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*.