CellPress

Molecular Cell

Review

Compendium of human transcription factor effector domains

Luis F. Soto,[1] Zhaorong Li,[2] Clarissa S. Santoso,[3,4] Anna Berenson,[3,4] Isabella Ho,[3] Vivian X. Shen,[3] Samson Yuan,[3] and Juan I. Fuxman Bass[2,3,4,*]

[1]Escuela Profesional de Genética y Biotecnología, Facultad de Ciencias Biológicas, Universidad Nacional Mayor de San Marcos, Lima 15081, Perú

[2]Bioinformatics Program, Boston University, Boston, MA 02215, USA

[3]Biology Department, Boston University, Boston, MA 02215, USA

[4]Molecular Biology, Cellular Biology, and Biochemistry Program, Boston University, Boston, MA 02215, USA

*Correspondence: fuxman@bu.edu

https://doi.org/10.1016/j.molcel.2021.11.007

SUMMARY

Transcription factors (TFs) regulate gene expression by binding to DNA sequences and modulating transcriptional activity through their effector domains. Despite the central role of effector domains in TF function, there is a current lack of a comprehensive resource and characterization of effector domains. Here, we provide a catalog of 924 effector domains across 594 human TFs. Using this catalog, we characterized the amino acid composition of effector domains, their conservation across species and across the human population, and their roles in human diseases. Furthermore, we provide a classification system for effector domains that constitutes a valuable resource and a blueprint for future experimental studies of TF effector domain function.

INTRODUCTION

Transcription factors (TFs) play a central role in the regulation of gene expression and thereby affect diverse biological processes such as cell differentiation and de-differentiation (Takahashi et al., 2007; Tapscott et al., 1988), development (Davidson and Erwin, 2006), and immune regulation (Carrasco Pro et al., 2018; Santoso et al., 2020). Most TFs contain two main types of protein domains to accomplish their functions: DNA-binding domains (DBDs) and effector domains (EDs) (Frankel and Kim, 1991; Lambert et al., 2018; Vaquerizas et al., 2009). DBDs direct TFs to their target genomic regulatory regions by recognizing specific DNA sequences. DBDs are well-conserved structural classes and are often used to classify TFs into families. For example, the current list of 1,639 human TFs is classified into 25 DBD families, the largest of which are zinc fingers Cys2His2 (ZF-C2H2) and homeodomains (Lambert et al., 2018). Alternatively, EDs can activate or repress target gene expression through several mechanisms such as interactions with cofactors, enzymes, and mediator, leading to histone modifications, changes in DNA methylation states, and recruitment of RNA polymerase II (RNA Pol II) (Frietze and Farnham, 2011; Reiter et al., 2017) (Figure 1A). Broadly, we can classify these EDs as activator domains (ADs), also known as trans-activator domains, repressor domains (RDs), and bifunctional (Bif) domains (i.e., those that can activate or repress gene expression depending on the cellular and chromatin contexts).

While there are multiple resources and annotations of TF DBDs, there are currently no comprehensive annotations of TF EDs. This is because EDs are generally less conserved across paralogs and orthologs than DBDs and often do not have well-defined structures, rendering predictions based on sequence or structure largely ineffective (Staller et al., 2018). Therefore, EDs have mostly been identified by deletion experiments, and their annotation is scattered across the literature.

The transcriptional regulation field has made substantial contributions to our understanding of the molecular mechanisms of gene expression and the role of EDs in the recruitment of the pre-initiation complex, chromatin organization, cofactor recruitment, RNA Pol II regulation, and DNA methylation (Roeder, 2019). Given the extensive and important research by thousands of scientists in this field, the goal of this article is not to offer a historical perspective on these key contributions, but rather to synthesize the currently available information and provide a novel resource to obtain a big-picture comparative perspective on TF EDs.

Here, we review >3 decades of literature to manually annotate 924 EDs across 594 human TFs. We use this resource to characterize EDs and their amino acid (aa) composition, sequence conservation, and roles in human diseases. In addition, we implement a web server annotating the known EDs, as well as to predict EDs across paralogs and within TF isoforms. Collectively, our data and analyses provide a novel and important resource for future studies of TF EDs.

Methods to identify and characterize EDs

The ability of EDs to modulate transcriptional activity has been mapped and characterized using different experimental approaches (Figure 1B; Table S1). Most of these approaches require recruiting either a full-length TF or a TF fragment to a transcriptional control region, followed by quantifying the transcriptional activity of a downstream target gene. Recruitment of the TF can be achieved using the intrinsic DBD of the TF and a promoter region known to bind the TF (Han et al., 2020; Ma and

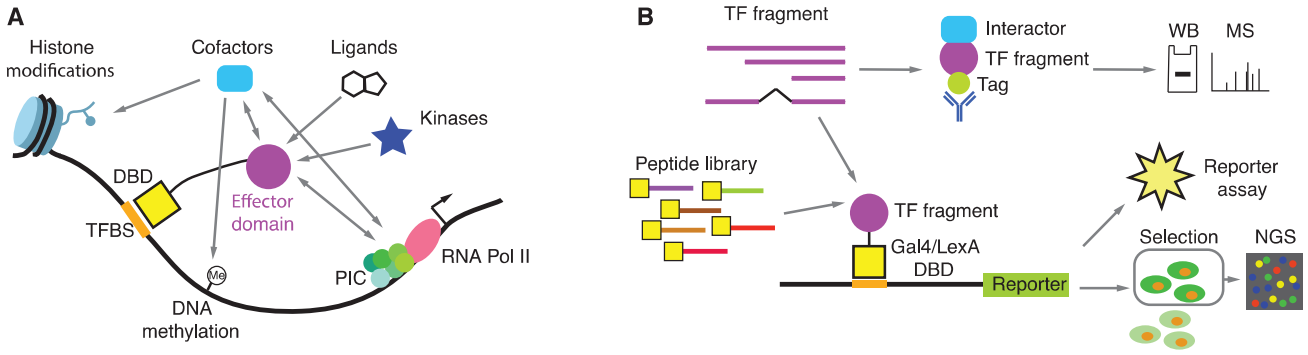


Figure 1. Effector domain (ED) identification, characterization, and function

(A) EDs can affect gene expression by interacting with cofactors and the preinitiation complex (PIC), by modifying histone tails, and by leading to changes in DNA methylation states. The activity of some EDs can be affected by interactions with ligands or by post-translational modifications.

(B) Experimental approaches to identify and characterize EDs. TF fragments or pool peptide libraries comprising tilling, random, or mutated peptides are fused to an exogenous DBD (e.g., Gal4, Gcn4, LexA, rTetR DBDs). Transcriptional activity is often measured using a reporter gene. In the case of high-throughput peptide screens, cells with different levels of reporter activity are sorted, and the enrichment for sequences corresponding to each peptide is determined by next-generation sequencing (NGS). Pull-down experiments are also used to identify interacting cofactors by western blot (WB) or mass spectrometry (MS).

Ptashne, 1987; Roose et al., 1998). The target gene can either be an endogenous target gene whose expression can be measured by qRT-PCR, or a reporter gene measured by enzymatic activity (e.g., luciferase, chloramphenicol acetyltransferase, β-galactosidase) (Ma and Ptashne, 1987; Meijer et al., 1992; Roose et al., 1998). These experiments involve protein deletions to identify the aa sequences that are necessary for activating or repressing transcription (i.e., if the regions are removed, the transcriptional effect is totally or partially lost). However, these assays rarely demonstrate that these sequences, on their own, are sufficient to elicit their transcriptional effect. To show sufficiency, complementary reporter assays are used in which TF fragments are fused to DBDs from heterologous TFs that have well-characterized DNA-binding sites, such as the yeast Gal4 and the bacterial LexA (Braun et al., 1990; Brent and Ptashne, 1985; Hope and Struhl, 1986). This allows for recruitment of TF fragments of any size to test their effect on reporter gene expression. Moreover, these experiments are not compromised by effects that deletions in the native TF may have on its ability to bind its natural DNA-binding sites. To avoid mapping regions that affect the overall function of the TF (i.e., necessary but not sufficient) or regions that are active in a heterologous context but not within the TF sequence (i.e., sufficient but not necessary), both types of experiments showing necessity and sufficiency are recommended.

Most of the assays listed above are low throughput, in particular protein deletion experiments, as they require custom-designed sequences for each TF tested. In addition, studies of different TFs may require different cell types expressing the appropriate cofactors and varying experimental conditions such as different ligands and stimuli (Figure 1A). Recently, exogenous DBD or dCas9 fusion experiments have been adapted for high-throughput transcriptional activity screens using libraries coding for thousands of peptide sequences (Figure 1B). These peptide libraries can include fragments of protein-coding genes (including TFs), comprehensive mutagenesis of selected peptide sequences to identify key amino acids within the peptides involved in transcriptional activity, or random peptides to screen for activating and repressive functions (Alerasool et al., 2021; Arnold et al., 2018; Erijman et al., 2020; Ravarani et al., 2018; Staller et al., 2018; Tycko et al., 2020). In these experiments, the reporter used allows for the separation of cells harboring a transcriptionally active (or repressive) DBD-peptide fusion within a pool (e.g., GFP reporter using fluorescence-activated cell sorting, a surface marker using magnetic separation), followed by sequencing of the enriched peptide sequences.

In addition, protein-protein interaction (PPI) studies have provided indirect evidence of transcriptional regulatory activity by identifying TF fragments that interact with cofactors or other proteins that modulate transcription (Figure 1B). For example, pull-down assays have been used extensively to identify the interactions of EDs with cofactors and chromatin remodeling complexes (Giraud et al., 2002; Neely et al., 1999; Xu et al., 2018). When integrated with reporter studies, these PPIs can provide a mechanism for observed transcriptional effects.

Few computational approaches have been developed to predict TF EDs. This is mainly because there are no comprehensive databases annotating experimentally determined EDs, because EDs are poorly conserved between paralogs, and because the sequence rules for transcriptional activity have not been fully established. EDs are thus relatively difficult to predict from aa sequences compared to DBDs (Mistry et al., 2021). 9aaTAD is a predictor based on different experimentally determined 9-mer ADs; however, this tool is limited to short ADs (Piskacek et al., 2007). Since sequence alignment proved to be of limited use to predict ADs, novel machine learning predictors have been developed. For example, ADpred is a deep learning model that uses the aa composition and the secondary structure of known ADs to predict ADs between 9 and 30 amino acids (Erijman et al., 2020). PADDLE, a deep convolutional neural network model, uses 53 aa tiles to predict the location of ADs within a TF sequence, its key residues, and its transactivation strength (Sanborn et al., 2021). However, most experimentally determined ADs are longer, as we found from our curation (median = 91 amino acids). Although this could be associated with imprecise boundary definition for some ADs, many carefully mapped ADs are indeed longer, somewhat limiting the applicability of

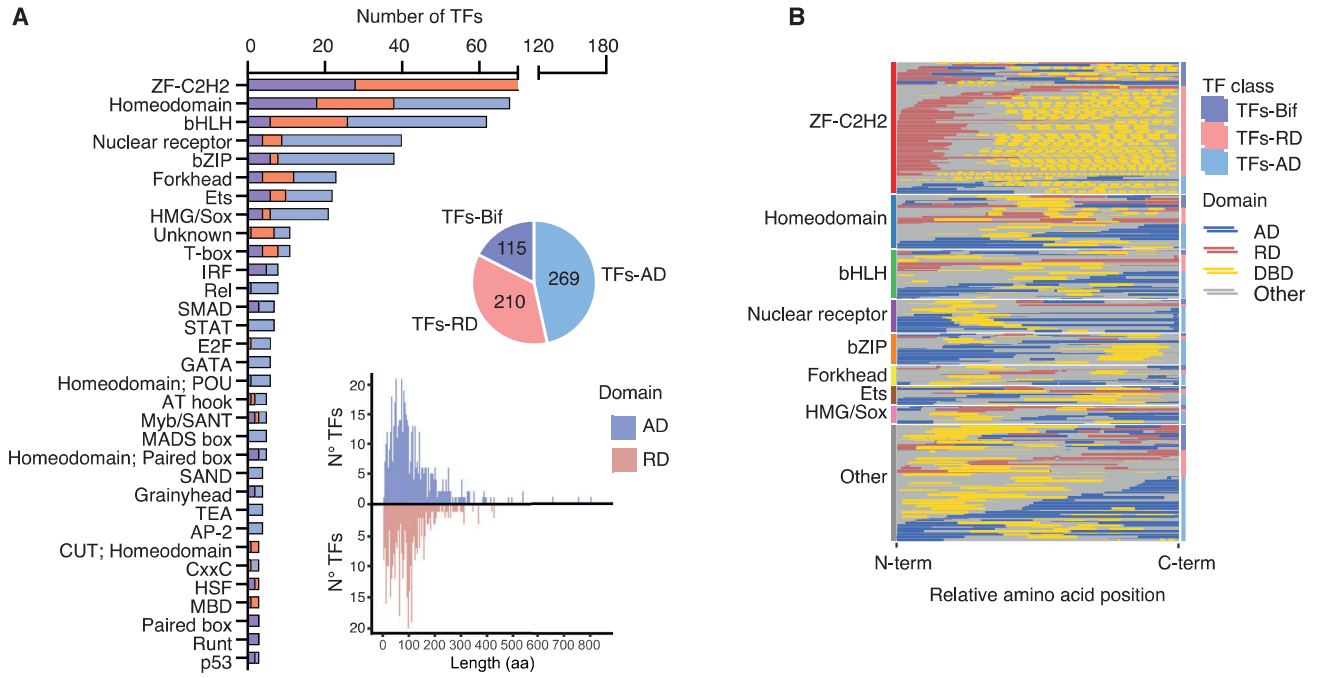


Figure 2. Distribution of EDs

(A) Number of TFs per family with annotated EDs classified as TFs-AD (if they only have ADs annotated), TFs-RD (if they only have RDs annotated), and TFs-Bif (if they have both ADs and RDs or bifunctional domains annotated). Only TF families with ≥3 annotated TFs are included. The pie chart indicates the number of TFs in each class. The histogram indicates the amino acid length distribution for ADs (blue) and RDs (red).

(B) Relative position of ADs, RDs, and DBDs within the TF amino acid sequence (from N to C termini). TF families are indicated by the left bars. Within each family, TFs are sorted based on whether they are classified as TFs-Bif, TFs-RD, or TFs-AD (indicated by the right bars). Within each class, TFs are sorted by the relative position of the ED in the TF sequence.

See also Figure S1.

current computational predictions. Furthermore, to our knowledge, with the exception of KRAB and POZ/BTB domains, there are currently no predictors developed for repression domains. Therefore, there is a need for improved computational approaches to predict EDs, which will be in part driven by new large-scale experimental datasets.

A census of human TF EDs

To generate a large-scale resource of experimentally validated EDs, we searched for ED evidence across the literature for the 1,639 annotated human TFs. We manually curated and extracted experimental evidence for 924 EDs from 594 TFs (Figure 2A; Table S2). Of these, only 94 EDs belonging to 79 TFs were reported in the Pfam domain database (mostly corresponding to KRAB and BTB/POZ domains), illustrating the lack of structural classification for most EDs (Mistry et al., 2021). We implemented a webtool called TFRegDB (https://tfregdb.bu.edu/tfregdb/), annotating available information about human TF EDs, including aa sequence, coordinates in different isoforms, experimental methods used to determine the EDs, whether they are necessary or sufficient for transcriptional activity, a confidence score, and links to supporting evidence. We also implemented a BLAST search functionality, in which a query sequence can be submitted to detect EDs in TF isoforms or to predict EDs based on aa sequence similarity.

We annotated EDs in all major families of TFs, including ZF-C2H2 (170 TFs), homeodomains (68 TFs), and bHLHs (62 TFs) (Figure 2A). Of the 594 TFs in our database, 40% have ≥2 EDs annotated (Table S2). Based on the ED activity, TFs can be classified into 3 groups: those that contain only ADs (269 "TFs-AD"), those that have only RDs (210 "TFs-RD"), and those with both ED types (115 "TFs-Bif"). As expected, most ZF-C2H2 are TFs-RD, as many of these TFs contain the well-characterized KRAB and BTB/POZ domains involved in transcriptional repression (Collins et al., 2001). Conversely, most TFs in the basic helix-loop-helix (bHLH), nuclear receptor, and homeodomain families are classified as TFs-AD (Figures 2A and S1A). However, many of these TFs, such as nuclear receptors, may switch from repression to activation upon interaction with ligands, while the activities of others are affected by post-translational modifications. The classification into TFs-AD, TFs-RD, and TFs-Bif is solely based on reported ED activity in the conditions tested. Therefore, many of these TFs could be bifunctional in other conditions, or if other aa regions of the TF are considered.

Reported ED sizes range from 4 to 1,248 amino acids, with a median of 91 amino acids (Figure 2A). Although some of these differences are likely due to varying mechanisms of action, in many cases size differences likely arise from variation in the stringency of the deletion experiments used to identify the EDs. Overall, we found that 30% of the EDs were located at the N terminus, 28% at the C terminus, and 42% in internal

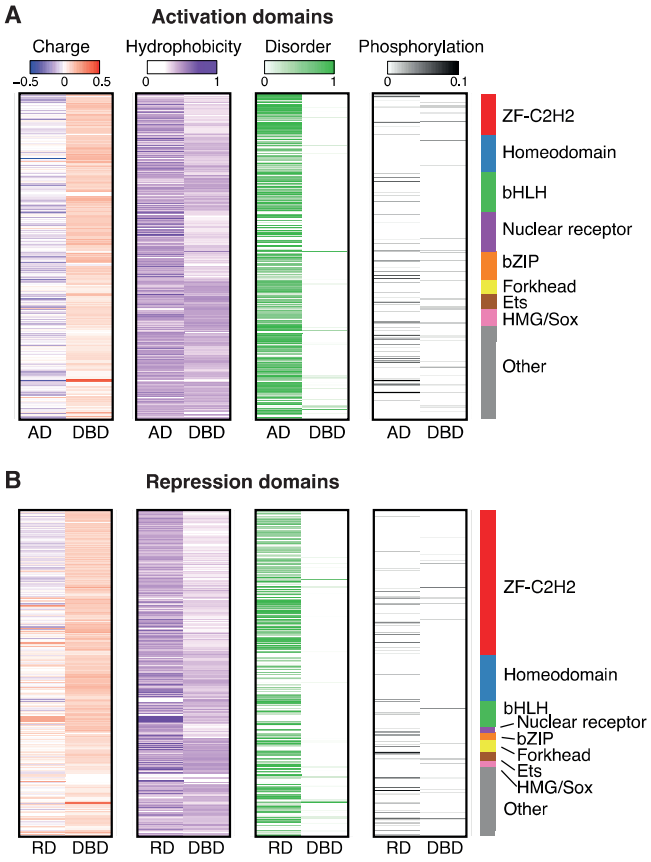


Figure 3. Sequence features of EDs

For each activation domain (A) and repression domain (B), the charge density (charge/amino acid length), hydrophobicity, disorder (determined using AlphaFold), and phosphorylation density (number of phosphorylation events/amino acid length) are indicated.

See also Figures S2 and S3.

regions within the TF aa sequence (Figure 2B). However, the positioning of EDs differs among TF families (Figures 2B and S1B). To illustrate, repressor domains in ZF-C2H2 predominantly reside at the N termini, while activation domains in bZIP and HMG/Sox families mainly reside at the N and C termini, respectively. This suggests that alternative splicing, transcription starts, and polyadenylation sites may differentially affect ADs and RDs from different TF families.

aa composition of EDs

It has been broadly determined that DBDs are enriched in basic amino acids that increase TF affinity for the negatively charged DNA (Lambert et al., 2019), but less is known about the aa composition of EDs from different TF families. Since Paul Sigler proposed in 1988 the acid blob and negative model, positing that acidic ADs interact with RNA polymerase electrostatically (Sigler, 1988), significant progress has been made in characterizing the aa composition of the ADs of some TFs, as well as determining the rules for transcriptional activity (Erijman et al., 2020; Sanborn et al., 2021). Seminal studies on yeast TFs reported that ADs are acidic, disordered, and hydrophobic (Drysdale et al., 1995; Hope and Struhl, 1986; Ravarani et al., 2018; Staller et al., 2018). However, predictions suggest that ADs of human TFs are not as highly enriched in acidic amino acids as yeast ADs (Erijman et al., 2020). For example, HOXA13 and ONECUT1 have basic ADs enriched in lysine/arginine and histidine, respectively (Table S2), consistent with the identification of basic ADs in high-throughput screens (Arnold et al., 2018). Furthermore, although acidity may be important for some human ADs, acidity is not sufficient to predict AD function, as appropriate levels of hydrophobicity and disorder are also required (Staller et al., 2018; Tycko et al., 2020). A current model, known as the exposure model, indicates that acidic residues that surround hydrophobic motifs are necessary to repel one another, promoting interaction between exposed hydrophobic residues with positively charged cofactors (Ferreira et al., 2005; Hermann et al., 2001; Staller et al., 2018; Warfield et al., 2014). These contacts between hydrophobic residues may mediate high-affinity PPIs via the hydrophobic effect (Levy and Onuchic, 2006). This model was initially proposed based on ADs from yeast TFs, and recently supported by mutational studies in five human TFs (Staller et al., 2018; Staller et al., 2021). Whether these models extend to other human ADs and RDs remains to be determined.

To establish whether the reported sequence characteristics are present in most of the annotated EDs, we evaluated the acidity, hydrophobicity, and disorder of ADs and RDs of TFs from different families. We confirmed that ADs and RDs are more acidic than DBDs and found that globally, ADs are statistically more acidic than RDs ($p < 2.2 \times 10^{-16}$, Wilcoxon test) (Figures 3A and 3B). Interestingly, not all TF families show the same level of acidity in their EDs, which could be partially driven by sequence homology between paralogs. For example, homeodomain ADs are less acidic than ADs from other families, while RDs in ZF-C2H2 are the most acidic among the RDs. Furthermore, there are marked differences even within families (e.g., the HES1-7 bHLH subfamily has more basic RDs than other bHLHs) (Figure 3B). We also found that both ADs and RDs are more hydrophobic than DBDs (Figures 3A and 3B). Although no specific hydrophobic enrichment was observed for any TF family, in bHLH, the RDs of the HES1-7 subfamily are more hydrophobic than RDs from other families. Most TF EDs are highly acidic and hydrophobic; however, there are some TFs whose EDs are basic and highly hydrophobic (HES TF subfamily) or acidic but lowly hydrophobic (e.g., HOXB7, HMGA1).

Several studies of individual TFs have shown that EDs are enriched in disordered regions (Liu et al., 2006; Oldfield and Dunker, 2014). Disordered regions have been associated with the occurrence of PPIs, as their flexibility allows disordered regions to bind multiple structurally diverse protein partners (Oldfield and Dunker, 2014). This disorder allows EDs to assume different conformations when bound to cofactors, facilitating the dynamic exposure of hydrophobic motifs (Dyson and Wright, 2016; Staby et al., 2017; Warfield et al., 2014). For example, the disordered ADs of TP53, HIF1A, REL, STAT, and other TFs interact with well-structured domains of coactivators, such as CREBBP and EP300 (Dyson and Wright, 2016). Using AlphaFold (Jumper et al., 2021), we found that disorder is a property shared by both ADs and RDs (Figures 3A and 3B). ADs are significantly more disordered than RDs ($p = 1.9 \times 10^{-11}$, Wilcoxon test), and both are more disordered than DBD ($p < 2.2 \times 10^{-16}$, Wilcoxon

test) (Figures 3A and 3B). This disorder in EDs is a shared feature across all major TF families. Nevertheless, we observed a large variability within TF families, with some EDs being 100% disordered (e.g., the ADs of SP1 and SP3), while others are highly ordered (e.g., the RDs of MXI1 and MNT, which are alpha helixes).

Short linear motifs (SliMs), which are involved in PPIs and are generally enriched in hydrophobic amino acids, could be more important than overall high levels of hydrophobicity (Tompa et al., 2014). Many examples of SliMs have been reported in non-human EDs, but few cases have been studied in human EDs (Dinkel et al., 2014). In general, it has been shown that disordered structures facilitate the interaction mediated by these SliMs (Staller et al., 2018), but more in-depth studies are needed to determine their role across TF families.

In addition to general charge, hydrophobicity, and disorder features, many EDs, ADs in particular, have been shown to display aa compositional bias (Figure S2A). For example, many ADs across TF families are enriched in proline, serine, glutamine, glycine, and alanine, as has been previously described (Gerber et al., 1994; Husberg et al., 2001; Meijer et al., 1992; Paulsen et al., 1992; Pei and Shih, 1991; Raney et al., 1991). Although there is also a compositional bias for some RDs, in particular those enriched in proline and serine, these are less frequent than for ADs (Figure S2B). These enriched amino acids are generally present in the EDs of TFs from many different families.

Post-translational modifications (PTMs) are known to regulate TF functions by affecting PPIs, cellular localization, and ultimately, their regulatory activity. Furthermore, the dysregulation of TF PTMs has been associated with several pathological conditions (Filtz et al., 2014; Qian et al., 2020; Tootle and Rebay, 2005). In particular, phosphorylation is known to play a significant role in the activation of many TFs and their interaction with cofactors and other protein complexes (Filtz et al., 2014). Phosphorylation introduces negative charges, thus changing charge and solubility properties of EDs. As negative charges spaced between hydrophobic residues help keep domains exposed to solvent, phosphorylation may act as a switch changing the ability of ADs and RDs to interact with other proteins and cofactors. This is the case of IRF5 and IRF3, whose phosphorylation stimulates dimerization and interaction with the coactivators CREBBP/EP300 (Chen et al., 2008), while the phosphorylation of ELK1 promotes mediator recruitment to promoter sequences (Cantin et al., 2003). Similarly, phosphorylation in the ADs of TP53 was reported to increase the binding to different domains of EP300 and reduce binding to the negative regulator Mdm2 (Teufel et al., 2009). Since most of these studies were performed on specific TFs or functional domains, we used PhosphoSitePlus (Hornbeck et al., 2019), a curated phosphorylation site database, to analyze the frequency of phosphorylation events in EDs and DBDs. Across most major TF families, we found that EDs are more highly phosphorylated than DBDs (Figures 3A and 3B), even when normalizing by the frequency of serines, threonines, and tyrosines (Figures S2C and S2D). Our analysis suggests that 21% of EDs may be regulated by phosphorylation. This is likely an underestimate, as some EDs may be phosphorylated in conditions not yet tested. Except for a few cases, the overall role of other PTMs in ED regulation remains to be determined.

Role of EDs in liquid-liquid phase separation

Recent evidence suggests that TF EDs contribute to gene regulation by facilitating liquid-liquid phase separation (LLPS), during which chromatin-bound TFs, co-regulators, and other transcription machinery form dynamic condensates within the nucleus (Boija et al., 2018; Hnisz et al., 2017; Sabari et al., 2018; Shrinivas et al., 2019). Forming these distinct transcription "factories" is thought to enhance transcriptional efficiency by increasing the effective concentration of required proteins within the crowded milieu of the nucleus. LLPS can be driven by two main types of interactions: (1) specific interactions between folded molecular domains or between folded and unfolded domains, or (2) non-specific interactions between intrinsically disordered low-complexity domains (LCDs) (Chiesa et al., 2020). As many TF EDs contain LCDs, it is hypothesized that the regulatory functions of these EDs depend on their ability to participate in LLPS by forming LCD-LCD interactions with co-regulators. Boija et al. (2018) have shown that the TFs OCT4, GCN4, and estrogen receptor form phase-separated condensates with the co-regulatory protein Mediator and that the processes of LLPS and transcriptional activation by these TFs require the same key AD residues.

It is important to note that while Boija et al. (2018) showed that LLPS induction by OCT4, GCN4, and estrogen receptor requires activation domain residues, others have shown that LCDs are not required for all LLPS events involving TFs (Chiesa et al., 2020). Li et al. (2020) showed that the DBD of mouse TF Sox2, and not the LCDs, are required for the incorporation of Sox2 and coactivator Brd4 into transcriptional clusters, suggesting spatial clustering of *cis*-regulatory elements. In addition, it has yet to be shown whether transcriptional activation or repression involving EDs requires the formation of liquid droplets. Chong et al. (2018) observed that while TF overexpression resulted in LLPS, expression at physiological concentrations resulted in the formation of LCD-LCD interaction-dependent transcriptionally active protein "hubs" without observable phase separation. This suggests that TF EDs can exert their transcriptional regulatory functions by forming transcription centers without the requirement for LLPS. However, given that EDs and LCDs are not synonymous, more work is required to fully understand how activation and repression domains of TFs exert their functions and the involvement of LLPS.

To evaluate whether ADs and RDs are associated with a propensity for phase separation, we compared the LLPS score between TFs classes using two different phase separation predictors (van Mierlo et al., 2021; Vernon et al., 2018). We found that TFs-AD and TFs-Bif have higher LLPS scores and probabilities than TFs-RD (Figures S3A and S3B), suggesting that ADs may play an important role in the LLPS. Moreover, we observed that 15.8% of ADs contain aa contexts predicted to promote LLPS (Vernon et al., 2018), versus 5.5% for RDs and 1.5% for DBDs (Figure S3C). ADs with LLPS-promoting aa contexts were found in TFs well known for promoting the formation of phase-separated condensates (e.g., SOX2, POU5F1, NANOG). Although domains without effector function can also be involved in LLPS, our results suggest that many ADs likely promote LLPS.

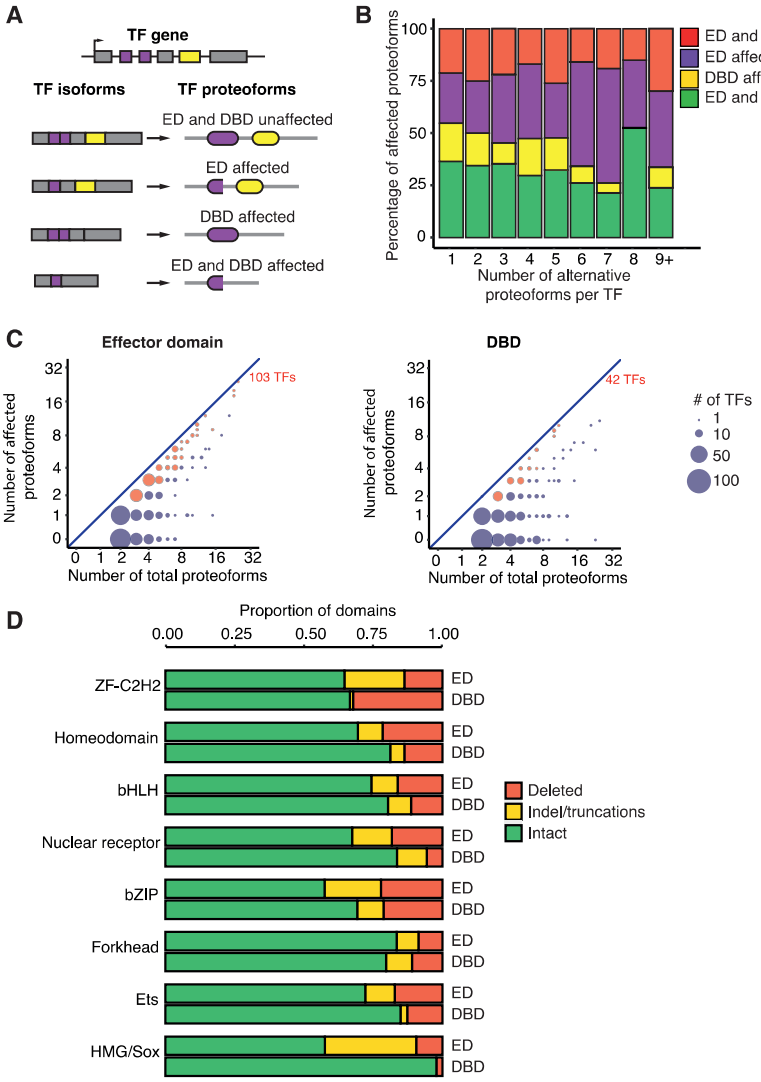


Figure 4. EDs affected in TF proteoforms

(A) Schematic of different proteoforms with ED affected, DBD affected, or with both domains affected or unaffected. Purple rectangles indicate ED coding exons; yellow rectangles indicate DBD coding exons; purple and yellow ovals indicate the ED and DBD, respectively.

(B) Fraction of proteoforms with ED, DBD, or both domains affected. TFs are binned based on the number of alternative proteoforms.

(C) Comparison between the number of proteoforms where the ED (left) or the DBD (right) are affected versus the total number of proteoforms of a TF. The size of the circles indicates the number of TFs. Red circles indicate TFs with >50% of proteoforms with affected domain.

(D) Proportion of EDs and DBDs where the domains are intact, have indels, or are deleted across proteofoms for each TF family.

See also Figure S4.

EDs are preferentially affected in TF proteoforms

TF proteoforms produced by alternative promoters, splicing, and polyadenylation can differ in both DNA binding and effector activity, potentially leading to variation in gene regulatory networks across tissues or pathological conditions (Figure 4A) (Epstein et al., 1994; Foulkes et al., 1991; Kozmik et al., 1993; López, 1995; Venkatanarayan et al., 2015). For example, *in silico* studies found that alternative splicing in murine TFs preferentially affects DBDs (Taneri et al., 2004). In regard to EDs, experimental studies on individual TFs found that different proteoforms of mouse Pou2f2 (Stoykova et al., 1992), human PAX8 (Kozmik et al., 1993), and human RUNX1 (Tanaka et al., 1997) have reduced transcriptional activity due to AD loss. Beyond individual examples, how EDs are affected in different TF proteoforms is not currently known on a TF-wide scale.

To gain more insight into how EDs and DBDs are affected in different TF proteoforms, we used a curated TF proteoform database derived from GENCODE version 30 (Frankish et al., 2019). We considered a TF proteoform to be affected if its functional domain (either DBD or ED) was fully deleted, had truncations, or had insertions/deletions (indels). We found that EDs were affected in a higher proportion of proteoforms than DBDs, regardless of the number of proteoforms per TF (Figure 4B) and regardless of domain length (Figure S4). Among the TFs with >2 proteoforms, there were only 42 (10.8%) TFs with affected DBDs in most of their proteoforms, while for EDs, this was the case for 103 TFs (26.6%) ($p = 1.9 \times 10^{-8}$ by proportion comparison test) (Figure 4C). This suggests that transcriptional activity is more frequently affected across proteoforms than DNA binding. EDs were more affected by indels, truncations, and full domain deletions than DBDs ($p = 7.7 \times 10^{-16}$, Kolmogorov-Smirnov test) across most major TF families, except for Forkhead and ZF-C2H2 (Figure 4D). In the case of ZF-C2H2, this could be related to the loss of individual ZFs in multiple proteoforms. Interestingly, the EDs of the Forkhead family were the least affected by indels or deletions, while EDs of bZIP and HMG/Sox TFs were the most affected (Figure 4D). We detected widespread variability in EDs that may contribute to differences in transcriptional activities between proteoforms as shown in multiple examples in the literature. Alternatively, these results may derive from a lower impact of deletions and truncations in EDs on overall transcriptional activity.

Evolutionary and population-wide divergence of EDs

While DBDs are highly conserved across multiple species and TF families, anecdotal examples have suggested that EDs are lowly conserved (Staller et al., 2018). To evaluate ED conservation in our resource, we aligned human EDs across TF orthologs in 27 vertebrate species and found a lower aa sequence conservation

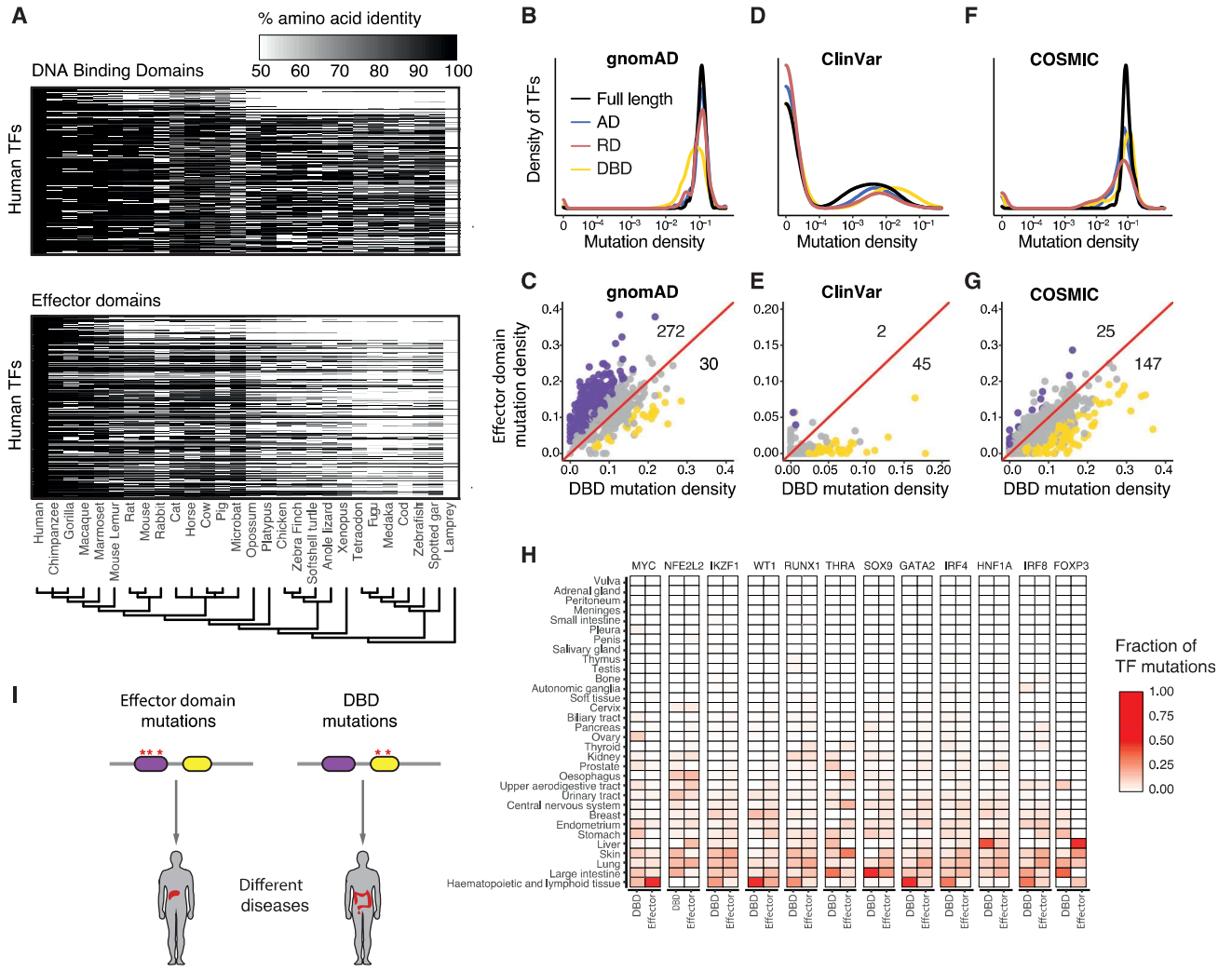


Figure 5. ED conservation and association with disease

(A) Conservation of DBDs and EDs between human and 27 vertebrate species. The percentage amino acid identity of the corresponding domains between the human sequence and the sequence in the indicated species is shown in shades of gray. The evolutionary relation among species is indicated as a phylogenetic tree.

(B, D, and F) Density distributions of the number of mutations in the indicated domains per coding sequence length for variants reported in gnomAD (B), and the mutations reported in ClinVar (D) and COSMIC (F).

(C, E, and G) Scatter plot between the density of mutations in DBDs and EDs in gnomAD (C), ClinVar (E), and COSMIC (G). Purple and yellow dots indicate TFs with a significant enrichment of mutations in EDs and DBDs, respectively. Significant TFs were identified by a Fisher's exact test followed by Benjamini-Hochberg (BH) correction and $q < 0.1$ as a cutoff. The numbers above and below the diagonal indicate the number of purple and yellow dots, respectively.

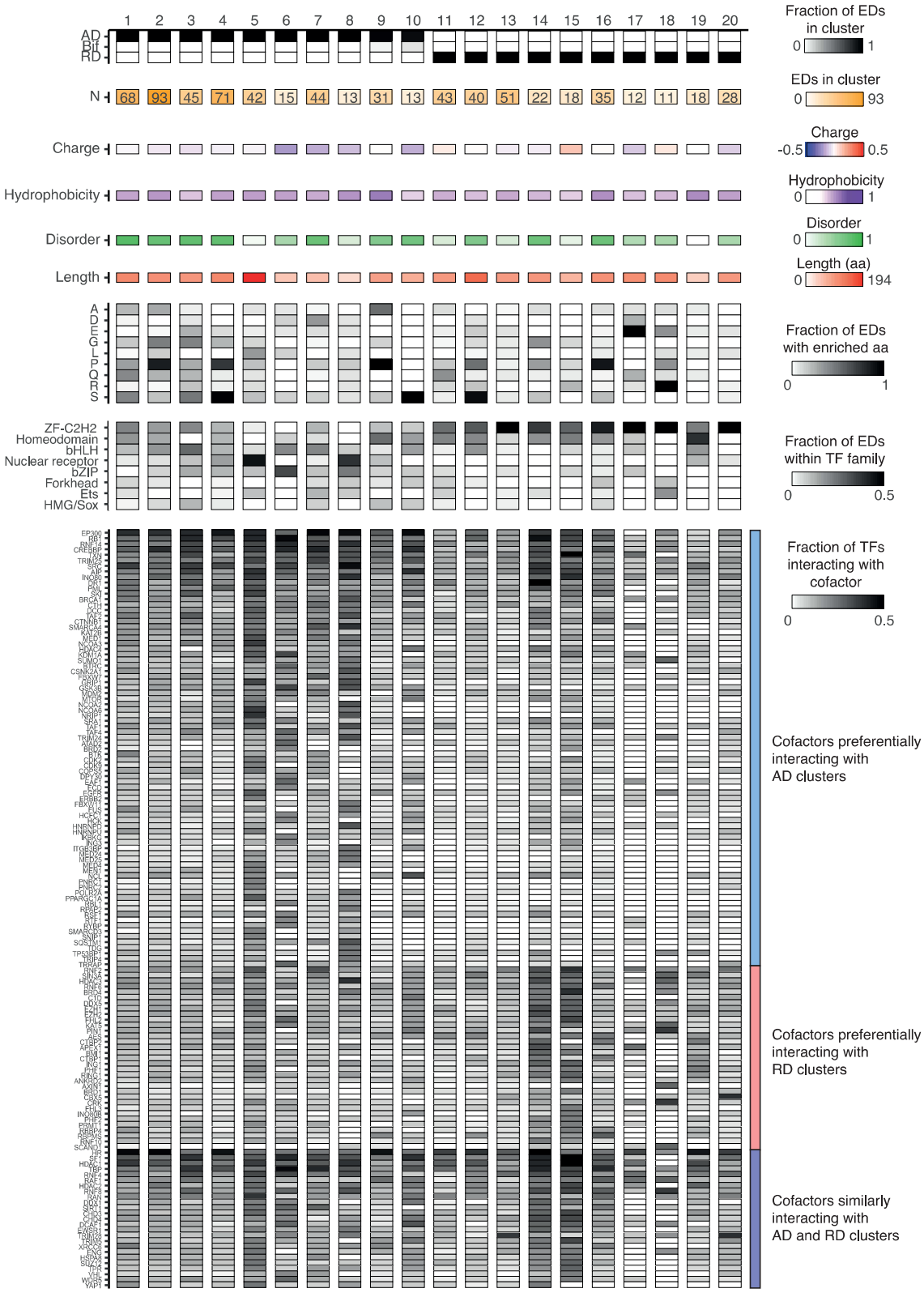
(H) Fraction of mutations in the DBD and ED for each indicated TF that has been detected in the indicated cancer types. Each column sums to 1.

(I) Schematic showing that mutations in EDs and DBDs could lead to different diseases.

See also Figure S5.

compared to DBDs across all families (Figure 5A). As expected, both DBDs and EDs are less conserved as the divergence time increases; however, ED conservation diminished more drastically (Figure S5). Although this low conservation can be partially explained by ED boundaries being less well determined than DBDs (e.g., we observed that short EDs are more conserved than long EDs), it is also likely that EDs are more plastic than DBDs. It has been reported that EDs can evolve rapidly, conferring greater evolutionary, structural, and functional plasticity on the interactome (Sanborn et al., 2021; Tompa et al., 2014; Wang et al., 2012).

Previous studies reported that the DBDs of most human TFs are depleted of common genetic variation (Barrera et al., 2016), likely because small changes in DBDs can lead to marked changes in affinity or specificity that could have detrimental effects. Mutagenesis studies in a few TFs have suggested that ADs more readily tolerate aa substitutions than DBDs (Ravarani et al., 2018; Sainz et al., 1997; Staller et al., 2018). To explore the presence of common variants in EDs, we used the gnomAD database (Karczewski et al., 2020) to compare the proportion of missense variants within EDs and DBDs for each TF. In general, we observed a higher proportion of missense variants in



(legend on next page)

EDs (both ADs and RDs) than in DBDs (12.3 variants/100 nt versus 8.4 variants/100 nt, $p < 2.2 \times 10^{-16}$ by Wilcoxon test) (Figure 5B). In particular, we found 272 TFs with a significantly higher proportion of variants in EDs than in DBDs, while 30 TFs had a significantly higher proportion of variants in DBDs than in EDs (Figure 5C). This higher proportion of variants in EDs was not dependent on minor allele frequencies and was not observed for synonymous variants (not shown). We observed similar results when analyzing genetic variants from the 1000 Genomes Project (Auton et al., 2015). These results suggest that there is a stronger negative selection for mutations in DBDs than in EDs. Among the TFs whose EDs are more tolerable to mutations, DUX4 and ZNF595 showed the greatest proportion of missense variants in their AD and RD (58.6% and 37.9% of missense variants, respectively). Although most EDs are found to tolerate missense variants, there are several highly conserved EDs. For example, RARB and RBPJ showed the lowest proportion of missense variants in their AD and RD, respectively (1.9 variants/100 nt and 3.4 variants/100 nt), although their DBDs are highly mutated.

Mutations in EDs and association with disease

Mutations in TF EDs have long been associated with many genetic diseases and cancers (Bradner et al., 2017). While mutations in DBDs can alter the targets of a TF by modifying its DNA-binding affinity and specificity (Barrera et al., 2016; Sahni et al., 2015), mutations in EDs can alter the ability of a TF to activate or repress gene expression by affecting its interactions with cofactors, mediator, or chromatin-modifying enzymes (Frietze and Farnham, 2011; Lambert et al., 2018). However, the extent to which mutations affect EDs has not been comprehensively determined.

To determine the prevalence of germline mutations associated with disease within EDs, we considered pathogenic and likely pathogenic mutations from the ClinVar database (Landrum et al., 2020). We found disease-associated variants both in EDs and DBDs, although DBDs were preferentially mutated in disease (Figure 5D). We found 44 TFs significantly enriched in DBD mutations ($q < 0.1$, Fisher's exact test) and only 2 TFs (SMAD3 and SMAD4) enriched in ED mutations (Figure 5E). This suggests either that fewer mutations in EDs are pathogenic or that multiple mutations may be concurrently needed to produce a phenotype, which is consistent with the high tolerance for the variants observed in EDs in the human population (Figures 5B and 5C).

Mutations in DBDs and EDs of different TFs (e.g., MYC, TP53, ESR1) have been identified or predicted as cancer drivers (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; Martínez-Jiménez et al., 2020). However, the prevalence of these ED mutations in relation to DBD mutations has not been comprehensively determined. By exploring the COSMIC database (Tate et al., 2019), we found that the proportion of cancer-associated somatic mutations in EDs is lower than in DBDs ($p < 2.2 \times 10^{-16}$, by Wilcoxon test) (Figure 5F). We found 25 and 147 TFs with statistically enriched somatic mutations in EDs and DBDs, respectively (Figure 5G). However, many EDs have a density of somatic mutations comparable to or higher than that of many DBDs (Figure 5F). For example, the EDs of MYC, SMAD4, SMAD3, AR, and SIM1 are enriched in cancer-associated mutations. Interestingly, we identified 12 TFs for which mutations in their respective EDs and DBDs are associated with different types of cancer ($q < 0.1$, by Fisher's exact test) (Figure 5H). For example, somatic mutations in the AD of MYC are preferentially associated with hematopoietic and lymphoid cancers, whereas mutations in the DBD of MYC are associated with many different cancer types such as hematopoietic, lymphoid, large intestine, and stomach. Similarly, while mutations in the DBD of FOXP3 are enriched in large intestine cancer, mutations in its RD are associated with liver cancers. This suggests that, at least for some TFs, mutations in different functional domains can lead to different diseases (Figure 5I).

These results show that EDs are more tolerable to common genetic variation and that they are less frequently associated with disease mutations than DBDs. Nevertheless, there are still numerous examples of disease-associated mutations in EDs.

Classification of EDs

EDs have traditionally been identified based on regulatory activity (activation versus repression), biophysical features (e.g., charge, hydrophobicity, disorder), the enrichment of certain amino acids (e.g., proline, serine, glutamine), and sequence conservation. To provide a functional classification of EDs, we leveraged these features to calculate pairwise similarities between EDs (see Data S1), which we then used to identify clusters of EDs with similar features. After an initial clustering into 63 clusters, we retained 20 containing at least 10 EDs, which comprise 77% of the EDs we annotated (Figure 6; Table S3).

We identified 10 clusters of ADs and 10 clusters of RDs (Figure 6). These clusters differ in the biophysical features and the enrichment of certain amino acids within their sequences. For example, clusters 15 and 18 comprise basic RDs, enriched in arginine residues, whereas cluster 4 comprises mildly acidic and disordered ADs enriched in serine and proline residues (Figure 6). Some clusters are enriched in TFs from certain families, such as clusters 5 (nuclear receptor), 6 (bZIP), 8 (nuclear receptor), 13 (ZF-C2H2), 17 (ZF-C2H2), 18 (ZF-C2H2), 19 (homeodomain), and 20 (ZF-C2H2). However, many clusters contain EDs from different TF families without a clear TF family enrichment, suggesting that the ED classification does not directly match TF classifications based on DBDs. This is consistent with the high variability in ED regulatory activity, localization within the TF aa sequence, biophysical features observed even

Figure 6. Classification of EDs

EDs were classified into 20 clusters based on biophysical features, amino acid enrichment, and sequence similarity. The number of EDs per cluster is indicated in shades of orange. The charge density, hydrophobicity, and disorder were determined as in Figure 2. The length in amino acids is indicated in shades of red. The fractions of EDs per cluster enriched in each amino acid, TF family, or interacting with a cofactor are indicated in shades of gray. Cofactors interacting with at least 20% of TFs in at least 1 cluster are shown.

See also Figure S6.

within TF families (Figures 2 and 3), and the modular organization of TF protein domains.

Next, we evaluated whether EDs from different clusters preferentially interacted with specific cofactors, and thus may share mechanisms of action. Although interactions between EDs and cofactors have not been comprehensively determined, we leveraged PPIs from BioGRID, HuRI, and Lit-BM (Luck et al., 2020; Oughtred et al., 2021; Rolland et al., 2014) between cofactors and the TFs containing the EDs. As expected, some coactivator hubs such as CREBBP, EP300, and RB1, as well as general TFs such as TAF1 and TAF2, preferentially interact with AD-containing clusters, while co-repressor hubs such as RNF2 and SIN3A preferentially interact with RD-containing clusters (Figures 6 and S6). Other cofactors are more specific to certain ED clusters. For example, mediator complex subunits and nuclear receptor coactivators preferentially interact with clusters 5 and 8, which are enriched in nuclear receptors (Figure 6). Similarly, co-repressor TRIM28 interacts with TFs from clusters 13 and 20, which are highly enriched in KRAB domain-containing ZF-C2H2, as has been previously reported (Friedman et al., 1996), whereas heterochromatin protein CBX5 preferentially interacts with TFs from cluster 20 (Figure 6). Several cofactors are shared between activation and repression domain clusters (e.g., HR, SF1, HDAC1) (Figures 6 and S6). This may be because several EDs can interact with both coactivators and co-repressors, which modulate transcriptional activity under different conditions. However, some of these cases may be related to the fact that PPIs are considered at the whole-protein level and 115 TFs contain both activation and repression domains.

Perspectives and future directions

Most studies of EDs have been conducted on individual TFs, showing that EDs are generally acidic, disordered, and hydrophobic. However, many EDs are not defined by these general rules, making it difficult to predict, identify, and classify EDs and elucidate their functions. Recently, high-throughput studies have been used to identify EDs by tiling through protein sequences genome-wide and to determine the aa features responsible for transcriptional activity. However, EDs that belong to different clusters and that interact with different cofactors may be governed by different sequence features, without a one-rule-fits-all. Furthermore, since ED activity may differ between cell types or may be influenced by ligands and PTMs, many EDs cannot be determined or characterized in single screens. Although our resource is the most comprehensive to date, this only represents ~35% of all human TFs. Further studies, using high-throughput approaches in different cell types and conditions, are needed to identify and characterize EDs for the remaining ~1,000 human TFs. It is unclear how many of these TFs will contain EDs, as several TFs are known to lack EDs and affect transcriptional activity through dimerization or interactions with other TFs.

TFs often have more than one ED. In most cases, how they functionally interact with one another, cofactors, or the mediator complex remain to be determined. Most PPIs between TFs and cofactors have been determined for full-length TFs, rather than EDs, limiting our understanding of the molecular mechanisms by which individual or sets of EDs in a TF function. Systematic interaction mapping assays such as yeast two-hybrid, proximity ligation, and affinity purification followed by mass spectrometry are needed to identify cofactor-ED interactions to increase our understanding of the mechanisms of action of EDs. This, coupled with high-density mutational screens and structure-based modeling, will also provide insights into the molecular consequences of disease mutations in EDs.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2021.11.007.

ACKNOWLEDGMENTS

This work was funded by the National Institutes of Health grants R35 GM128625 and U01 CA232161, awarded to J.I.F.B. We thank Drs. Trevor Siggers, Martha Bulyk, Thomas Gilmore, Matthew Weirauch, and Ana Fiszbein for critically reading and commenting on the manuscript.

AUTHOR CONTRIBUTIONS

J.I.F.B. conceived the project. L.F.S. performed the data analysis. J.I.F.B., C.S.S., I.H., V.X.S., S.Y., and L.F.S. performed the literature curation of EDs. Z.L. generated the database TFRegDB. L.F.S., J.I.F.B., and A.B. wrote the manuscript. All of the authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPORTING CITATIONS

The following reference appears in the supplemental information: Hu et al. (2019).

REFERENCES

Alerasool, N., Lin, Z.-Y., Gingras, A.-C., and Taipale, M. (2021). Identification and functional characterization of transcriptional activators in human cells. bioRxiv. https://doi.org/10.1101/2021.07.30.454360.

Arnold, C.D., Nemčko, F., Woodfin, A.R., Wienerroither, S., Vlasova, A., Schleiffer, A., Pagani, M., Rath, M., and Stark, A. (2018). A high-throughput method to identify trans-activation domains within transcription factor sequences. EMBO J. *37*, e98896.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

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.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842–1855.e16.

Bradner, J.E., Hnisz, D., and Young, R.A. (2017). Transcriptional Addiction in Cancer. Cell *168*, 629–643.

Braun, T., Winter, B., Bober, E., and Arnold, H.H. (1990). Transcriptional activation domain of the muscle-specific gene-regulatory protein myf5. Nature *346*, 663–665.

Brent, R., and Ptashne, M. (1985). A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. Cell *43*, 729–736.

Cantin, G.T., Stevens, J.L., and Berk, A.J. (2003). Activation domain-mediator interactions promote transcription preinitiation complex assembly on promoter DNA. Proc. Natl. Acad. Sci. USA *100*, 12003–12008.

Carrasco Pro, S., Dafonte Imedio, A., Santoso, C.S., Gan, K.A., Sewell, J.A., Martinez, M., Sereda, R., Mehta, S., and Fuxman Bass, J.I. (2018). Global landscape of mouse and human cytokine transcriptional regulation. Nucleic Acids Res. *46*, 9321–9337.

Chen, W., Lam, S.S., Srinath, H., Jiang, Z., Correia, J.J., Schiffer, C.A., Fitzgerald, K.A., Lin, K., and Royer, W.E., Jr. (2008). Insights into interferon regulatory factor activation from the crystal structure of dimeric IRF5. Nat. Struct. Mol. Biol. *15*, 1213–1220.

Chiesa, G., Kiriakov, S., and Khalil, A.S. (2020). Protein assembly systems in natural and synthetic biology. BMC Biol. *18*, 35.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., and Tjian, R. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*, eaar2555.

Collins, T., Stone, J.R., and Williams, A.J. (2001). All in the family: the BTB/POZ, KRAB, and SCAN domains. Mol. Cell. Biol. *21*, 3609–3615.

Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science *311*, 796–800.

Dinkel, H., Van Roey, K., Michael, S., Davey, N.E., Weatheritt, R.J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kuban, M., et al. (2014). The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res. *42*, D259–D266.

Drysdale, C.M., Dueñas, E., Jackson, B.M., Reusser, U., Braus, G.H., and Hinnebusch, A.G. (1995). The transcriptional activator GCN4 contains multiple activation domains that are critically dependent on hydrophobic amino acids. Mol. Cell. Biol. *15*, 1220–1233.

Dyson, H.J., and Wright, P.E. (2016). Role of Intrinsic Protein Disorder in the Function and Interactions of the Transcriptional Coactivators CREB-binding Protein (CBP) and p300. J. Biol. Chem. *291*, 6714–6722.

Epstein, J.A., Glaser, T., Cai, J., Jepeal, L., Walton, D.S., and Maas, R.L. (1994). Two independent and interactive DNA-binding subdomains of the Pax6 paired domain are regulated by alternative splicing. Genes Dev. *8*, 2022–2034.

Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., and Hahn, S. (2020). A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. Mol. Cell *78*, 890–902.e6.

Ferreira, M.E., Hermann, S., Prochasson, P., Workman, J.L., Berndt, K.D., and Wright, A.P. (2005). Mechanism of transcription factor recruitment by acidic activators. J. Biol. Chem. *280*, 21779–21784.

Filtz, T.M., Vogel, W.K., and Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. Trends Pharmacol. Sci. *35*, 76–85.

Foulkes, N.S., Borrelli, E., and Sassone-Corsi, P. (1991). CREM gene: use of alternative DNA-binding domains generates multiple antagonists of cAMP-induced transcription. Cell *64*, 739–749.

Frankel, A.D., and Kim, P.S. (1991). Modular structure of transcription factors: implications for gene regulation. Cell *65*, 717–719.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47* (D1), D766–D773.

Friedman, J.R., Fredericks, W.J., Jensen, D.E., Speicher, D.W., Huang, X.P., Neilson, E.G., and Rauscher, F.J., 3rd (1996). KAP-1, a novel corepressor for the highly conserved KRAB repression domain. Genes Dev. *10*, 2067–2078.

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

Gerber, H.P., Seipel, K., Georgiev, O., Höfferer, M., Hug, M., Rusconi, S., and Schaffner, W. (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science *263*, 808–811.

Giraud, S., Bienvenu, F., Avril, S., Gascan, H., Heery, D.M., and Coqueret, O. (2002). Functional interaction of STAT3 transcription factor with the coactivator NcoA/SRC1a. J. Biol. Chem. *277*, 8004–8011.

Han, B.Y., Seah, M.K.Y., Brooks, I.R., Quek, D.H.P., Huxley, D.R., Foo, C.S., Lee, L.T., Wollmann, H., Guo, H., Messerschmidt, D.M., and Guccione, E. (2020). Global translation during early development depends on the essential transcription factor PRDM10. Nat. Commun. *11*, 3603.

Hermann, S., Berndt, K.D., and Wright, A.P. (2001). How transcriptional activators bind target proteins. J. Biol. Chem. *276*, 40127–40132.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13–23.

Hope, I.A., and Struhl, K. (1986). Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. Cell *46*, 885–894.

Hornbeck, P.V., Kornhauser, J.M., Latham, V., Murray, B., Nandhikonda, V., Nord, A., Skrzypek, E., Wheeler, T., Zhang, B., and Gnad, F. (2019). 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms. Nucleic Acids Res. *47* (D1), D433–D441.

Hu, H., Miao, Y.R., Jia, L.H., Yu, Q.Y., Zhang, Q., and Guo, A.Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res *47*, D33–D38.

Husberg, C., Murphy, P., Martin, E., and Kolsto, A.B. (2001). Two domains of the human bZIP transcription factor TCF11 are necessary for transactivation. J. Biol. Chem. *276*, 17641–17652.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Kozmik, Z., Kurzbauer, R., Dörfler, P., and Busslinger, M. (1993). Alternative splicing of Pax-8 gene transcripts is developmentally regulated and generates isoforms with different transactivation properties. Mol. Cell. Biol. *13*, 6024–6035.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665.

Lambert, S.A., Yang, A.W.H., Sasse, A., Cowley, G., Albu, M., Caddick, M.X., Morris, Q.D., Weirauch, M.T., and Hughes, T.R. (2019). Similarity regression predicts evolution of transcription factor sequence specificity. Nat. Genet. *51*, 981–989.

Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res. *48* (D1), D835–D844.

Levy, Y., and Onuchic, J.N. (2006). Water mediation in protein folding and molecular recognition. Annu. Rev. Biophys. Biomol. Struct. *35*, 389–415.

Li, J., Hsu, A., Hua, Y., Wang, G., Cheng, L., Ochiai, H., Yamamoto, T., and Pertsinidis, A. (2020). Single-gene imaging links genome topology, promoter-enhancer communication and transcription control. Nat. Struct. Mol. Biol. *27*, 1032–1040.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. (2006). Intrinsic disorder in transcription factors. Biochemistry *45*, 6873–6888.

López, A.J. (1995). Developmental role of transcription factor isoforms generated by alternative splicing. Dev. Biol. *172*, 396–411.

Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408.

Ma, J., and Ptashne, M. (1987). Deletion analysis of GAL4 defines two transcriptional activating segments. Cell *48*, 847–853.

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al. (2020). A compendium of mutational cancer driver genes. Nat. Rev. Cancer *20*, 555–572.

Meijer, D., Graus, A., and Grosveld, G. (1992). Mapping the transactivation domain of the Oct-6 POU transcription factor. Nucleic Acids Res. *20*, 2241–2247.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49* (D1), D412–D419.

Neely, K.E., Hassan, A.H., Wallberg, A.E., Steger, D.J., Cairns, B.R., Wright, A.P., and Workman, J.L. (1999). Activation domain-mediated targeting of the SWI/SNF complex to promoters stimulates transcription from nucleosome arrays. Mol. Cell *4*, 649–655.

Oldfield, C.J., and Dunker, A.K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. Annu. Rev. Biochem. *83*, 553–584.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. *30*, 187–200.

Paulsen, R.E., Weaver, C.A., Fahrner, T.J., and Milbrandt, J. (1992). Domains regulating transcriptional activity of the inducible orphan receptor NGFI-B. J. Biol. Chem. *267*, 16491–16496.

Pei, D.Q., and Shih, C.H. (1991). An "attenuator domain" is sandwiched by two distinct transactivation domains in the transcription factor C/EBP. Mol. Cell. Biol. *11*, 1480–1487.

Piskacek, S., Gregor, M., Nemethova, M., Grabner, M., Kovarik, P., and Piskacek, M. (2007). Nine-amino-acid transactivation domain: establishment and prediction utilities. Genomics *89*, 756–768.

Qian, M., Yan, F., Yuan, T., Yang, B., He, Q., and Zhu, H. (2020). Targeting post-translational modification of transcription factors as cancer therapy. Drug Discov. Today *25*, 1502–1512.

Raney, A.K., Easton, A.J., Milich, D.R., and McLachlan, A. (1991). Promoter-specific transactivation of hepatitis B virus transcription by a glutamine- and proline-rich domain of hepatocyte nuclear factor 1. J. Virol. *65*, 5774–5781.

Ravarani, C.N., Erkina, T.Y., De Baets, G., Dudman, D.C., Erkine, A.M., and Babu, M.M. (2018). High-throughput discovery of functional disordered regions: investigation of transactivation domains. Mol. Syst. Biol. *14*, e8190.

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

Roeder, R.G. (2019). 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. Nat. Struct. Mol. Biol. *26*, 783–791.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

Roose, J., Molenaar, M., Peterson, J., Hurenkamp, J., Brantjes, H., Moerer, P., van de Wetering, M., Destrée, O., and Clevers, H. (1998). The Xenopus Wnt effector XTcf-3 interacts with Groucho-related transcriptional repressors. Nature *395*, 608–612.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*, eaar3958.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647–660.

Sainz, M.B., Goff, S.A., and Chandler, V.L. (1997). Extensive mutagenesis of a transcriptional activation domain identifies single hydrophobic and acidic amino acids important for activation in vivo. Mol. Cell. Biol. *17*, 115–122.

Sanborn, A.L., Yeh, B.T., Feigerle, J.T., Hao, C.V., Townshend, R.J., Lieberman Aiden, E., Dror, R.O., and Kornberg, R.D. (2021). Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. eLife *10*, e68068.

Santoso, C.S., Li, Z., Lal, S., Yuan, S., Gan, K.A., Agosto, L.M., Liu, X., Pro, S.C., Sewell, J.A., Henderson, A., et al. (2020). Comprehensive mapping of the human cytokine gene regulatory network. Nucleic Acids Res. *48*, 12055–12073.

Shrinivas, K., Sabari, B.R., Coffey, E.L., Klein, I.A., Boija, A., Zamudio, A.V., Schuijers, J., Hannett, N.M., Sharp, P.A., Young, R.A., and Chakraborty, A.K. (2019). Enhancer Features that Drive Formation of Transcriptional Condensates. Mol. Cell *75*, 549–561.e7.

Sigler, P.B. (1988). Transcriptional activation. Acid blobs and negative noodles. Nature *333*, 210–212.

Staby, L., O'Shea, C., Willemoës, M., Theisen, F., Kragelund, B.B., and Skriver, K. (2017). Eukaryotic transcription factors: paradigms of protein intrinsic disorder. Biochem. J. *474*, 2509–2532.

Staller, M.V., Ramirez, E., Holehouse, A.S., Pappu, R.V., and Cohen, B.A. (2021). Design principles of acidic transcriptional activation domains. bioRxiv, 10.28.359026.

Staller, M.V., Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R.V., and Cohen, B.A. (2018). A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. Cell Syst. *6*, 444–455.e6.

Stoykova, A.S., Sterrer, S., Erselius, J.R., Hatzopoulos, A.K., and Gruss, P. (1992). Mini-Oct and Oct-2c: two novel, functionally diverse murine Oct-2 gene products are differentially expressed in the CNS. Neuron *8*, 541–558.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Tanaka, T., Tanaka, K., Ogawa, S., Kurokawa, M., Mitani, K., Yazaki, Y., Shibata, Y., and Hirai, H. (1997). An acute myeloid leukemia gene, AML1, regulates transcriptional activation and hemopoietic myeloid cell differentiation antagonistically by two alternative spliced forms. Leukemia *11* (*Suppl 3*), 299–302.

Taneri, B., Snyder, B., Novoradovsky, A., and Gaasterland, T. (2004). Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. Genome Biol. *5*, R75.

Tapscott, S.J., Davis, R.L., Thayer, M.J., Cheng, P.F., Weintraub, H., and Lassar, A.B. (1988). MyoD1: a nuclear phosphoprotein requiring a Myc homology region to convert fibroblasts to myoblasts. Science *242*, 405–411.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. *47* (D1), D941–D947.

Teufel, D.P., Bycroft, M., and Fersht, A.R. (2009). Regulation by phosphorylation of the relative affinities of the N-terminal transactivation domains of p53 for p300 domains and Mdm2. Oncogene *28*, 2112–2118.

Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A million peptide motifs for the molecular biologist. Mol. Cell *55*, 161–169.

Tootle, T.L., and Rebay, I. (2005). Post-translational modifications influence transcription factor activity: a view from the ETS superfamily. BioEssays *27*, 285–298.

Tycko, J., DelRosso, N., Hess, G.T., Aradhana, Banerjee, A., Mukund, A., Van, M.V., Ego, B.K., Yao, D., Spees, K., et al. (2020). High-Throughput Discovery and Characterization of Human Transcriptional Effectors. Cell *183*, 2020–2035.e16.

van Mierlo, G., Jansen, J.R.G., Wang, J., Poser, I., van Heeringen, S.J., and Vermeulen, M. (2021). Predicting protein condensate formation using machine learning. Cell Rep. *34*, 108705.

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.

Venkatanarayan, A., Raulji, P., Norton, W., Chakravarti, D., Coarfa, C., Su, X., Sandur, S.K., Ramirez, M.S., Lee, J., Kingsley, C.V., et al. (2015). IAPP-driven metabolic reprogramming induces regression of p53-deficient tumours in vivo. Nature *517*, 626–630.

Vernon, R.M., Chong, P.A., Tsang, B., Kim, T.H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J.D. (2018). Pi-Pi contacts are an overlooked protein feature relevant to phase separation. eLife *7*, e31486.

Wang, F., Marshall, C.B., Yamamoto, K., Li, G.Y., Gasmi-Seabrook, G.M., Okada, H., Mak, T.W., and Ikura, M. (2012). Structures of KIX domain of CBP in complex with two FOXO3a transactivation domains reveal promiscuity and plasticity in coactivator recruitment. Proc. Natl. Acad. Sci. USA *109*, 6078–6083.

Warfield, L., Tuttle, L.M., Pacheco, D., Klevit, R.E., and Hahn, S. (2014). A sequence-specific transcription activator motif and powerful synthetic variants that bind Mediator using a fuzzy protein interface. Proc. Natl. Acad. Sci. USA *111*, E3506–E3513.

Xu, Y., Milazzo, J.P., Somerville, T.D.D., Tarumoto, Y., Huang, Y.H., Ostrander, E.L., Wilkinson, J.E., Challen, G.A., and Vakoc, C.R. (2018). A TFIID-SAGA Perturbation that Targets MYB and Suppresses Acute Myeloid Leukemia. Cancer Cell *33*, 13–28.e8.

Molecular Cell, Volume *82*

Supplemental information

Compendium of human transcription factor effector domains

Luis F. Soto, Zhaorong Li, Clarissa S. Santoso, Anna Berenson, Isabella Ho, Vivian X. Shen, Samson Yuan, and Juan I. Fuxman Bass

Document S1

Literature curation of effector domains

For each of the 1,639 human TFs reported in Lambert et al (Lambert et al., 2018), publications reporting effector domains were identified by searching in PubMed for the TF name in combination with at least one key word associated with the effector domain function ("activation", "transactivation", "repression") or a functional assay ("luciferase", "Gal4", "LexA", "reporter"). Only effector domains tested individually (low-throughput experiments) were considered in our annotation, while domains determined in high-throughput pooled screens were excluded. This is to reduce the impact of false positive predictions from high-throughput screens and because these screens evaluate peptides of defined lengths in single cell types/conditions which may not match the ones where particular effector domains are functionally active.

The amino acid location of the effector domain was obtained by analyzing the text and figures of the manuscripts from protein deletion or effector domain-DBD fusion experiments. The full length of the TF isoform used in the experiments, or the reported amino acid sequence of the domain, were used to match an isoform reported in UniProt (UniProt Consortium, 2021). The sequence and amino acid location of the domain was then obtained from the corresponding UniProt isoform. In cases where experiments were performed using the TF from another vertebrate species, the amino acid sequence reported (or inferred from amino acid location) was aligned to the human sequence to extract the corresponding amino acid sequence and location in a human isoform of the TF. For each effector domain, we annotated the regulatory activity (activation, repression, or bifunctional), the amino acid sequence and location in a UniProt isoform, the assay used to identify the domain, the species in which the domain was identified, whether the effector domain was necessary and/or sufficient, the level of activity, and the PubMed ID. To reduce the chances of missing effector domains, we complemented the PubMed search

by searching in Google images and in reviews for images containing effector domain locations within each TF, followed by PubMed searches for experimental evidence.

We also annotated a confidence score as high (58%), moderate (30%), or low (12%). Highly confident annotations correspond to effector domains that are sufficient with high/moderate transcriptional activity. Moderately confident annotations correspond effector domains that are sufficient with low activity, or necessary with high/moderate activity. Low confident annotations correspond to necessary effector domains with low activity, or cases where no experimental evidence is provided. This general classification was in some cases modified based on additional evidence (e.g., interactions with cofactors) or if the sequence identity of the domain tested was not high compared to the human effector domain sequence.

Localization of TF effector domains

To determine the location of domains within the amino acid sequence of each TF (**Figure 2B**), we calculated the relative position of each domain (activation domains - ADs, repression domains - RDs, and DNA binding domains - DBDs) in each TF where 0 corresponds to the N-terminus and 100 to the C-terminus. To do this, for each TF we calculated a normalization factor as their respective length in amino acids divided by 100. Then, to obtain the relative position of each domain, the N-terminal and C-terminal positions of each domain were divided by their respective normalization factor.

To show this graphically (**Figure 2B**), TFs were arranged in descending order by TF families based on the number of TFs within each family. We only showed TF families with more than 20 TFs in our resource while the remaining TFs were considered as “Others”. Then, in each TF family, TFs were ordered as follows: TFs with only ADs, TFs with only RD, TFs with ADs and

RDs. Finally, inside each subgroup, TFs were ordered based on the first appearance of an effector domain. Each TF was represented as a horizontal line where AD, RD, DBD and the rest of the protein was colored with blue, red, yellow, and grey, respectively. Similarly, we showed the un-normalized sequence and domain positions, centered in the DBDs (**Supplementary Figure S1B**).

Additionally, in each TF family, effector domains were classified based on their relative location within the TFs. Effector domains whose normalized start position was less than or equal to 3 were considered N-terminal, while those with normalized end position greater than or equal to 97 were considered C-terminal. Other cases were considered as internal.

Characterization and amino acid composition of effector domains

The TFs families were obtained from The Human Transcription Factors database (http://humantfs.ccbr.utoronto.ca/) and were used to annotate TFs with effector domains in major families (Lambert et al., 2018). To determine whether effector domains were previously annotated in Pfam (Finn et al., 2016), we downloaded the Pfam database and considered effector domains that: 1) were longer than 10 amino acids, and 2) displayed at least 90% of the effector domain overlapping with a domain annotated in Pfam.

DBD amino acid sequences and coordinates were obtained from CisBP2.0 (Lambert et al., 2019). Disorder, hydrophobicity, charge, and proportion of phosphorylations were calculated for ADs, RDs, and DBDs. Bifunctional domains were excluded from these analyses as only 11 domains are annotated in our database. For TFs with more than one effector domain, the properties were calculated for each domain individually. In the case of multiple DBDs, as in ZF-C2H2 TFs, we concatenated all DBDs into one DBD.

The disorder of effector domains and DBDs was calculated using the AlphaFold Database (Jumper et al., 2021). First, we determined the disordered regions for each TF based on the per-residue confidence score (pLDDT) using the TF .cif files. Regions with two or more amino acids with scores lower than 50 were considered as disordered regions. Then, for each domain (effector domain and DBD), we calculated the proportion of disordered amino acids as the fraction of amino acids in the domain belonging to a disordered region.

The hydrophobicity score was obtained as the proportion of hydrophobic amino acids (F, I, L, M, W, A, Y, P) relative to the domain length. The charge was calculated using the Localcider Python package (http://pappulab.github.io/localCIDER/) (Ginell and Holehouse, 2020). Phosphorylation sites were downloaded from PhosphoSitePlus (https://www.phosphosite.org/staticDownloads) (Hornbeck et al., 2019) and the proportion of phosphorylation was calculated as the number of phosphorylation sites in each domain divided by their length in amino acids. We considered only phosphorylation events reported with at least 5 references in the PhosphoSitePlus database. A Wilcoxon-test was performed to compare charge, disorder, and hydrophobicity between effector domains (either ADs and RDs) and DBDs.

To annotate regions in effector domains that have amino acid composition bias, we used fLPS (Harrison, 2017) with the following parameters -m 5 -M 100 -o short, and considered those regions that are enriched with a single or multiple amino acids. We considered only enriched regions that were longer than 10 amino acids. Then, for each effector domain, a score of 1 was assigned for each amino acid if there was at least one region inside the effector domain enriched with that amino acid. Otherwise, it was 0. Finally, amino acid density was calculated for each effector domain and DBD as the number of each amino acid divided by the domain length.

Liquid-Liquid Phase Separation (LLPS) prediction

To evaluate if effector domains are involved in LLPS we performed two different analyses. First, we compared the LLPS promotion scores between TFs-AD, TFs-RD, and TFs-Bif. We used PSAP (van Mierlo et al., 2021) to obtain the LLPS probability for each human TF. Similarly, we obtained the LLPS score from another predictor based on pi-interactions (Vernon et al., 2018). Then, we performed a Wilcoxon test to compare both the LLPS probability and score between TFs-AD, TFs-RD, and TFs-Bif. Second, we evaluated the proportion of effector domains and DBDs containing amino acid contexts predicted to promote LLPS. To do this, we obtained the score for each amino acid in effector domains and DBDs (Vernon et al., 2018), and calculated the proportion of ADs, RDs, and DBDs containing at least one amino acid with a score greater than 4. Significance was evaluated using a proportion comparison test.

Effector domains in proteoforms

Transcripts with available experimental evidence (Minimum Transcription Support Level and in GENCODE Basic) were obtained from the GENCODE v.30 database (Frankish et al., 2019). Transcripts that are predicted to produce the same amino acid sequence, or sequences that differ due to genetic variation, were merged into the same proteoform. For each TF, we calculated the number of proteoforms that have (1) effector domain and DBD unaffected, (2) effector domain affected, (3) DBD affected, and (4) both domains affected by deletions, truncations, or indels. A similar calculation was performed for DBDs. The affected domains were identified by aligning these domains with their different proteoforms using Needle-Wunschman global alignment in BioPython (Cock et al., 2009) with the following parameters: gapopen = 10, gapextend = 0.5 matrix = BLOSUM62. Then, we used an in-house Python script to calculate an identity-based score for each alignment (effector domain or DBD versus proteoform). This was defined as the number of identical amino acids divided by the length of the aligned sequence. If the identity-based score of the domain in a proteoform was < 90%, the domain was considered affected in

the corresponding proteoform. A Kolmogorov-Smirnov test was performed between the distributions of identity-based score of effector domains and DBDs. A domain was classified as "intact domain" within a proteoform if the domain had an identity-based score higher than 90% and at most only one substitution, as "domain with indels" if the identity-based score was 30-90%, and as "deleted domain" if the identity-based score was lower than 30%.

To evaluate any bias due the domain length, we calculated the number of proteoforms with the affected domain across bins of different domain lengths. The bin selected was 50 amino acids with a step of 10 amino acids. For each bin, we calculated the proportion of proteoforms with affected effector domains or DBDs.

Evolutionary and population conservation of effector domains

We used the Ensembl rest API to obtain the orthologs of TFs in 27 vertebrate species. Then, we performed a global alignment between each domain (effector domains and DBDs) and each ortholog TF using the BioPython package (Cock et al., 2009). If a TF had multiple effector domains (or multiple DBDs), they were concatenated into one sequence. The alignment was performed between each domain (effector domain or DBD) and each ortholog TF with the following parameters gapopen =10, gapextend= 0.5 and BLOSUM62 matrix. Then, we assigned the percentage identity to each alignment as the number of identical amino acids divided by the length of the respective domain. To obtain the species dendogram, we retrieved the species relation from the Ensembl project (Howe et al., 2021) and generated the dendogram using the package "phytools" and "ape" in R v4.05. The divergence time between each species and *Homo sapiens* was obtained from TimeTree (Kumar et al., 2017), and the amino acid conservation for both effector domains and DBDs at each evolutionary time was represented.

To determine the density of genomic variants, we first obtained the genomic coordinates of each effector domain, DBD, and the full length protein. To do this, we retrieved the ENST code for each TF using the GENCODE.v38 database. When available, we used the UNIPROT ID of each TF to find their respective ENST code. In other cases, we used an in-house Python script to map the amino acid sequences to each isoform reported in GENCODE until we found the perfect match. Then, we used these transcript IDs to obtain the nucleotide coordinates for each exon, and lastly, obtain the nucleotide coordinate for each domain (ADs, RDs, Bifs, DBDs) from their respective amino acid positions. All nucleotides coordinates were translated to their respective amino acid sequence as a verification step.

To map the genomic variants into the domains, we downloaded the gnomAD database (Karczewski et al., 2020) and used "Bedtools intersect" to determine the variants in each effector domain, DBD, and full length protein. Then, we removed variants that correspond to more than one nucleotide and classified the single nucleotide variants into synonymous and non-synonymous using a Python script. The density of non-synonymous variants for each domain (AD, RD, DBD) and full length TF were calculated as the number of non-synonymous variants in the corresponding amino acid region divided by its length in nucleotides. Multiple effector domains (or DBDs) in a TF were concatenated and considered as a unique domain. Variants residing in the same genomic position were considered different. To determine statistical enrichment of variants in the effector domain versus DBD of a TF, we performed a Fisher's exact test considering the number of variant and non-variant nucleotides in each domain, and performed a Benjamini-Holchberg correction with a cutoff of 0.1 to correct for multiple hypothesis testing.

Density of mutations in effector domains

To evaluate the density of mutations in effector domains and DBDs associated with diseases and cancer, we downloaded mutations from the ClinVar (Landrum et al., 2020) and COSMIC (Tate et al., 2019) databases, respectively. We calculated the density of mutations in the effector domains, DBDs, and full length TF from ClinVar and COSMIC mutations as we did for gnomAD genetic variants. Only variants that were annotated as “Pathogenic” in COSMIC, and “Pathogenic” and “Likely Pathogenic” in the ClinVar database were considered. In addition, we evaluated whether mutations in effector domains and DBDs are associated with different cancer types. To do this, we considered the “primary site“ as the main cancer type of each somatic mutation in each sample using the COSMIC annotation file. In cases where a TF contained more than one effector domain, these were concatenated in one group to be evaluated as effector domains. Multiple DBDs in a TF were concatenated in a similar manner. In cases where a mutation was associated with multiple cancer types, all of these were considered. Then, we performed a Fisher’s exact test for each TF comparing the number of mutations associated with different cancer types in effector domains and DBDs and p-values were adjusted by Benjamini-Hochberg correction considering a cutoff of 0.1.

Classification of effector domains

To classify effector domains, we built 6 similarity matrices (6 x 924 x 924) leveraging different characteristics of effector domains. (1) Sequence similarity matrix: We calculated a sequence identity score between 0-1 for each pair of effector domains using a global Needleman-Wunsch alignment. All identity scores lower than 0.5 were replaced with 0 to avoid high background noise when clustering. (2) Regulatory function matrix: We assigned a score of 1 for a pair of effector domains that have the same regulatory function (AD-AD, RD-RD, Bif-Bif), a score of 0.5 if the effector domains share a regulatory function (AD-Bif, RD-Bif), and a score of 0 if the effector

domains do not share a regulatory function (AD-RD). (3) Amino acid composition matrix: First, we used flps (https://github.com/pmharrison/flps) to detect low complexity sequences (i.e., enriched amino acids in short stretch regions) in each effector domain. The software was run using default parameters and we considered regions that were enriched with single and multiple amino acids. Then, we generated a matrix where rows are effector domains, columns are amino acids, and the values 1 or 0 indicate enrichment or no enrichment of the amino acid in the effector domain, respectively. Finally, we calculated the Jaccard index (Fuxman Bass et al., 2013) for each pair of effector domains to generate the amino acid composition similarity matrix. (4-6) Charge , disorder, and hydrophobicity matrices: First, we calculated the charge, disorder, and hydrophobicity for each effector domain. Then, for each parameter, we generated a matrix where a similarity score was calculated for each pair of effector domains as follows (example shown for charge calculation):

$$Score_{charge}\ (x_1, x_2) = 1 - \frac{|charge(x_1) - charge(x_2)|}{Max(charge\ differences)}$$

where x1 = effector domain 1 and x2 = effector domain 2

For example if two effector domains have charge values of 0.7 and 0.3 and the maximum differences in all possible combinations of effector domains is 1.4, the charge similarity score would be 1 - (0.4/1.4) = 0.714.

To give each matrix a similar weight, we normalized each matrix by dividing each value by the standard deviation between the values with the corresponding matrix. Then, we generated an effector domain similarity matrix by adding each of these four matrices with the following weights: Sequence Similarity Matrix = 2, Regulatory Function Matrix = 2, Amino acid composition matrix = 1, Charge matrix = 1, Disorder matrix = 1, Hydrophobicity matrix = 1. This matrix was then converted to a distance matrix using the “sim2dist” function in R v4.05. Using this effector

domain distance matrix, we performed hierarchical clustering using the "hclust" function in R with the "complete" agglomeration method. To select an appropriate number of clusters, we obtained clusters using the "cutree" function with the parameter k from 2 to 100 and selected the minimal k value where the maximum number of effector domains in any cluster was less than 100 (k = 63). Only clusters containing more than 10 effector domains are shown and included in the analyses. For each of the 20 clusters obtained, we showed eight characteristics: (1) the number of effector domains, their median (2) charge, (3) hydrophobicity, (4) disorder, and (5) length, (6) enrichment of amino acids, (7) proportion of domains in each TF family, and (8) proportion of domains interacting with cofactors.

To annotate interactions with cofactors, we first downloaded the list of cofactors from AnimalTFDB 3.0 (Hu et al., 2019) and protein-protein interactions between TFs and cofactors from HuRI (Luck et al., 2020), Lit-BM (Luck et al., 2020), and BioGRID (Oughtred et al., 2021) databases. From BioGRID, we only considered interactions with at least one report of physical evidence.

References

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-1423.

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*, D279-285.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., *et al.* (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res *47*, D766-D773.

Fuxman Bass, J.I., Diallo, A., Nelson, J., Soto, J.M., Myers, C.L., and Walhout, A.J. (2013). Using networks to measure similarity between genes: association index selection. Nat Methods *10*, 1169-1176.

Ginell, G.M., and Holehouse, A.S. (2020). Analyzing the Sequences of Intrinsically Disordered Regions with CIDER and localCIDER. Methods Mol Biol *2141*, 103-126.

Harrison, P.M. (2017). fLPS: Fast discovery of compositional biases for the protein universe. BMC Bioinformatics *18*, 476.

Hornbeck, P.V., Kornhauser, J.M., Latham, V., Murray, B., Nandhikonda, V., Nord, A., Skrzypek, E., Wheeler, T., Zhang, B., and Gnad, F. (2019). 15 years of PhosphoSitePlus(R): integrating post-translationally modified sites, disease variants and isoforms. Nucleic Acids Res *47*, D433-D441.

Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J.*, et al.* (2021). Ensembl 2021. Nucleic Acids Res *49*, D884-D891.

Hu, H., Miao, Y.R., Jia, L.H., Yu, Q.Y., Zhang, Q., and Guo, A.Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res *47*, D33-D38.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A.*, et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P.*, et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol *34*, 1812-1819.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650-665.

Lambert, S.A., Yang, A.W.H., Sasse, A., Cowley, G., Albu, M., Caddick, M.X., Morris, Q.D., Weirauch, M.T., and Hughes, T.R. (2019). Similarity regression predicts evolution of transcription factor sequence specificity. Nat Genet *51*, 981-989.

Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C.*, et al.* (2020). ClinVar: improvements to accessing data. Nucleic Acids Res *48*, D835-D844.

Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B.*, et al.* (2020). A reference map of the human binary protein interactome. Nature *580*, 402-408.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F.*, et al.* (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci *30*, 187-200.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., *et al.* (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res *47*, D941-D947.

UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res *49*, D480-D489.

van Mierlo, G., Jansen, J.R.G., Wang, J., Poser, I., van Heeringen, S.J., and Vermeulen, M. (2021). Predicting protein condensate formation using machine learning. Cell Rep *34*, 108705.

Vernon, R.M., Chong, P.A., Tsang, B., Kim, T.H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J.D. (2018). Pi-Pi contacts are an overlooked protein feature relevant to phase separation. Elife *7*.