an inhibitor when binding to DNA as a heterodimer with MNT or MXD1 and as an activator when binding as a heterodimer with MYC (reviewed in Amati and Land [1994]). A recent study used a complex pool of >4 million sequences to survey the effect on gene expression of the relative positions of various TF-binding sites in diverse contexts, uncovering numerous motifs capable of both activation and repression in the same cell type (Ernst et al., 2016).

Because effects on transcription are so frequently context dependent, more precise terminology may be warranted, in general—for example, reflecting the biochemical activities of TFs and their cofactors. On a global level, however, there is no comprehensive catalog of cofactors recruited by TFs. Moreover, the biochemical functions required for gene activation or communication between enhancers and promoters remain largely unknown (Zabidi and Stark, 2016). As many as 443 different chromatin modification proteins have been cataloged in human, and many interactions among cofactors and chromatin proteins have been described (e.g., Marcon et al. [2014]). But, the same studies detected few TFs, suggesting that TF-cofactor interactions are weak/transient or that relative stoichiometry is skewed against TFs. Given the large number of factors involved, it is conceivable that a complex network of thousands of interactions among TFs and cofactors exists, providing a ready explanation for context dependency.

The Human TF Repertoire

A key starting point in the global analysis of human TFs and gene regulation is a simple index of high-confidence human TFs and what is known about them. There is no one-size-fits-all solution to automate the generation of such a list: domain structures do not perfectly predict TFs, the literature is highly heterogeneous, and electronic annotations are non-uniform. To our knowledge, the latest comprehensive reviews of human TFs were published in 2009 (Fulton et al., 2009; Vaquerizas et al., 2009). Fulton et al. curated a list of putative mouse and human TFs based on evidence of TF activity, including both DNA binding and regulation of transcription, identifying a total of 535 human TFs. Vaquerizas et al. annotated putative DBDs and proteins that contain them with confidence levels based on selectivity for known TFs and their likelihood of involvement in transcription. This list was

then appended with Gene Ontology and TRANSFAC TF annotations to yield a total of 1,391 human TFs. In recent years, the field has advanced substantially with dramatic expansions in data collection, including hundreds of motifs generated *in vitro* (Badis et al., 2009; Jolma et al., 2013; Wei et al., 2010; Weirauch et al., 2013, 2014; Yin et al., 2017). There have also been updates to gene annotations. We therefore undertook a revised manual curation of the human TF collection, which forms the basis of the remainder of this review.

The overall approach is depicted in Figure S1A. We manually examined 2,765 proteins compiled by combining putative TF lists from several sources: the aforementioned papers (Fulton et al., 2009; Vaquerizas et al., 2009), domain searches (using HMMs and parameters from CisBP [Weirauch et al., 2014] and Interpro, as well as the TRANSFAC-related database TFClass [Wingender et al., 2015]), Gene Ontology, and crystal and NMR structures of proteins in complex with DNA taken from the PDB (Berman et al., 2000). We created a web page for each protein containing all relevant information and links to external databases. We then assigned two curators (among the authors of this manuscript) to classify the protein's status as a TF ("TF with a known motif," "TF with a motif inferred from a close homolog," "likely TF" [due to presence of a DBD or literature information], "ssDNA/RNA binding protein," or "unlikely TF"), and its DNA-binding mode (binds as a monomer or homomultimer, binds as an obligate heteromer, binds with low specificity, or does not bind DNA). Curators could also submit notes and citations supporting their assessments. Using data from CisBP and other sources, we recorded whether motifs are known for each TF (or a close homolog) along with the availability of a protein-DNA structure. We considered global sequence alignments and known DNA-binding residues to make decisions for poorly characterized proteins within families where only a subset binds DNA (e.g., ARID, HMG, and Myb/SANT). To make the task feasible, we did not explore or record complexities such as protein modifications or binding partners. Three senior authors (T.R.H., M.T.W., J.T.) resolved cases of disagreement between reviewers and manually reviewed all cases where both curators agreed that a protein without a canonical DBD is a likely TF. Table S1 contains the full curation results. The "HumanTFs" website (http://humantfs.ccbr.utoronto.ca/) displays the results, with a separate page for each TF, along with all known motifs and information and sequence alignments for each DBD type. The site also has an option for users to submit additional information.

The final tally encompasses 1,639 known or likely human TFs. Most contain at least one of only two DBD types (C2H2-ZFs [747] and Homeodomains [196]). Nearly half of the remainder (46%) are accounted for by an additional six (bHLH [108], bZIP [54], Forkhead [49], nuclear hormone receptor [46], HMG/Sox [30], and ETS [27]) (Figure 1B). There are far fewer Myb/SANT and HMG domain TFs than previously estimated (Vaquerizas et al., 2009) (14 versus 38 and 40 versus 55, respectively) after accounting for known subclasses that lack DNA-sequence specificity. The vast majority (93%) of the 1,639 TFs are known or expected to bind DNA as either a monomer or homomultimer. Many contain multiple copies of the same DBD type (Figure 1C), but most of these are C2H2-ZFs, which bind DNA as an array (Figure 1A). The number of C2H2-ZFs per protein varies substantially, depending partly on the effector domain (Figure 1B). The large numbers of C2H2-ZFs in the KRAB-containing subtype may be due to the specificity required to target individual transposable elements (see below). Only a small fraction of TFs (47, or ~3%) contain more than one type of DBD, with POU:Homeodomain being the most prevalent (Figure 1C). Most human TFs also contain additional protein domains (Figure 1D): in total, 391 different types of non-DNA-binding domains are represented, consistent with the notion of a diverse and extensive network of TF effector functions.

This survey includes 348 TFs not included in the Vaquerizas list. Notable additions include 134 C2H2-ZFs, 22 bHLHs, 14 AT-hooks, 13 Homeodomains, and the 12 recently described THAP finger proteins [Campagne et al., 2010]) (Figure S1B). The individual proteins in previous lists are, however, almost completely reconfirmed. 1,292 out of the 1,391 proteins (93%) identified by Vaquerizas et al. were also in our compilation, with 50 removed due to changes in gene annotations (pseudogenes and duplicates) and 49 removed using the guidelines above. Likewise, 98% of the TFs identified by manual curation by Fulton et al. (523/535) are considered to be TFs in our study.

It is likely that our current TF list is still incomplete, and entire DBD families may remain undiscovered. Indeed, 69 of the TFs in our list are categorized as "Unknown family," due to the lack of a canonical DBD. Most of these proteins lack motifs (see below), crystal structures are largely unavailable, and the evidence for DNA binding typically includes only a handful of sequences identified in a single manuscript. Thus, TFs in this category should be treated with caution until further experimental data are available.

In addition, some known DBD families might be larger than is currently appreciated. For example, the simple AT-hook domain (represented by a 13 amino acid [aa] consensus) is predicted to be present in 3 and 21 human genes according to the Interpro and SMART databases, respectively. A more lenient definition, however, requiring only the presence of a GRP tripeptide flanked by multiple basic residues over a 22-base window (Aravind and Landsman, 1998) is present in hundreds of human proteins, each of which could represent a bona fide TF. The set of C2H2-ZFs will also warrant revisiting as better models emerge for recognizing these short (~23 aa) domains and distinguishing those involved in DNA binding from those facilitating interactions with RNA or other proteins (Brayer and Segal, 2008), although most do appear to bind DNA in large surveys (Imbeault et al., 2017; Schmitges et al., 2016).

Sequence Specificities of the Human TFs

Roughly three-quarters (1,211) of the human TFs currently have a binding motif (1,107 "known," i.e., measured experimentally, and a further 104 inferred from a closely-related homolog) (Weirauch et al., 2014). 913 of the known motifs were obtained from high-throughput *in vitro* assays such as HT-SELEX or PBM and hence provide a profile of their intrinsic relative preferences to many DNA sequences. Figure 1B illustrates that most classes of TFs have high or complete motif coverage, while a handful have major gaps. Almost all Homeodomains (188/196), for example, have a known or inferred motif, likely due to their relative ease of study *in vitro* and their deep conservation

enabling inference by homology. The C2H2-ZF class, in contrast, currently lacks hundreds of motifs (267/747) (Figure 1B, inset), possibly because they are difficult to study *in vitro* (many are large proteins) and relatively few are well conserved (Stubbs et al., 2011). By proportion, the AT-hook proteins, THAP finger, BED-ZF, and those with no known DBD are also poorly characterized.

Among the 1,107 proteins with a known motif, less than 2% (19) lack a canonical DBD, with only 6 of 69 such proteins having an *in vitro* derived motif—the other 13 are based on experiments such as ChIP-seq and thus may describe binding through a cofactor. Nevertheless, the additional 50 non-canonical TFs were included in our list due to some evidence for direct sequence-specific DNA binding. An example of a bona fide non-canonical TF is NRF1, which was initially characterized in 1993 (Virbasius et al., 1993), with further high-throughput characterization occurring 20 years later (Jolma et al., 2013). Some of the likely TFs that do not contain a canonical DBD are obligate heterodimers that contribute to protein–DNA contacts in crystal structures of sequence-specific protein complexes but are unlikely to bind DNA on their own (e.g., NFYB and NFYC, which form a trimeric complex with NFYA [Nardini et al., 2013]).

Many TFs recognize similar motifs, typically corresponding to TF families or subfamilies, consistent with intuition and with many previous studies (e.g., Badis et al. [2009] and Wei et al. [2010]) (Figure 2A). Notably, C2H2-ZF proteins contribute most of the diversity to the motif collection (Figure 2B) as expected from previous studies and from the diversity in their DNA-contacting residues (Emerson and Thomas, 2009; Imbeault et al., 2017; Najafabadi et al., 2015; Schmitges et al., 2016; Stubbs et al., 2011). Figure 2C shows motifs for the NHR family, illustrating that TF diversity can involve changes in both monomeric DNA-sequence preference and protein-complex formation—many motifs in Figure 2C are recognized by dimers. In total, over 500 motif specificity groups are present in human (Table S2), indicative of the wide range of DNA sequences capable of functioning as human TF-binding sites.

Conservation and Evolution of Human TFs

Evolution of TFs is typically much slower than evolution of their regulatory sites. TF orthologs between human and *Drosophila* often display virtually identical sequence specificity (Nitta et al., 2015). Physiological roles of TFs are also often conserved—the HOX proteins, which specify the anterior-posterior body plan, are perhaps the best-known example (Bürglin, 2011)—but there are numerous others, e.g., the regulation of cilia genes by RFX TFs (Choksi et al., 2014). Nonetheless, TFs do evolve, changing their motifs, binding partners, and expression patterns (Arendt et al., 2016; Grove et al., 2009; Lynch and Wagner, 2008; Schmitges et al., 2016). A striking example of duplication and divergence among human TFs is the hundreds of KRAB-containing C2H2-ZF proteins encoded by most mammalian genomes, many of which display hallmarks of diversifying selection (Emerson and Thomas, 2009) with complex orthology patterns even between human and mouse (Huntley et al., 2006). In human, KRAB C2H2-ZF proteins generally bind transposable elements (TEs) (mainly LINEs and endogenous retroviruses), presumably silencing them, at least initially, via the repressive function of the KRAB domain (Imbeault et al., 2017; Jacobs et al., 2014; Rowe et al., 2010; Schmitges et al., 2016). An "arms race" between the TEs and TFs provides a ready explanation for their rapid diversification. A "domestication" model is also supported, however, in which the KRAB-TE interaction is evolutionarily maintained to co-opt the TE for host gene regulation long after TEs degrade beyond pathogenic potential (Imbeault et al. [2017], reviewed in Ecco et al. [2017]).

Based on their distribution across eukaryotic genomes (Figure 3A), the 1,639 TFs in our updated catalog fall into major groups with close relatives extending to metazoans, vertebrates, tetrapods, placental mammals, or primates. Strikingly, nearly all Homeodomain proteins have recognizable counterparts across vertebrates, while virtually all of the mammal-specific proteins contain C2H2-ZFs. Indeed, the divergence times between Ensembl-defined human TF-TF paralogs display a bimodal division: a first wave of duplications across diverse TF families occurred at the base of Bilateria, and a second wave of duplications, dominated by KRAB C2H2-ZFs, began in Amniota (Figure 3B, left). The earlier wave, with duplications across diverse TF families, is consistent with the postulation that two rounds of whole-genome duplication occurred at or near the base of vertebrates (Dehal and Boore, 2005). This event is roughly coincident with the expansion of cell-type diversity, possibly facilitated by duplicated TFs available to regulate novel cell types (Nitta et al., 2015; Arendt et al., 2016). The expansive KRAB radiation may be partly explained by the increased opportunity for retroviral transmission facilitated by the placenta (Hayward et al., 2015). Remarkably, TF-TF duplications during the KRAB radiation era dominate the distribution of all human paralog pairs arising over the last 300 million years (Figure 3B, right).

Expression of Human TFs across Tissues and Cell Types

Tissue- and cell-type-specific expression of genes, including TFs, is often indicative of corresponding specific functions. We examined expression patterns for 1,554 TFs detected in 37 adult tissues using RNA-seq (RNA sequencing) data from the Human Tissue Atlas (Figure 4A), adopting its quantitative definitions for tissue specific expression (tissue enriched, group enriched, or tissue enhanced) (Uhlén et al., 2015). This global view of gene expression patterns captures known roles for many well-characterized TFs. For example, SOX2, OLIG1, and POU3F2 (OCT7) are expressed almost exclusively in the cerebral cortex, and GATA4 and TBX20 are highly expressed only in cardiac muscle. Roughly one-third (543) of the human TFs in this dataset displayed tissue-specific expression, including many with poorly characterized physiological roles.

Comparing between TF classes, a striking trend emerges, mimicking the evolutionary sequence analysis above. C2H2-ZFs are markedly depleted for tissue specificity—only 19% versus 49% for other types of TFs ($p < 10^{-13}$, Bonferroni-corrected Fisher's Exact Test) (also visible at right in Figure 4A). Only 12% (41/339) of KRAB-containing C2H2-ZFs are tissue specific, possibly due to their role in the repression of transposable elements, which may be beneficial broadly across cell types. The majority are testes-specific (26/41), consistent with

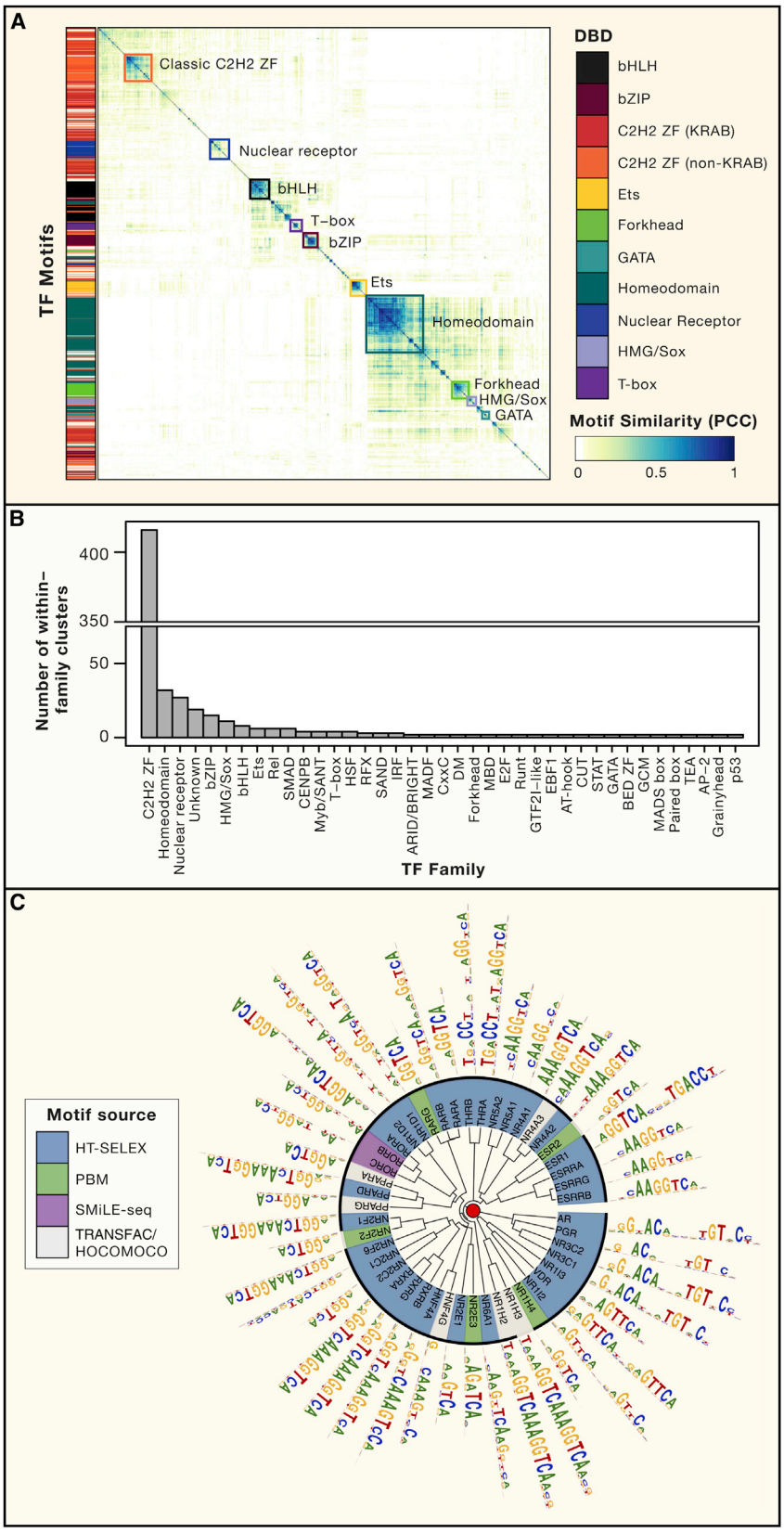


Figure 2. DNA-Binding Specificities of the Human Transcription Factors

(A) Heatmap showing similarity of human TF DNA binding motifs. Representative motif(s) were selected for each TF from the set of motifs directly determined by a high-throughput *in vitro* assay. Pairwise motif similarities were calculated using MoSBAT energy scores (Lambert et al., 2016) and arranged by hierarchical clustering using Pearson dissimilarity and average linkage.

(B) Motif diversity within each family, as measured by the number of clusters supported by the optimal silhouette value (Lovmar et al., 2005).

(C) Detailed view of representative motifs for nuclear hormone receptors, displayed on a phylogram according to DBD sequence similarity using motifStack (Ou et al., 2018).

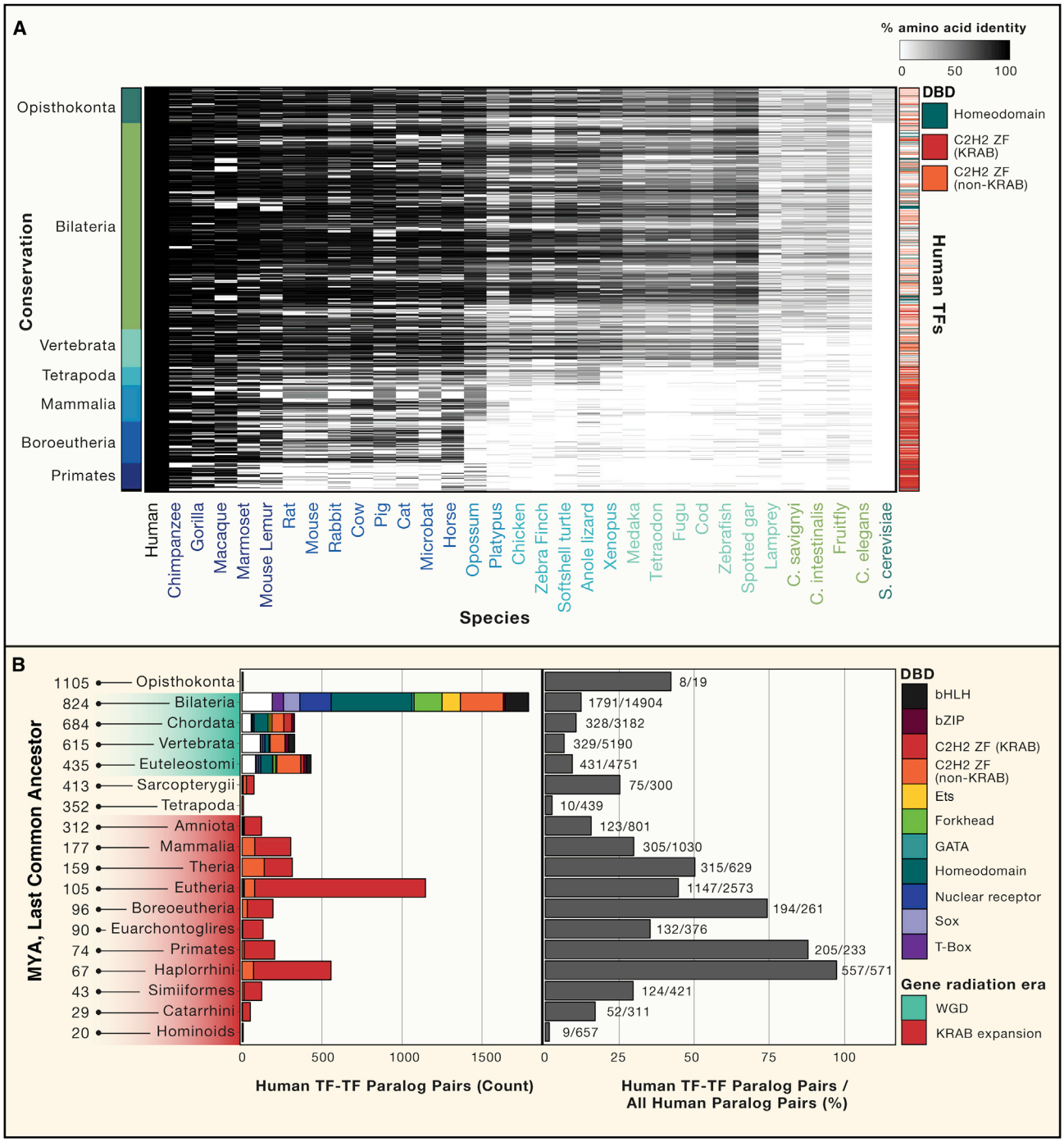


Figure 3. Orthologs and Paralogs of the Human Transcription Factors

(A) Presence and absence of human TF orthologs across eukaryotic species. Amino acid percent identity is plotted for the most similar non-human TF gene in 32 eukaryotic species (from Ensembl Compara database [Herrero et al., 2016]). TFs are ordered first by conservation level (approximated gene age), based on similarity to expected conservation patterns for each of the clades plotted. For an interactive version of this panel, see http://www.cell.com/cell/9995.

(B) Left: Number of human TF-TF paralog pairs that diverged in each clade shown. Right: Proportion of all human paralog pairs from each clade that are a TF-TF pair.

a role for KRAB C2H2-ZFs in retroelement silencing during gametogenesis (Ecco et al., 2017). Homeodomain TFs, in contrast, are highly enriched for tissue-specific expression (133/162, 82%, $p < 10^{-13}$) and are also the only group overrepresented in the list of TFs that is not detected in the Human Tissue Atlas dataset (34/84; $p < 10^{-7}$), presumably reflecting

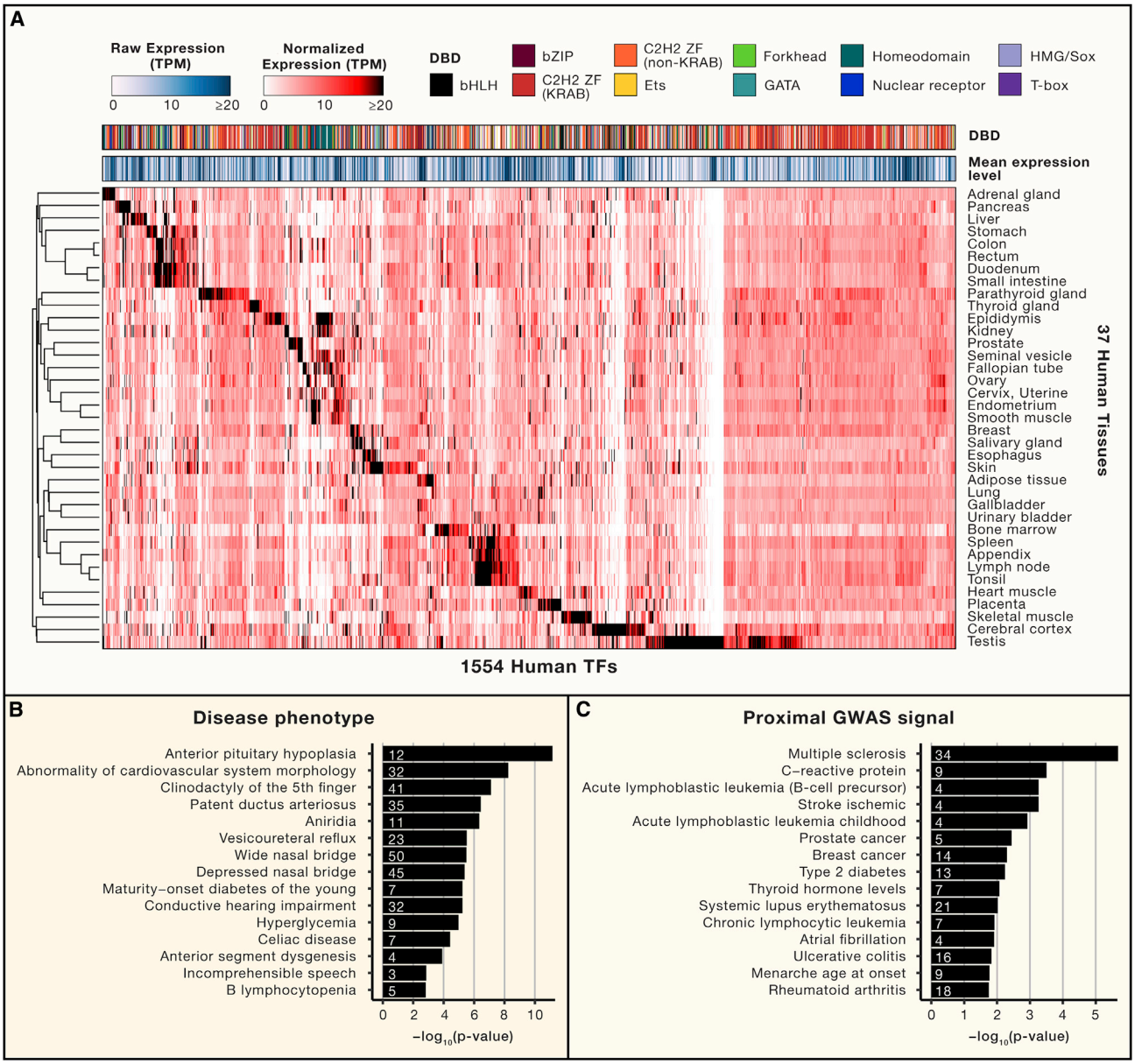


Figure 4. Functional Properties of the Human Transcription Factors

(A) RNA-seq gene expression profiles for 1,554 human TFs across 37 human tissues (from the Human Tissue Atlas version 17 [Uhlén et al., 2015]), normalized by row and column. Tissues and TFs are arranged using hierarchical clustering by Pearson correlation. Mean expression level indicates the mean pre-normalization mRNA expression level of each TF (in TPM) across all tissues in which the TF was expressed (TPM $\geq$ 1). For an interactive version of this panel, see http://www.cell.com/cell/9995.

(B) TF gene set over-representation for human disease phenotypes (Köhler et al., 2014). y axis indicates the significance of the size of the intersection between the set of human TFs and the indicated gene set. Values indicate the number of TFs in the gene set.

(C) Diseases with GWAS signal ($p < 5x10^{-8}$) located proximal to TF-encoding genes. Loci containing multiple variants were restricted to the single most strongly associated variant, and subsequently expanded to incorporate variants in strong linkage disequilibrium (LD) ($r^2 > 0.8$) with this variant using Plink (Purcell et al., 2007). The full set of genetic variants and sources for each disease are provided in Tables S3 and S4. Each resulting variant was assigned to its nearest gene, creating a gene set for each disease. For each gene set, the significance of its overlap with the list of human TFs was estimated using the hypergeometric distribution. p values were corrected using Bonferroni's method. Values indicate the number of TF-encoding loci associated with the given disease.

well-established roles in early embryonic-cell-fate specification and/or roles in the maintenance and differentiation of specialized cell types (Bürglin, 2011; Dunwell and Holland, 2016).

Across all other TF families, half (49%) are tissue specific, providing a clue as to their specific physiological functions. Higher-resolution data—e.g., from single-cell RNA-seq, which can resolve the different cell types that comprise tissues—will almost certainly lead to a more refined view of the associations between TFs, cell identity, and the genes regulated by the TFs.

Human TFs in Genetics and Disease

TFs represent ~8% of all human genes and are associated with a wide array of diseases and phenotypes. TF mutations are often highly deleterious, presumably explaining why genomic loci-encoding TFs are enriched for ultraconserved elements (Bejerano et al., 2004) and depleted of common variation within their DBDs (Barrera et al., 2016). The genetic analysis of TFs can be complicated by functional redundancies inherent to gene regulatory networks because phenotypes might be difficult to detect or manifest only under specific conditions or because variants with highly deleterious effects will be absent at the population level. Nonetheless, a global perspective on human TFs in clinical phenotypes does reveal common themes. Figure 4B illustrates human disease phenotypes that involve a significant number of mutations within or near genes encoding TFs, as compiled by The Human Phenotype Ontology (Köhler et al., 2014). The strongest enrichment is observed for anterior pituitary hypoplasia, which occurs in association with congenital growth-hormone deficiency—of the 15 genes known to be involved in this phenotype, 12 are TFs ($p < 10^{-11}$), including multiple Homeodomain and Sox family TFs. Overall, 313 (19.1%) of the human TFs are currently associated with at least one phenotype, a significantly higher fraction than that observed for all genes (16.2%) ($p = 0.002$, proportions test). In contrast, TFs are depleted from the core set of essential genes in human cancer cell lines, based on data from recent CRISPR screens (3% versus 10% (Hart et al., 2015)), perhaps because the human TF repertoire is utilized mainly for developmental or tissue-specific functions. Phenotypes have been associated with genetic perturbations of 304 (18.6%) of the 1,198 one-to-one human/mouse TF orthologs in mouse (Blake et al., 2017), often yielding phenotypes that are consistent with the TF's known function in human. For example, six of the ten Rel family TFs result in "decreased B cell proliferation."

Genome-wide association study (GWAS) signals for some polygenic diseases are also enriched for loci-encoding TFs (Figure 4C). Many of these diseases have a strong immune-dependent component, suggesting a prominent role for the many immune-responsive TFs (reviewed in Smale [2014]). In addition, many individual TF loci harbor strong GWAS signals for multiple diseases. For example, variants within the loci encoding the Ikaros-family C2H2-ZFs IKZF1 and IKFZ3, which play critical roles in the adaptive immune response (John and Ward, 2011), reach genome-wide significance in ten different GWAS studies; most of these studies involve autoimmune diseases with strong B and T cell-specific genetic signals (Hu et al., 2011).

The modular structure of TFs facilitates identification of the mechanistic impact of mutations. DBD mutations can alter sequence specificity; such mutations in HOXD13 have been associated with limb malformations (Barrera et al., 2016). Profound effects on gene expression can also result from mutations located outside of the DBD. For example, multiple variants within the TP53 protein affect its activity by altering protein interactions (reviewed in Muller and Vousden [2013]). In cancer, chromosomal abnormalities can create onco-fusion proteins with novel functions, such as the Ets factors ERG and FLI1 fusing with the RNA-binding protein EWSR1 (Sizemore et al., 2017). Similarly, as for any gene, a mutation can fall within a regulatory region controlling the expression of a TF, ultimately resulting in altered TF function. For instance, weakening of a TCF7L2 (TCF-4)-binding site within an enhancer that drives expression of *MYC* can decrease risk for tumorigenesis in the colon (reviewed in Sur and Taipale [2016]).

TFs are unique as a gene class in that they represent the proteins whose binding sites are impacted by variation or mutation in regulatory DNA. Numerous such examples have been established, covering a wide range of TF families and diseases (reviewed in Deplancke et al. [2016]). For example, an intronic obesity-associated polymorphism in the *FTO* locus alters enhancer function by modulating the binding of ARID5B, leading to an increase in *IRX3* and *IRX5* expression, changing adipocyte cell fate and overall mitochondrial thermogenesis in adipose tissue (Claussnitzer et al., 2015). Deeper knowledge of how TFs find their targets and control gene expression patterns will be vastly beneficial for our understanding of the estimated 85%–93% of common disease-associated genetic variation that is likely to impact gene regulation (Hindorff et al., 2009; Maurano et al., 2012).

Perspective: Learning to Read the Genome

In 2003, Eric Lander presented a seven-word nano-lecture summary: "Genome: bought the book, hard to read," emphasizing the difficulty of mechanistic interpretation of DNA sequence (https://www.improbable.com/airchives/paperair/volume9/v9i6/nano/nano_6.html). 15 years later, the task of interpreting the function of noncoding sequence is still challenging—the "futility theorem" still holds. As an illustration, it is now known that many TFs bind preferentially within open chromatin, but the open chromatin itself is presumably controlled by TFs, and there is currently no algorithm that predicts open chromatin directly from sequence with both high sensitivity and precision: a leading model achieves 20%–35% sensitivity at a 20% false discovery rate (Kelley et al., 2016) and is most effective at identifying promoters.

This ongoing challenge can no longer be explained by a general lack of motifs for known TFs (Table S1). A clear hurdle to be addressed now is how to learn relevant combinations of binding sites and other sequence features. On a global scale, TF-TF cooperativity and TF-nucleosome interactions are largely unmapped, although both are likely to be prevalent. Because the number of factors involved is high, the number of functionally interacting combinations may be astronomical—the limited size of the human genome will likely pose challenges for the systematic detection of such higher-order interactions, due to a lack of statistical power.

Most of the functional DNA in the genome is likely regulatory (Kellis et al., 2014), with TFs playing a central role in its recognition and utilization. There is a clear role for TFs in many human diseases, highlighting the importance of continued efforts for understanding TF-mediated gene-regulatory mechanisms. Other current challenges include addressing synergy and redundancy among multiple elements regulating the same gene, predicting enhancer-promoter contacts, the relevance of large-scale arrangement of regulatory features along chromosomes and in three dimensions, and various types of epigenetic memory.

Computational methods examining these themes are a topic of ongoing research, and experimental techniques probing the role of TFs in nucleating and mediating these phenomena likewise continue to be developed. These advances will be instrumental in conquering what is likely to be the next frontier in human genetics: decoding the genome the way TFs do.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure and four tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.01.029.

ACKNOWLEDGMENTS

We apologize to colleagues whose important primary studies could not be cited due to space constraints. This work was supported by NIH R01 NS099068-01A1, Lupus Research Alliance "Novel Approaches," Cincinnati Children's Hospital "Center for Pediatric Genomics pilot study," "Trustee Award," and "Endowed Scholar award" (M.T.W.), and a CIHR Foundation Award (T.R.H.). A.J. was supported by Swedish Research Council "Vetenskapsrådet" postdoctoral grant (2016-00158). S.A.L. and L.F.C. are supported by NSERC PGS-D scholarships. T.R.H. is a Senior Fellow of CIFAR and the Billes Chair of Medical Research at the University of Toronto.

REFERENCES

Adams, C.C., and Workman, J.L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol. Cell Biol. *15*, 1405–1421.

Akerblom, I.E., Slater, E.P., Beato, M., Baxter, J.D., and Mellon, P.L. (1988). Negative regulation by glucocorticoids through interference with a cAMP responsive enhancer. Science *241*, 350–353.

Amati, B., and Land, H. (1994). Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death. Curr. Opin. Genet. Dev. *4*, 102–108.

Aravind, L., and Landsman, D. (1998). AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res. *26*, 4413–4421.

Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D., and Wagner, G.P. (2016). The origin and evolution of cell types. Nat. Rev. Genet. *17*, 744–757.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Barrera, L.A., Vedenko, A., Kurland, J.V., Rogers, J.M., Gisselbrecht, S.S., Rossin, E.J., Woodard, J., Mariani, L., Kock, K.H., Inukai, S., et al. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science *351*, 1450–1454.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Blake, J.A., Eppig, J.T., Kadin, J.A., Richardson, J.E., Smith, C.L., and Bult, C.J.; the Mouse Genome Database Group (2017). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucleic Acids Res. *45* (D1), D723–D729.

Brayer, K.J., and Segal, D.J. (2008). Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. Cell Biochem. Biophys. *50*, 111–131.

Bulyk, M.L., Johnson, P.L., and Church, G.M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. *30*, 1255–1261.

Bürglin, T.R. (2011). Homeodomain subtypes and functional diversity. Subcell. Biochem. *52*, 95–122.

Campagne, S., Saurel, O., Gervais, V., and Milon, A. (2010). Structural determinants of specific DNA-recognition by the THAP zinc finger. Nucleic Acids Res. *38*, 3466–3476.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25–36.

Chen, J., Zhang, Z., Li, L., Chen, B.C., Revyakin, A., Hajj, B., Legant, W., Dahan, M., Lionnet, T., Betzig, E., et al. (2014). Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell *156*, 1274–1285.

Choksi, S.P., Lauter, G., Swoboda, P., and Roy, S. (2014). Switching on cilia: transcriptional networks regulating ciliogenesis. Development *141*, 1427–1441.

Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. *373*, 895–907.

Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. PLoS Genet. *10*, e1004226.

Damante, G., Fabbro, D., Pellizzari, L., Civitareale, D., Guazzi, S., Polycarpou-Schwartz, M., Cauci, S., Quadrifoglio, F., Formisano, S., and Di Lauro, R. (1994). Sequence-specific DNA recognition by the thyroid transcription factor-1 homeodomain. Nucleic Acids Res. *22*, 3075–3083.

Dehal, P., and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. *3*, e314.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. Cell *166*, 538–554.

Dunwell, T.L., and Holland, P.W. (2016). Diversity of human and mouse homeobox gene expression in development and adult tissues. BMC Dev. Biol. *16*, 40.

Ecco, G., Imbeault, M., and Trono, D. (2017). KRAB zinc finger proteins. Development *144*, 2719–2729.

Emerson, R.O., and Thomas, J.H. (2009). Adaptive evolution in zinc finger transcription factors. PLoS Genet. *5*, e1000325.

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S., and Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat. Biotechnol. *34*, 1180–1190.

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztányi, Z., El-Gebali, S., Fraser, M., et al. (2017). InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. *45* (D1), D190–D199.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44* (D1), D279–D285.

Fong, A.P., and Tapscott, S.J. (2013). Skeletal muscle programming and reprogramming. Curr. Opin. Genet. Dev. *23*, 568–573.

Font, J., and Mackay, J.P. (2010). Beyond DNA: zinc finger domains as RNA-binding modules. Methods Mol. Biol. *649*, 479–491.

Frietze, S., and Farnham, P.J. (2011). Transcription factor effector domains. Subcell. Biochem. *52*, 261–277.

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

Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., Wasserman, W.W., Roach, J.C., and Sladek, R. (2009). TFCat: the curated catalog of mouse and human transcription factors. Genome Biol. *10*, R29.

Garza, A.S., Ahmad, N., and Kumar, R. (2009). Role of intrinsically disordered protein regions/domains in transcriptional regulation. Life Sci. *84*, 189–193.

Gebhardt, J.C., Suter, D.M., Roy, R., Zhao, Z.W., Chapman, A.R., Basu, S., Maniatis, T., and Xie, X.S. (2013). Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. Nat. Methods *10*, 421–426.

Geertz, M., Shore, D., and Maerkl, S.J. (2012). Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. Proc. Natl. Acad. Sci. USA *109*, 16540–16545.

Gertz, J., Reddy, T.E., Varley, K.E., Garabedian, M.J., and Myers, R.M. (2012). Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. Genome Res. *22*, 2153–2162.

Gotea, V., Visel, A., Westlund, J.M., Nobrega, M.A., Pennacchio, L.A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. *20*, 565–577.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J. (2009). A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. Cell *138*, 314–327.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell *163*, 1515–1526.

Hayward, A., Cornwallis, C.K., and Jern, P. (2015). Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc. Natl. Acad. Sci. USA *112*, 464–469.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources*2016* (Database (Oxford)), p. baw053.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Hu, S., Xie, Z., Onishi, A., Yu, X., Jiang, L., Lin, J., Rho, H.S., Woodard, C., Wang, H., Jeong, J.S., et al. (2009). Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. Cell *139*, 610–622.

Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am. J. Hum. Genet. *89*, 496–506.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. *43*, D117–D122.

Huntley, S., Baggott, D.M., Hamilton, A.T., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E., and Stubbs, L. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. *16*, 669–677.

Imbeault, M., Helleboid, P.Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *543*, 550–554.

Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H., and Zaret, K.S. (2016). The pioneer transcription factor FoxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation. Mol. Cell *62*, 79–91.

Jacobs, F.M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R., and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature *516*, 242–245.

John, L.B., and Ward, A.C. (2011). The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. Mol. Immunol. *48*, 1272–1278.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Johnson, P.F., and McKnight, S.L. (1989). Eukaryotic transcriptional regulatory proteins. Annu. Rev. Biochem. *58*, 799–839.

Jolma, A., and Taipale, J. (2011). Methods for Analysis of Transcription Factor DNA-Binding Specificity In Vitro. Subcell. Biochem. *52*, 155–173.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999.

Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. USA *111*, 6131–6138.

Kim, S., Broströmer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y.Q., et al. (2013). Probing allostery through DNA. Science *339*, 816–819.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. *42*, D966–D974.

Lambert, S.A., Albu, M., Hughes, T.R., and Najafabadi, H.S. (2016). Motif comparison based on similarity of binding affinity profiles. Bioinformatics *32*, 3504–3506.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. *43*, D257–D260.

Li, H., Liefke, R., Jiang, J., Kurland, J.V., Tian, W., Deng, P., Zhang, W., He, Q., Patel, D.J., Bulyk, M.L., et al. (2017). Polycomb-like proteins link the PRC2 complex to CpG islands. Nature *549*, 287–291.

Lovmar, L., Ahlford, A., Jonsson, M., and Syvänen, A.C. (2005). Silhouette scores for assessment of SNP genotype clusters. BMC Genomics *6*, 35.

Lynch, V.J., and Wagner, G.P. (2008). Resurrecting the role of transcription factor change in developmental evolution. Evolution *62*, 2131–2154.

Malik, S., and Roeder, R.G. (2010). The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nat. Rev. Genet. *11*, 761–772.

Marcon, E., Ni, Z., Pu, S., Turinsky, A.L., Trimble, S.S., Olsen, J.B., Silverman-Gavrila, R., Silverman-Gavrila, L., Phanse, S., Guo, H., et al. (2014). Human-chromatin-related protein interactions identify a demethylase complex required for chromosome segregation. Cell Rep. *8*, 297–310.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *44* (D1), D110–D115.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108–D110.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Meijsing, S.H., Pufall, M.A., So, A.Y., Bates, D.L., Chen, L., and Yamamoto, K.R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. Science *324*, 407–410.

Mesika, A., Ben-Dor, S., Laviad, E.L., and Futerman, A.H. (2007). A new functional motif in Hox domain-containing ceramide synthases: identification of a novel region flanking the Hox and TLC domains essential for activity. J. Biol. Chem. *282*, 27366–27373.

Morgunova, E., and Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. Curr. Opin. Struct. Biol. *47*, 1–8.

Muller, P.A., and Vousden, K.H. (2013). p53 mutations in cancer. Nat. Cell Biol. *15*, 2–8.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. *33*, 555–562.

Nardini, M., Gnesutta, N., Donati, G., Gatta, R., Forni, C., Fossati, A., Vonrhein, C., Moras, D., Romier, C., Bolognesi, M., and Mantovani, R. (2013). Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. Cell *152*, 132–143.

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E., and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife *4*, e04837.

Ou, J., Wolfe, S.A., Brodsky, M.H., and Zhu, L.J. (2018). motifStack for the analysis of transcription factor binding site evolution. Nat. Methods *15*, 8–9.

Panne, D. (2008). The enhanceosome. Curr. Opin. Struct. Biol. *18*, 236–242.

Plevin, M.J., Mills, M.M., and Ikura, M. (2005). The LxxLL motif: a multifunctional binding sequence in transcriptional regulation. Trends Biochem. Sci. *30*, 66–69.

Polach, K.J., and Widom, J. (1996). A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. J. Mol. Biol. *258*, 800–812.

Ptashne, M. (2011). Principles of a switch. Nat. Chem. Biol. *7*, 484–487.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Reece-Hoyes, J.S., and Marian Walhout, A.J. (2012). Yeast one-hybrid assays: a historical and technical perspective. Methods *57*, 441–447.

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

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell *147*, 1408–1419.

Rippe, K., Schrader, A., Riede, P., Strohner, R., Lehmann, E., and Längst, G. (2007). DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. Proc. Natl. Acad. Sci. USA *104*, 15635–15640.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Rosenfeld, M.G., Lunyak, V.V., and Glass, C.K. (2006). Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. Genes Dev. *20*, 1405–1428.

Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P.V., Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature *463*, 237–240.

Schmitges, F.W., Radovani, E., Najafabadi, H.S., Barazandeh, M., Campitelli, L.F., Yin, Y., Jolma, A., Zhong, G., Guo, H., Kanagalingam, T., et al. (2016). Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Res. *26*, 1742–1752.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Singh, H., Khan, A.A., and Dinner, A.R. (2014). Gene regulatory networks in the immune system. Trends Immunol. *35*, 211–218.

Sizemore, G.M., Pitarresi, J.R., Balakrishnan, S., and Ostrowski, M.C. (2017). The ETS family of oncogenic transcription factors in solid tumours. Nat. Rev. Cancer *17*, 337–351.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., and Mann, R.S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. *39*, 381–399.

Smale, S.T. (2014). Transcriptional regulation in the immune system: a status report. Trends Immunol. *35*, 190–194.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555–568.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. *11*, 751–760.

Stubbs, L., Sun, Y., and Caetano-Anolles, D. (2011). Function and evolution of C2H2 zinc finger arrays. Subcell. Biochem. *52*, 75–94.

Sur, I., and Taipale, J. (2016). The role of enhancers in cancer. Nat. Rev. Cancer *16*, 483–493.

Swinstead, E.E., Miranda, T.B., Paakinaho, V., Baek, S., Goldstein, I., Hawkins, M., Karpova, T.S., Ball, D., Mazza, D., Lavis, L.D., et al. (2016a). Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. Cell *165*, 593–605.

Swinstead, E.E., Paakinaho, V., Presman, D.M., and Hager, G.L. (2016b). 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 *38*, 1150–1157.

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

Takahashi, K., and Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. Nat. Rev. Mol. Cell Biol. *17*, 183–193.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Virbasius, C.A., Virbasius, J.V., and Scarpulla, R.C. (1993). NRF-1, an activator involved in nuclear-mitochondrial interactions, utilizes a new DNA-binding domain conserved in a family of developmental regulators. Genes Dev. *7* (12A), 2431–2445.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Wang, J., Zhuang, J., Iyer, S., Lin, X.Y., Greven, M.C., Kim, B.H., Moore, J., Pierce, B.G., Dong, X., Virgil, D., et al. (2013). Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. *41*, D171–D176.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. *5*, 276–287.

Wei, G.H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. *29*, 2147–2160.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al.; DREAM5 Consortium (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.

Weirauch, M.T., and Hughes, T.R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. *26*, 66–74.

Weirauch, M.T., and Hughes, T.R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell. Biochem. *52*, 25–73.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Wingender, E., Schoeps, T., Haubrock, M., and Dönitz, J. (2015). TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic Acids Res. *43*, D97–D102.

Wong, K.H., and Struhl, K. (2011). The Cyc8-Tup1 complex inhibits transcription primarily by masking the activation domain of the recruiting protein. Genes Dev. *25*, 2525–2539.

Worsley Hunt, R., and Wasserman, W.W. (2014). Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. Genome Biol. *15*, 412.

Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. *25*, 434–440.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

Zabidi, M.A., and Stark, A. (2016). Regulatory enhancer-core-promoter communication via transcription factors and cofactors. Trends Genet. *32*, 801–814.