Article

Large-scale mapping and mutagenesis of human transcriptional effector domains

https://doi.org/10.1038/s41586-023-05906-y

Received: 14 July 2022

Accepted: 1 March 2023

Published online: 05 April 2023

Check for updates

Nicole DelRosso[1], Josh Tycko[2], Peter Suzuki[3], Cecelia Andrews[4], Aradhana[2], Adi Mukund[1], Ivan Liongson[5], Connor Ludwig[3], Kaitlyn Spees[2], Polly Fordyce[2,3,6,7], Michael C. Bassik[2] & Lacramioara Bintu[3]✉

Human gene expression is regulated by more than 2,000 transcription factors and chromatin regulators[1,2]. Effector domains within these proteins can activate or repress transcription. However, for many of these regulators we do not know what type of effector domains they contain, their location in the protein, their activation and repression strengths, and the sequences that are necessary for their functions. Here, we systematically measure the effector activity of more than 100,000 protein fragments tiling across most chromatin regulators and transcription factors in human cells (2,047 proteins). By testing the effect they have when recruited at reporter genes, we annotate 374 activation domains and 715 repression domains, roughly 80% of which are new and have not been previously annotated[3–5]. Rational mutagenesis and deletion scans across all the effector domains reveal aromatic and/or leucine residues interspersed with acidic, proline, serine and/or glutamine residues are necessary for activation domain activity. Furthermore, most repression domain sequences contain sites for small ubiquitin-like modifier (SUMO)ylation, short interaction motifs for recruiting corepressors or are structured binding domains for recruiting other repressive proteins. We discover bifunctional domains that can both activate and repress, some of which dynamically split a cell population into high- and low-expression subpopulations. Our systematic annotation and characterization of effector domains provide a rich resource for understanding the function of human transcription factors and chromatin regulators, engineering compact tools for controlling gene expression and refining predictive models of effector domain function.

Large scale efforts have mapped where in the human genome transcription factors (TFs) and chromatin regulators (CRs) bind[6,7]. However, equivalent maps of transcriptional effector domains (EDs) are incomplete: we are currently missing ED annotations for about 60% of human TFs[8]. Moreover, the sequence characteristics of what makes a good human activation or repression domain are still under investigation. One useful assay for characterizing individual EDs and testing specific sequence requirements consists of recruitment of domains and mutants to reporter genes (reviewed in ref. 8). This approach has been extended from recruiting single domains to high-throughput assays in yeast[9–12], *Drosophila*[13–16] and human cells with a subset of transcriptional domains[4,17] or a subset of full-length TFs[5]. These works have extended our list of EDs and have set the stage for systematically mapping EDs across the thousands of human transcriptional proteins.

To map the human EDs at unprecedented scale and resolution, we synthesized DNA sequences encoding 80 amino acid (aa) segments that tile across 1,292 human TFs[1] and 755 CRs[2] (hereafter CRTF tiling library) with a 10-aa step size between segments (Fig. 1a, Extended Data Fig. 1a, Supplementary Table 1 and Methods). This library, consisting of 128,565 sequences, was cloned into a lentiviral vector, where each protein tile is expressed as a fusion protein with rTetR (a doxycycline inducible DNA-binding domain) and delivered as a pool to K562 cells at a low lentiviral infection rate, such that each cell contains a single rTetR-tile. The cells contain a reporter with binding sites for rTetR. This reporter consists of a synthetic surface marker that allows facile magnetic separation of cells for high-throughput measurements, and the fluorescent protein citrine for flow cytometry quantification during individual validations. The reporter gene is driven by either a minimally active minCMV promoter for identifying activators, or constitutively active pEF promoter for finding repressors. To simultaneously measure the effector function of these sequences, we used a high-throughput recruitment assay we recently developed: HT-recruit[4]. After treating the cells with doxycycline, which recruits each CRTF tiling library member to the reporter, we magnetically separated the cells into ON and OFF populations and sequenced the tiles to identify sequences enriched in each cell population (Methods and Extended Data Fig. 1b,c). Each screen

[1]Biophysics Program, Stanford University, Stanford, CA, USA. [2]Department of Genetics, Stanford University, Stanford, CA, USA. [3]Department of Bioengineering, Stanford University, Stanford, CA, USA. [4]Department of Developmental Biology, Stanford University, Stanford, CA, USA. [5]Department of Biology, Stanford University, Stanford, CA, USA. [6]ChEM-H Institute, Stanford University, Stanford, CA, USA. [7]Chan Zuckerberg Biohub, San Francisco, CA, USA. ✉e-mail: lbintu@stanford.edu

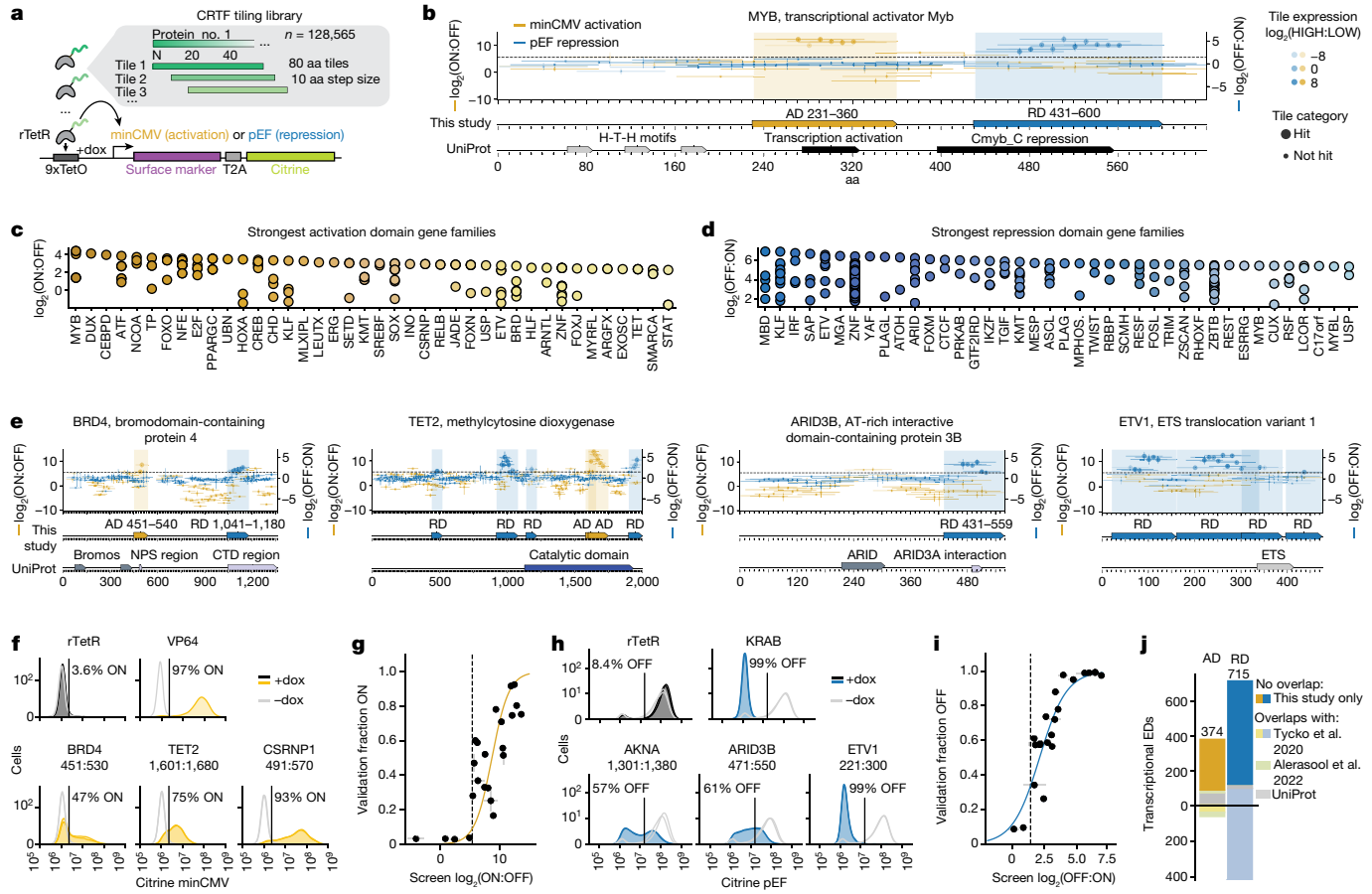


Fig. 1 | High-throughput tiling screen across 2,047 human TFs and CRs finds hundreds of EDs. a, Schematic of HT-recruit. A pooled library of protein tiles is synthesized, cloned as a fusion to rTetR-3xFLAG and delivered to reporter cells. The reporter includes fluorescent citrine and a synthetic surface marker for magnetic bead separation of ON and OFF cells. **b**, MYB's activation and repression enrichment scores. Each horizontal line is a tile, and each vertical bar is the range of measurements from two biologically independent screens. Dashed horizontal line is the hit calling threshold based on random controls (Methods). Points with larger marker sizes are hits in the validation screen. Marker hues indicate FLAG-stained expression levels. **c**,**d**, Distribution of the strongest EDs from the top 40 gene families: activation (**c**) and repression (**d**). Average enrichment scores are from the maximum tile within each domain measured in the validation screen ($n = 2$). All points shown are above the hit thresholds. **e**, Tiling results for BRD4, TET2, ARID3B and ETV1 ($n = 2$ screens, dots are the mean, vertical bars the range). **f**, Citrine fluorescence distributions from flow cytometry for cell lines expressing individual activating tiles ($n = 2$). Vertical line is the citrine gate used to determine the fraction of cells ON (written above each distribution). **g**, Comparison between screen measurements and individually recruited tiles at minCMV ($n = 2$, dots are the mean, bars the range) with logistic model fit plotted as solid line ($r^2 = 0.67$, $n = 23$). Dashed line is the hits threshold. **h**, Flow cytometry citrine distributions for individual validations of repressing tiles ($n = 2$). **i**, Comparison between screen measurements and individually recruited tiles at pEF ($n = 2$, dots are the mean, bars the range) with logistic model fit as solid line ($r^2 = 0.84$, $n = 22$). **j**, ED counts identified in this study are shown above the black line, and domain counts from previous work not tested in this study are shown below. RDs are annotated from tiles that were hits in both pEF and PGK promoter screens (Extended Data Fig. 4).

was reproducible across two biological replicates (Extended Data Fig. 1d,e). We drew thresholds for calling hits on the basis of the scores of random negative controls (Extended Data Fig. 1d,e and Methods). Of the positive control domains for activation and repression, 90 and 92% (ref. 4), respectively, were hits above this threshold, as expected (Supplementary Table 1). Among the tiles shared with our previous screen[4], we identified an extra subset of tiles that were only hits in this repression screen and whose activity validated in individual flow cytometry experiments (Extended Data Fig. 1f,g), showing this screen had better sensitivity. Overall, these results demonstrated HT-recruit reliably identified EDs while using an order-of-magnitude larger library than our previous experiments[4].

Measured transcriptional strength depends not only on the intrinsic potential of the sequence but also on the levels at which individual tiles are expressed. All library members contain a 3xFLAG tag, allowing us to measure each fusion protein's expression levels by staining with an anti-FLAG antibody, FACS sorting the cells into FLAG HIGH and LOW populations (Extended Data Fig. 2a and Supplementary Fig. 3), and measuring the abundance of each member in the two populations by sequencing the domains (Extended Data Fig. 2b). These FLAG scores from the high-throughput measurements can identify proteins that are not expressed, as determined from individual validations using western blotting (Extended Data Fig. 2c), and were used when annotating EDs, allowing us to filter out false negative library members that have lower activation or repression scores due to low expression (Extended Data Fig. 2d and Methods).

To further confirm all the hits and help remove false positives, we screened a smaller library containing only the activating and repressive hit tiles (hereafter validation screen; Supplementary Table 1 and Methods). Because of their small size, these screens had better separation purity (Extended Data Fig. 3a,b) and could be screened at tenfold higher coverage, which resulted in higher reproducibility than the original, larger screens (Extended Data Fig. 3c,d) and even better correlation between screen scores and individual validations (Extended Data Fig. 3e,f). About 80% of the hits were confirmed as hits in these validation screens (Supplementary Table 1 and Extended Data Fig. 3c,d).

We only considered these confirmed sequences in subsequent analyses (Supplementary Table 1).

Using these filtered tiling data, we annotated EDs from contiguous hit tiles (Methods, Fig. 1b and Supplementary Table 2). Doing so can accurately identify EDs previously annotated in UniProt, for example MYB's EDs (Fig. 1b). Some of the strongest EDs come from gene families with some family members already containing annotated activation domains (ADs) (for example, ATF and NCOA) and repression domains (RDs) (for example, KLF and ZNF), making us more confident our screens returned reliable results (Fig. 1c,d). TFs from certain gene families (for example, KLF and KMT) contain both strong ADs and RDs, which highlights our results can identify bifunctional transcriptional regulators. In total, 12% of the proteins screened are bifunctional and 77% of proteins have at least one ED (Supplementary Table 2).

Furthermore, this method allows us to discover previously unannotated EDs (Fig. 1e). For example, we found both a new AD and four new RDs within the DNA demethylating protein, TET2. We validated tens of these new EDs by individually cloning them, creating stable cell lines and measuring their effect using flow cytometry after dox-induced recruitment (Fig. 1f,h and Supplementary Table 3). In these experiments, fluorescence distributions are often not unimodal, most likely due to stochastic gene expression: bursting in the case of activation[18,19] and stochastic silencing in the case of repression[20]. We used these results to validate our screen thresholds: all tiles above the thresholds had activity and no tiles below did (Fig. 1g,i).

Forty-five of the proteins tiled here were recently screened for activation in human embryonic kidney 293T (HEK293T) cells, but tiled with smaller fragments[5]. The two studies show good agreement: 19 proteins do not activate in both screens, and 15 proteins do (Extended Data Fig. 4a). The proteins that only activate in one of the studies could represent activators that are unique to the specific context (cell type for example) but could also reflect the difference in length. For example, KLF6 tiles that only activated with smaller fragments overlapped a RD in our measurements with longer tiles (Supplementary Fig. 2). Whereas longer tiles can possibly capture large ADs, shorter peptides are more likely to find small ADs that are near RDs.

Previous screens in yeast have led to the development of a machine learning model (PADDLE[12]) capable of predicting activation levels from sequence alone with an area under the precision-recall curve of 81%. If the sequence properties that drive activation in humans are like those in yeast, we would expect PADDLE to predict human ADs with similar accuracy. Whereas PADDLE was able to predict 70% of our ADs, the domains that PADDLE predicted to be activating were more negatively charged than the ADs it missed (Extended Data Fig. 4b), suggesting that in human cells there are more non-acidic activator classes compared to yeast.

Because there are no other comprehensive studies in human cells or predictive models with which we can compare our RDs, we repeated the repressive measurements with the entire CRTF library at a second constitutive promoter, PGK. Even though this promoter is weaker than pEF (Extended Data Fig. 4c), we were able to magnetically separate the silent and active cells (Extended Data Fig. 4d) and observed good reproducibility (Extended Data Fig. 4e). Of the hit tiles that showed up in the pEF and PGK screens, 92% also showed up as hits in the pEF validation screen (Extended Data Fig. 4f), suggesting higher confidence results when combining both screens. Taking the maximum tile's enrichment scores within each RD showed that 715 RDs were shared across both screens (Extended Data Fig. 4g,h). Together, these results indicate that at the 80 aa scale there are more sequences across the CRs and TFs that can work as repressors versus activators. In total, 291 out of 374 ADs and 592 out of 715 RDs are new compared to previous annotations[3–5] (Fig. 1j).

Activation domain sequence characteristics

ADs have been classified by the abundance of particular amino acids such as acidic, glutamine- and proline-rich sequences[21,22]. Acidic residues are essential for function in all yeast ADs[12] and some human ADs[17]. Certain human ADs have compositional biases that are not present in other organisms, often containing stretches of homotypic repeats[23] (for example, QQQQ). Furthermore, some aa are enriched in human ADs, in particular the hydrophobic residues, aromatics (W, F, Y) and leucines (L)[17]. It remains unclear whether these enriched residues are necessary for activation.

Our large set of new ADs provides a great opportunity to systematically quantify the prevalence of each of these sequence properties. 45% of activating tiles contain a compositional bias (Fig. 2a), in which serine and proline are the most abundant. Consistent with these observations, when we further normalize the aa frequencies in the AD sequences by the non-hit sequences, there is an enrichment in certain hydrophobic, acidic, serine and proline residues (Fig. 2b).

Despite being well-documented[22,24], we found very few Q-rich ADs (Fig. 2a, $n = 10$ and Supplementary Table 2). Annotated Q-rich ADs are longer than 80 aa (ref. 25), suggesting our tiling approach might have missed them. Alternatively, Q-rich ADs could be relatively weak and require other TFs to activate. Recruitment of SP1's two annotated Q-rich ADs[25] (longer than 80 aa) did not activate minCMV (Extended Data Fig. 5a). However, including a short, acidic AD upstream of the Q-rich domains was sufficient for SP1's 'tAD A' to activate (Extended Data Fig. 5a). This result supports the previous observations that acidic and Q-rich domains work synergistically in human cells[26].

To determine which amino acids are necessary for activation, we took a deletion scanning approach[27]: we measured the activity of mutant ADs containing consecutive small deletions (Extended Data Fig. 5b and Methods). Although most (61%) deletions do not affect activation, we found at least one deletion that was well-expressed and could abolish activator function in most of the pilot ADs (20 out of 24 with activity at minCMV) (Supplementary Table 1). To confirm whether this approach could resolve residues necessary for activity, we compared our deletion scan data from P53 to UniProt and found residues 20–22 (DLW) within one region and residue W52 within another were necessary for activity, corresponding to UniProt-annotated TAD I and TAD II[28] (Extended Data Fig. 5b). Furthermore, individual validations of deletions including these residues confirmed complete loss of activity (Extended Data Fig. 5c).

Confident in our deletion scan approach, we designed a second library of 10 aa deletions across the maximum activating tile from each AD, resulting in 304 total deletion scans (Supplementary Table 4). We measured activation using the minCMV reporter and HT-recruit workflow described in Fig. 1a (Extended Data Fig. 5d–f) and filtered out mutants that were poorly expressed on the basis of FLAG staining (Extended Data Fig. 5g,h). Across each of these expression-filtered deletion scans we classified deletions according to their effect on activation (Fig. 2c). Using these data, we can determine which compositionally biased residues are important for function and which are not: for example, whereas NFAT5's AD has a patch of four serines near the C terminus, deleting those residues had no effect on activation (Fig. 2c and Extended Data Fig. 6a). Applying this analysis to all ADs containing a homotypic repeat, we find serine, proline, acidic, glutamine and glycine homotypic repeats were more often found in deletions that had no effect on activation than in deletions that decreased activation (Fig. 2d). Therefore, homotypic repeats of these amino acids are generally not necessary for activation.

The deletion scans also identify the necessary sequence for activation of each tile: sequences that, once removed, completely abolished activation (Fig. 2c). We were able to annotate at least one necessary sequence (median length of 10 aa) in most (69%) of our screened ADs, and most (61%) ADs have several necessary sequences (Fig. 2c and Supplementary Table 4). Nearly every necessary sequence (96%) contained a W, F, Y or L.

To validate this enrichment of specific hydrophobic residues, we rationally designed mutant libraries in which we systematically

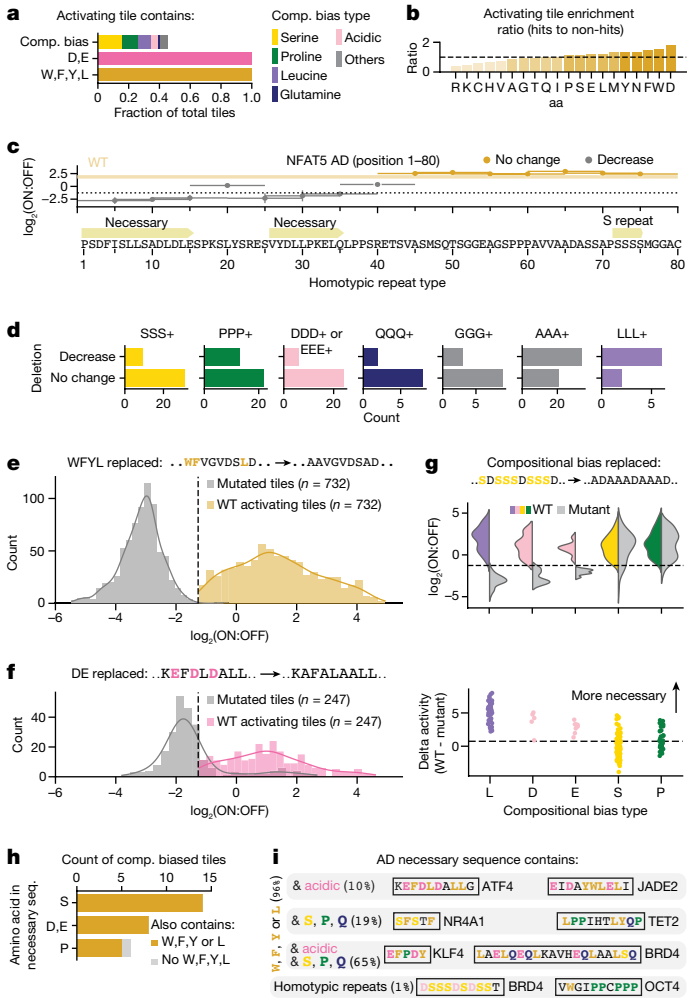


Fig. 2 | Hydrophobic amino acids interspersed with acidic, serine, proline or glutamine residues are necessary for AD activity. a, Fraction of activating tiles that contain compositional (comp.) biases. **b**, Enrichment ratio for each aa across all activating tiles. Dashed line is at 1. **c**, Deletion scan across NFAT5's AD. Yellow rectangle is WT enrichment score, its height the range of two biologically independent screens. Each horizontal line represents those residues that were were deleted, dots are the mean, vertical bars the range and *P* values less than 0.05 (one-sided *z* test compared to WT) are labelled in grey as a decrease. **d**, Counts of deletion sequences containing a homotypic repeat of three or more aa of the indicated type binned according to their effect compared to WT (Fisher's exact test compared with AAA+ and LLL+ distribution, two-sided, Ser $P = 5.1 \times 10^{-5}$, Pro = 1.9×10^{-2}, acidic 1.2×10^{-4}, Gln = 1.5×10^{-2}, Gly = 2.3×10^{-2}). **e**, Distribution of average activation enrichment scores ($n = 2$) for WT and W,F,Y,L mutant tiles for all well-expressed W,F,Y,L-containing activating tiles (Mann–Whitney one-sided *U*-test, $P = 9.2 \times 10^{-241}$). **f**, Distribution of average activation enrichment scores ($n = 2$) for WT and D,E mutant tiles for all well-expressed D,E-containing activating tiles (Mann–Whitney one-sided *U*-test, $P = 2.6 \times 10^{-61}$). **g**, The top shows the distributions of average activation enrichment scores ($n = 2$) for WT (colours) and comp. bias mutants (grey). The bottom shows the mutant enrichment scores subtracted from WT plotted for each comp. bias that was replaced with Ala. The dashed line is twice the average standard deviation (across all mutants) above 0. The probability we would observe these distributions for L, 7.7×10^{-19}; D, 0.0006; E, 0.0005; S, 0.56 and P, 0.006 (Mann–Whitney one-sided *U*-test). **h**, Counts of all necessary regions within comp. biased tiles that lost activity on mutation, coloured by containing W, F, Y, L or not (Fisher's exact test, two-sided, compared to the same tiles' comp. biased sequences that had no change on deletion: Ser $P = 3.8 \times 10^{-4}$, acidic $P = 3.0 \times 10^{-3}$, Pro $P = 5.5 \times 10^{-1}$). **i**, Summary of findings showing that necessary AD sequences consist of hydrophobic residues that are interspersed with acidic, prolines, serines and/or glutamine residues.

replaced every aa of a particular type within the sequence with alanines (Supplementary Table 4). Replacement of all W, F, Y or Ls with alanine (range 3–24 aa replaced/80 aa tile, median of 10 aa) in all our activating tiles resulted in a total loss of activation (Fig. 2e). The one exception that remained active was within DUX4, and the mutation did make it weaker (Extended Data Fig. 6b). This systematic loss of activation was not due to a decrease in protein expression, as measured by FLAG staining (Extended Data Fig. 6c). There is no correlation between the overall count of these residues within tiles and a tile's activation strength (Extended Data Fig. 6d), probably suggesting these residues mediate interactions necessary for activity, and the placement of these residues is more important than the overall count. This means ADs from 258 different proteins require at least some aromatic or leucine residues to activate.

We next replaced all acidic residues with alanine in all activating tiles. More than half of the acidic mutants had reduced expression (Extended Data Fig. 6c). These results indicate acidic residues increase protein levels, at least in the context of ADs. Of the remaining 247 well-expressed activating tile mutants, most mutants lost the ability to activate (Fig. 2f, $n = 196$). The mutants with no change in activity had significantly fewer acidic residues than the tiles whose mutants had a decreasing effect (Extended Data Fig. 6e), supporting the idea that acidic ADs are not the only class of human ADs.

Intrigued by what other compositional biases could be functional in human ADs, we next tested the necessity of other frequently appearing residues by replacing them with alanine. Consistent with the results above, all tiles with leucine and acidic compositional biases lost activity once mutated (Fig. 2g). Removal of serine and proline compositional biases had more mild effects: most mutants still had activity (Fig. 2g, top), even though the strength of activation decreased for a subset of them (Fig. 2g, bottom).

Wanting to follow up more on the compositionally biased tiles that decreased activity on compositional bias removal (Fig. 2g), we next analysed the set of necessary sequences (as determined from the deletion scans) from the compositionally biased activating tiles that lost activity on bias removal (Fig. 2g, bottom). For each bias type, most necessary sequences also contain a W, F, Y or L (Fig. 2h), suggesting their placement next to hydrophobic residues is important for their function.

In summary, sequences that are necessary for activation consist of certain hydrophobic residues (W, F, Y and/or L) that are interspersed with acidic, proline, serine and/or glutamine residues (Fig. 2i and Extended Data Fig. 6f). Although previous work has shown that homopolymer stretches of glutamine and proline are sufficient to activate a weak synthetic reporter[23], we find most glutamine and proline repeats within ADs of the human CRs and TFs are not part of the sequence necessary for activation.

Repression domain sequence characteristics

Repressing tile sequences have significantly more predicted secondary structure than activating tile sequences (Extended Data Fig. 7a). Therefore, we needed to take a different approach for understanding RD sequence characteristics. Instead of looking at RD sequence compositions, we first set out to classify the RDs by their potential mechanism. We used the Eukaryotic Linear Motifs (ELM)[29] database to search for corepressor interaction motifs (Methods) and UniProt[3] to search for domain annotations. We observe 72% of the RDs overlap diverse annotations, such as sites for SUMOylation, zinc fingers (ZFs), SUMO-interacting motifs, corepressor binding motifs, DNA-binding domains (including homeodomains, consistent with previous results)[4] and dimerization domains (Fig. 3a). To address whether these annotations are necessary for repression, we rationally designed mutant libraries that replaced sections of 1,313 repressing tiles (Supplementary Table 5 and Methods) and screened this RD mutant library using the pEF reporter and workflow described in Fig. 1a (Extended Data Fig. 7b–d).

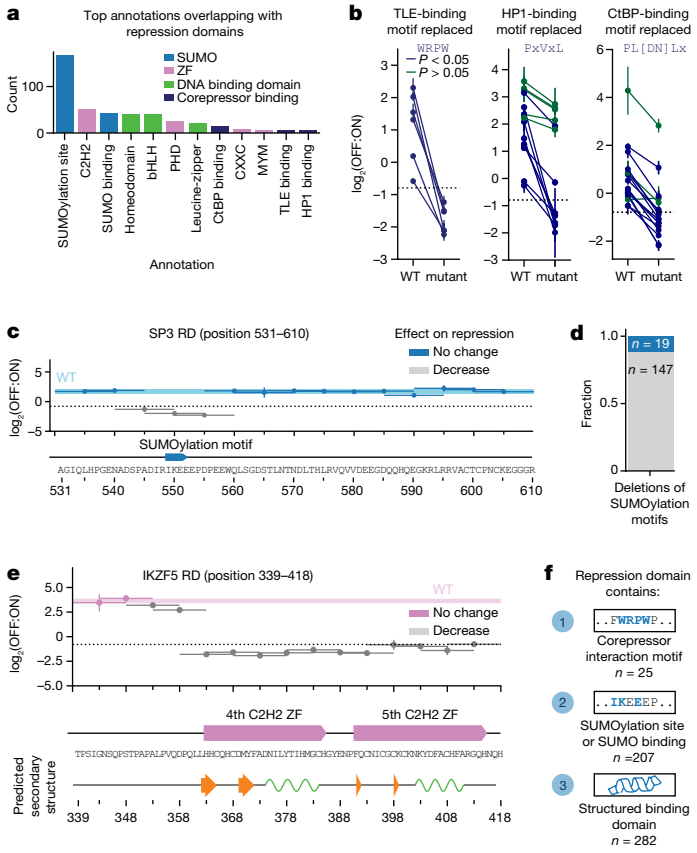


Fig. 3 | Most RD sequences contain sites for SUMOylation, short interaction motifs for recruiting corepressors or are structured binding domains for recruiting other repressive proteins. a, Count of RDs (repressive in both pEF and PGK promoter screens) that overlap annotations from UniProt and ELM[29]. Annotations that had at least six counts are shown. *P* values from a one-sided proportions *z* test stating how likely it is to find an annotation (for example, ZF) overlapping an activating tile versus a repressing tile: SUMO $P = 3.7 \times 10^{-26}$; ZF, 2.9×10^{-21}, DNA-binding domain, $P = 1.1 \times 10^{-22}$ and corepressor binding, $P = 4.7 \times 10^{-4}$. **b**, Repression enrichment scores ($n = 2$, dots are the mean, vertical bars the range) for tiles that contain a corepressor binding motif versus a replacement with Ala (Mutant). In TLE-binding: all 6 lost repressive activity on motif removal. Fraction of non-hit sequences containing motif was 0. HP1-binding: 8 out of 13 significantly decreased activity on motif removal (one-tailed *z* test). The fraction of non-hit sequences containing motif was 0.002. CtBP-binding: 14 out of 17 significantly decreased activity on motif removal. Fraction of non-hit sequences containing motif was 0.002. **c**, Deletion scan across SP3's RD. SUMOylation motif is 'IKEE'. The blue rectangle is the WT enrichment score, its height the range of two biologically independent screens. Each horizontal line represents those residues that were deleted, dots are the mean, vertical bars the range and $P < 0.05$ (one-sided *z* test compared to WT) are labelled in grey as a decrease. **d**, Fraction of deletion sequences containing a SUMOylation motif binned according to their effect on activity (blue means no change relative to WT, grey means decreased; one-tailed *z* test, $n = 166$ total RDs). **e**, Deletion scan across IKZF5's RD ($n = 2$, dots are the mean, bars the range). AlphaFold's predicted secondary structure (prediction from the whole protein sequence) shown below: alpha helices in green and beta sheets in orange. **f**, Summary of RD functional sequence categories (*n* indicated in the figure).

Futhermore, we stained for protein expression (Extended Data Fig. 7e,f) and filtered out mutants that had low FLAG enrichment scores.

We systematically replaced corepressor interaction motifs with alanine to test their contribution to activity (Fig. 3b). The TLE-binding motif, WRPW, appears exclusively in the C-terminal RDs of the HES family and all tiles containing this motif were repressive (Extended Data Fig. 7g). All tested TLE-binding motifs were necessary for repression (Fig. 3b, left). The HP1-binding motif, PxVxL, was necessary or contributed to repression in many of the tiles containing it (8 out of 13 tiles with decreasing effects Fig. 3b, middle). We found a more refined CtBP motif explained most tiles that lost activity on mutation (14 out of 17 tiles Fig. 3b, right, Extended Data Fig. 8a). Altogether, 78% of the 36 repressing tiles with a corepressor binding motif (TLE, HP1 or CtBP) decreased in repression strength when the motif was mutated, and 78% of 113 SUMO interaction motif- (SIM, the binding site to SUMOylated proteins) containing repressing tiles were similarly sensitive to mutation (Extended Data Fig. 8b).

We were intrigued by the many RDs that contain a SUMOylation site (site for covalent conjugation of a SUMO domain) (Fig. 3a). The ELM database classifies SUMOylation sites with the search pattern φKxE. Because this motif is short and flexible, some non-hit sequences (12.3%) also contain SUMOylation motifs. To investigate whether SUMOylation sites within non-hit sequences are functional, we used the AD deletion scan data. Deleting a SUMOylation motif within ADs rarely decreased activation (Extended Data Fig. 8c). Next, we asked if these motifs are functional in RDs using the same deletion scanning approach (Supplementary Table 5 and Fig. 3c). For example, residue K550 in the SP3 protein is a SUMOylation site and has been shown before to be important for repression[30]; indeed, we also find this site to overlap with the region necessary for repression (Fig. 3c). In a similar manner, we find SUMOylation motifs are important for the repression of at least 147 out of the 166 RDs where they are found (Fig. 3d and Supplementary Table 5). This result is concordant with our previous finding that a short 10 aa tile from the TF MGA, which contains this SUMOylation motif, IKEE, is itself sufficient to be a repressor[4]. SUMOylation of FOXP1 (which also shows up as a necessary region in our measurements, Supplementary Table 5), has been shown to promote repression by CtBP recruitment[31,32]. SUMOylation motif-containing TFs are enriched for binding corepressor KMT2D, as reported in a bioID interaction resource[33] ($P = 0.028$, one-sided proportions *z* test, compared to TFs with no EDs). We also identify a previously undescribed RD in KMT2D (Supplementary Fig. 2) containing a SIM, suggesting SUMOylation for these TFs drives repression by SIM-containing corepressor recruitment. Our results indicate the pervasive role, across more than 100 TFs, that SUMOylation plays in repression.

We next used our deletion scan data to gain better resolution of the region within RDs overlapping dimerization domains, such as basic helix-loop-helix domains (bHLHs). Within bHLHs, the basic region binds DNA and mutations in the HLH region are known to affect dimerization[34]. Deletion scans across tiles that overlap HLH domains reveal part of helix 1, the loop and helix 2 are necessary for repression (Extended Data Fig. 8d). HLHs lacking a basic region have previously been shown to negatively regulate transcription by forming complexes with other bHLHs and inhibiting their binding[35,36]. Alternatively, here we show that bHLHs containing basic regions can negatively regulate transcription when recruited at a promoter, probably by forming functional dimer complexes with another bHLH from a TF that contains RDs elsewhere in the protein. Most RDs that overlap bHLHs belong to Class II tissue specific bHLH TFs (Extended Data Fig. 8e) that can either activate or repress depending on the context[34]. Indeed, bHLH TFs can act as activators in other contexts: for example, NEUROG3, a Type II bHLH TF, acts as an activator when recruited full length to the minCMV promoter[5] and we find an activator tile that partially overlaps the bHLH RD (Supplementary Fig. 2). This context specificity to activation and repression of bHLH TFs might be expected given they can dimerize with different activating or repressing bHLH TFs.

Many RDs overlap annotated ZFs ($n = 124$), and some specifically overlap C2H2 ZFs ($n = 50$, compared to only three ADs that overlap C2H2 ZFs $P = 5.9 \times 10^{-24}$, one-sided proportions *z* test) (Fig. 3a). We wondered whether the C2H2 domain itself or the protein sequence flanking it was responsible for repression. For example, REST's ninth C2H2 ZF is repressive[37] and directly recruits the corepressor coREST[38]. In agreement with

these reports, our deletions in this RD of REST revealed the ninth ZF is necessary for repression (Extended Data Fig. 8f).

In addition to binding DNA and directly binding corepressors, ZFs dimerize with other ZFs[39]. We reasoned some ZFs could cause repression by binding to other ZF domains within endogenous repressive proteins. Support for this indirect recruitment of repressive TFs by means of ZFs comes from the IKZF family in which the N terminus of some members, such as IKZF1, directly recruits CtBP[40], whereas the C-terminal ZFs bind other IKZF family members[41]. Indeed, we recover the N-terminal repressive domains in IKZF1, and the associated sequence contains a CtBP-binding motif (Extended Data Fig. 8g). In addition, all IKZF family members show C-terminal RDs that overlap the last two ZFs (Extended Data Fig. 8g). These two ZFs are both necessary for repression in IKZF5 (Fig. 3e) and in all tested family members (Extended Data Fig. 8h), and therefore probably dimerize with the IKZFs that recruit CtBP. Whereas, in general, ZFs are well-known DNA-binding domains, our data expand the list of ZF sequences that are probably protein binding domains to other repressive TFs (Supplementary Table 5).

In summary, RDs can be categorized in the following way: (1) domains that contain short, linear motifs that directly recruit corepressors, (2) domains that contain SIMs or can be SUMOylated or (3) structured binding domains that probably recruit corepressors or other repressive TFs (Fig. 3f and Extended Data Fig. 8i).

Bifunctional activating and repressing domains

Transcriptional proteins are categorized as activating, repressing or bifunctional, where 115 proteins have previously been found to activate some promoters but repress others[8,42]. Here, we classify 248 proteins as bifunctional, CRs and TFs that have both an AD and RD (such as in Fig. 1b and Supplementary Table 2). Whereas most of these proteins contain both ADs and RDs at independent locations, a surprising fraction (92 out of 248) possess single domains apparently capable of both activation and repression (Fig. 4a–c and Supplementary Table 6), with many found within homeodomain TFs (Extended Data Fig. 9a).

To further investigate their behaviour, we individually recruited candidate bifunctional domains and quantified doxycycline-dependent minCMV activation and pEF repression (Fig. 4b and Supplementary Table 3). These validation measurements recapitulated initial screen observations, highlighting some domains with similar strengths of both repression and activation (for example, ARGFX-161:240 and NANOG-191:270), and others with preferential activities (for example, ARGFX-191:270, SREBF2-1:80; Fig. 4b and Extended Data Fig. 9b). Entire bifunctional domains could drive activation or repression, or specific regions within domains could mediate distinct activities. Systematic deletions of 10 aa segments within bifunctional domains further refined the necessary regions responsible for each activity (Supplementary Table 6 and Extended Data Fig. 9c–f). Whereas some bifunctional domains (23 of 92) possess independent activating and repressing regions (for example, NANOG; Extended Data Fig. 9g), others have fragments as small as 14 aa that are necessary for both activation and repression (69 of 92 domains, for example, ARGFX and the structurally related LEUTX) (Fig. 4d and Extended Data Fig. 10a–c).

Bifunctional domains could stably drive both activation and repression or could fluctuate between these activities over time. To distinguish between these possibilities, we quantified transcription driven by the bifunctional ARGFX tile 16 (Fig. 4b) at the minCMV promoter over 4 days and found that activation peaked at day 1 and then decreased over time (Extended Data Fig. 10d). Intrigued by these dynamics, we profiled activation dynamics for ARGFX tile 16 and several other bifunctional domains (FOXO1, NANOG and KLF7) recruited to a promoter of moderate strength (PGK) (Fig. 4e,f and Extended Data Fig. 10e). ARGFX tile 16 initially activated transcription at the PGK promoter from a low to a high state, but then the cell population split into two subpopulations: activated (high) or repressed (off). Other domains (for example, ARGFX tile 19 and FOXO1 tile 56) showed similar behaviour at the minCMV and PGK promoters, initially activating and then decreasing transcription over time. They also contained overlapping regions necessary for both activities (Supplementary Table 6). Several domains with bifunctional

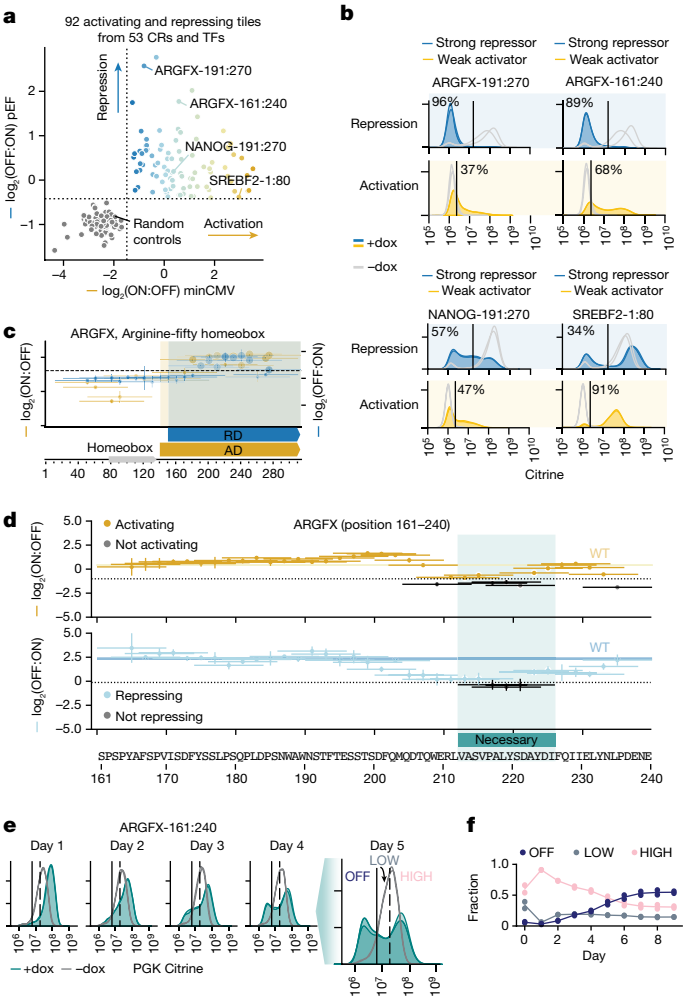


Fig. 4 | Discovery of bifunctional activating and repressing domains. **a**, Bifunctional tiles were discovered by observing both activation above the hits threshold (vertical dashed line) in the minCMV promoter CRTF validation screen (*x* axis) and repression above the hits threshold (horizontal dashed line) in the pEF promoter CRTF validation screen (*y* axis) ($n = 2$ biological replicates for each point). **b**, Citrine distributions from flow cytometry for individual validations of bifunctional tiles. Untreated cells (grey) and dox-treated cells (colours) ($n = 2$ biological replicates in each condition). Vertical line is the citrine gate used to determine the fraction of cells ON for activation and OFF for repression. **c**, Tiling plot for ARGFX ($n = 2$, dots are the mean, bars the range). Bifunctional domains are regions where the sequence is both activating at the minCMV promoter and repressing at the pEF promoter. **d**, Deletion scans across ARGFX-161:240 at minCMV promoter (top), and at pEF promoter (bottom). Yellow and blue rectangles represent WT enrichment scores, its height the range of two biologically independent screens. Each horizontal line represents those residues that were deleted, dots are the mean, vertical bars the range. The three deletions that caused no activation and no repression across both screens are highlighted in teal and the sequence annotated as necessary. **e**, Citrine distributions after recruitment of bifunctional tile ARGFX-161:240 to the PGK promoter ($n = 2$). Left vertical gate was used for measuring the fraction of cells OFF to its left. Right vertical gate was used for measuring the fraction of cells HIGH to its right. The fraction of LOW cells was measured by quantifying the number of cells between the two gates. **f**, Fraction of cells with citrine OFF (navy), LOW (grey) and HIGH (pink) over time after recruitment of ARGFX-161:240 ($n = 2$ biological replicates, average plotted as a line).

activities at the minCMV and pEF promoters did not significantly alter transcription when recruited to the PGK promoter, establishing that observed activities are promoter-dependent. For these domains, deletion scan measurements revealed independent regions necessary for activation and repression (Extended Data Fig. 9g and Supplementary Table 6). In summary, some bifunctional tiles that independently activate and repress different promoters are bifunctional even at a single promoter, and can dynamically split a cell population into high- and low-expressing cells.

Discussion

Compared to DNA-binding domains, many ED sequences are intrinsically disordered, poorly conserved and do not align well with one another in a multiple sequence alignment. When a new transcriptional protein is sequenced, homology models robustly identify DNA-binding domains and delineate their margins but cannot even identify EDs[43]. As a result, we lack comprehensive knowledge of the sequence patterns associated with EDs, and high-throughput experimental approaches are required to discover them and ultimately learn to predict their transcriptional activities.

Here, we report comprehensive measurements of human transcriptional EDs. By means of high-throughput tiling screens combined with deletion scans and rational mutagenesis, we collectively assigned EDs to 1,568 out of 2,047 (77%) of the CRs and TFs screened (Supplementary Table 1). Of these proteins, 1,193 were screened for activation as full-length proteins[5]. Despite the different methods, 83% had similar activities in both sets of measurements (Extended Data Fig. 10f). Of the ones that differed, 49 proteins only activated when full length, suggesting some proteins, such as PIN1, rely on a large catalytic fold that cannot be captured with 80 aa. Some proteins, such as CREB3L1, have large RDs (Supplementary Fig. 2) that might dominate over the ADs in the full-length protein. This might explain why 153 AD-containing proteins as determined by the CRTF tiling screen did not activate as full-length proteins[5]. These examples show how both methods measure complementary information.

Our sequential screening approach allowed us to first identify new EDs from a vast protein sequence space (>100,000 sequences), then follow up on these domains with orders of magnitude smaller validation screens (roughly 1,000–10,000 sequences) and nearly 100 individual validations, where we could not only confirm hits, but more accurately quantify each tile's transcriptional strength. Finally, by screening mutants of these high confidence activators and repressors, we identified sequence characteristics necessary for their functions.

In addition to the acidic exposure model[11], our data indicate further ways human ADs could conditionally promote hydrophobic exposure: serine could functionally mimic acidic residues only when phosphorylated and proline could favour exposure by breaking secondary structure. In this way, activation can be controlled by the relative activities of signalling proteins such as kinases, phosphatases and prolyl isomerases. Furthermore, ADs contain certain hydrophobic residues, but our data indicate those residues can be arranged in many ways, interspersed with serine, proline and/or acidic residues. Unlike RDs, we did not find any AD motifs, other than the previously reported LxxLL (Supplementary Table 4). Flexibility in composition might be related to promiscuity in binding, as many ADs bind many co-activators[44], probably because co-activators are a scarce resource within the cell[45]. Improving our understanding of ADs will require dissecting if and how their sequence composition dictates specific co-activator binding.

Here we propose several molecular mechanisms behind the function of RDs: corepressor binding motifs, SUMO interaction or SUMOylation sites and specific structured binding domains, accounting for 514 out of 715 of our RDs. This number is even larger (552 out of 715) if we consider the other TLE-binding motif EH1, the most abundant motif in a recent high-throughput study of *Drosophila* RDs[14]. This classification is in addition to the approximately 350 KRAB-structured domains that recruit KAP1, domains which we characterized in a previous high-throughput study[4] and excluded from the libraries used here. Previous investigation of several RDs revealed the presence of SUMOylation sites and established that SUMO-1 domain recruitment was sufficient for repression[30,32]. We find that SUMO-related repression is widespread: hundreds of RDs contain SUMOylation sites and deleting these sites ablates repression in roughly 90% of cases (Fig. 3d). SUMOylation could drive repression by recruiting SIM-containing corepressors (Extended Data Fig. 8i) or could localize SUMOylated TFs towards regions of heterochromatin[30,32].

ZF DNA-binding domains are the most common fold in the human proteome and can bind DNA, RNA or proteins[39]. Previous reports established that several ZF domain-containing proteins can interact with corepressors or repressive partner TFs[37–41], yet the relative prevalence of these interactions within CRs and TFs remained unknown. Here, we find 124 ZF domains that repress transcription, with the domain itself being necessary for repressive activity.

By systematically measuring both activation and repression of the same library, we were able to find EDs that can perform both roles. Whereas bifunctional TFs with separate domains have previously been observed[46–48], here we report bifunctional domains that are capable of simultaneously enhancing and silencing expression from a single promoter in different cells in the population. Bifunctional domains are particularly common in homeodomain TFs (Extended Data Fig. 9a), which are thought to compensate for a relative lack of DNA specificity by forming complexes with other DNA-binding partners and/or binding regulatory elements containing many motifs[49,50]. The bifunctional domain identified here within NANOG agrees with previous observations that this master regulator can both activate and repress at distinct target loci[51,52]. For homeodomain TFs such as CRX that switch from activating to repressing at regulatory elements containing several motifs, the downstream output could be dictated by particular stoichiometries or kinetics of partner protein binding[53,54].

Future work using these libraries and approaches in other cell types and under different signalling conditions will discover the context-specificities of this catalogue of EDs. The current work can be used to improve sequence prediction models of transcriptional EDs, understand the possible effects of CR and TF disease mutants, engineer better CRISPR systems[55], and move us one step closer to proteome-wide functional screening of protein domains.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05906-y.

1. Lambert, S. A. et al. The human transcription factors. *Cell* **175**, 598–599 (2018).
2. Medvedeva, Y. A. et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* **2015**, bav067 (2015).
3. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
4. Tycko, J. et al. High-throughput discovery and characterization of human transcriptional effectors. *Cell* **183**, 2020–2035.e16 (2020).
5. Alerasool, N., Leng, H., Lin, Z.-Y., Gingras, A.-C. & Taipale, M. Identification and functional characterization of transcriptional activators in human cells. *Mol. Cell* **82**, 677–695.e7 (2022).
6. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
7. Partridge, E. C. et al. Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583**, 720–728 (2020).
8. Soto, L. F. et al. Compendium of human transcription factor effector domains. *Mol. Cell* **82**, 514–526 (2022).
9. Keung, A. J., Bashor, C. J., Kiriakov, S., Collins, J. J. & Khalil, A. S. Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. *Cell* **158**, 110–120 (2014).

10. Erijman, A. et al. A high-throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning. *Mol. Cell* **78**, 890–902.e6 (2020).
11. Staller, M. V. et al. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.* **6**, 444–455.e6 (2018).
12. Sanborn, A. L. et al. Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *eLife* **10**, e68068 (2021).
13. Arnold, C. D. et al. A high-throughput method to identify trans-activation domains within transcription factor sequences. *EMBO J.* **37**, e98896 (2018).
14. Klaus, L. et al. Systematic identification and characterization of repressive domains in *Drosophila* transcription factors. *The EMBO Journal* **42.3**, e112100 (2023).
15. Stampfel, G. et al. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).
16. Neumayr, C. et al. Differential cofactor dependencies define distinct types of human enhancers. *Nature* **606**, 406–413 (2022).
17. Staller, M. V. et al. Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Syst.* **13**, 334–345.e5 (2022).
18. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
19. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional pulsing of a developmental gene. *Curr. Biol.* **16**, 1018–1025 (2006).
20. Bintu, L. et al. Dynamics of epigenetic regulation at the single-cell level. *Science* **351**, 720–724 (2016).
21. Ptashne, M. How eukaryotic transcriptional activators work. *Nature* **335**, 683–689 (1988).
22. Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).
23. Gerber, H. P. et al. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**, 808–811 (1994).
24. Gill, G., Pascal, E., Tseng, Z. H. & Tjian, R. A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the *Drosophila* TFIID complex and mediates transcriptional activation. *Proc. Natl Acad. Sci. USA* **91**, 192–196 (1994).
25. Courey, A. J. & Tjian, R. Analysis of Sp1 in vivo reveals mutiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* **55**, 887–898 (1988).
26. Escher, D., Bodmer-Glavas, M., Barberis, A. & Schaffner, W. Conservation of glutamine-rich transactivation function between yeast and humans. *Mol. Cell. Biol.* **20**, 2774–2782 (2000).
27. Tuttle, L. M. et al. Gcn4-mediator specificity is mediated by a large and dynamic fuzzy protein-protein complex. *Cell Rep.* **22**, 3251–3264 (2018).
28. Raj, N. & Attardi, L. D. The transactivation domains of the p53 protein. *Cold Spring Harb. Perspect. Med.* **7**, a026047 (2017).
29. Kumar, M. et al. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.* **50**, D497–D508 (2022).
30. Ross, S., Best, J. L., Zon, L. I. & Gill, G. SUMO-1 modification represses Sp3 transcriptional activation and modulates its subnuclear localization. *Mol. Cell* **10**, 831–842 (2002).
31. Rocca, D. L., Wilkinson, K. A. & Henley, J. M. SUMOylation of FOXP1 regulates transcriptional repression via CtBP1 to drive dendritic morphogenesis. *Sci Rep.* **7**, 877 (2017).
32. Verger, A., Perdomo, J. & Crossley, M. Modification with SUMO. A role in transcriptional regulation. *EMBO Rep.* **4**, 137–142 (2003).
33. Göös, H. et al. Human transcription factor protein interaction networks. *Nat. Commun.* **13**, 766 (2022).
34. Torres-Machorro, A. L. Homodimeric and heterodimeric interactions among vertebrate basic helix-loop-helix transcription factors. *Int. J. Mol. Sci.* **22**, 12855 (2021).
35. Sun, X. H., Copeland, N. G., Jenkins, N. A. & Baltimore, D. Id proteins Id1 and Id2 selectively inhibit DNA binding by one class of helix-loop-helix proteins. *Mol. Cell. Biol.* **11**, 5603–5611 (1991).
36. Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L. & Weintraub, H. The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. *Cell* **61**, 49–59 (1990).
37. Tapia-Ramírez, J., Eggen, B. J., Peral-Rubio, M. J., Toledo-Aral, J. J. & Mandel, G. A single zinc finger motif in the silencing factor REST represses the neural-specific type II sodium channel promoter. *Proc. Natl Acad. Sci. USA* **94**, 1177–1182 (1997).
38. Andrés, M. E. et al. CoREST: a functional corepressor required for regulation of neural-specific gene expression. *Proc. Natl Acad. Sci. USA* **96**, 9873–9878 (1999).
39. Brayer, K. J. & Segal, D. J. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem. Biophys.* **50**, 111–131 (2008).
40. Koipally, J. & Georgopoulos, K. A molecular dissection of the repression circuitry of Ikaros. *J. Biol. Chem.* **277**, 27697–27705 (2002).
41. McCarty, A. S., Kleiger, G., Eisenberg, D. & Smale, S. T. Selective dimerization of a C2H2 zinc finger subfamily. *Mol. Cell* **11**, 459–470 (2003).
42. Boyle, P. & Després, C. Dual-function transcription factors and their entourage: unique and unifying themes governing two pathogenesis-related genes. *Plant Signal. Behav.* **5**, 629–634 (2010).
43. Latchman, D. S. Eukaryotic transcription factors. *Biochem. J.* **270**, 281–289 (1990).
44. Dyson, H. J. & Wright, P. E. 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 (2016).
45. Gillespie, M. A. et al. Absolute quantification of transcription factors in human erythropoiesis using selected reaction monitoring mass spectrometry. *STAR Protocols* **1**, 100216 (2020).
46. Willy, P. J., Kobayashi, R. & Kadonaga, J. T. A basal transcription factor that activates or represses transcription. *Science* **290**, 982–985 (2000).
47. Majello, B., De Luca, P. & Lania, L. Sp3 is a bifunctional transcription regulator with modular independent activation and repression domains. *J. Biol. Chem.* **272**, 4021–4026 (1997).
48. Ma, J. Crossing the line between activation and repression. *Trends Genet.* **21**, 54–59 (2005).
49. Mann, R. S., Lelli, K. M. & Joshi, R. Hox specificity unique roles for cofactors and collaborators. *Curr. Top. Dev. Biol.* **88**, 63–101 (2009).
50. Bürglin, T. R. & Affolter, M. Homeodomain proteins: an update. *Chromosoma* **125**, 497–521 (2016).
51. Loh, Y.-H. et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).
52. Heurtier, V. et al. The molecular logic of Nanog-induced self-renewal in mouse embryonic stem cells. *Nat. Commun.* **10**, 1109 (2019).
53. White, M. A. et al. A simple grammar defines activating and repressing *cis*-regulatory elements in photoreceptors. *Cell Rep.* **17**, 1247–1254 (2016).
54. Friedman, R. Z. et al. Information content differentiates enhancers from silencers in mouse photoreceptors. *eLife* **10**, e67403 (2021).
55. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).

Methods

Cell culture

All experiments presented here were carried out in K562 cells (ATCC, CCL-243, female). Cells were cultured in a controlled humidified incubator at 37C and 5% CO_2, in RPMI 1640 (Gibco, 11-875-119) media supplemented with 10% fetal bovine serum (FBS) (Takara, 632180) and 1% penicillin streptomycin (Gibco, 15-140-122). HEK293T-LentiX (Takara Bio, 632180, female) cells, used to produce lentivirus, as described below, were grown in DMEM (Gibco, 10569069) media supplemented with 10% FBS (Takara, 632180) and 1% Penicillin Streptomycin Glutamine (Gibco, 10378016). minCMV and pEF reporter cell line generation is described in ref. 4. Briefly, pEF and minCMV promoter reporter cell lines were generated by TALEN-mediated homology-directed repair to integrate donor constructs (pEF promoter Addgene no. 161927, minCMV promoter Addgene no. 161928) into the *AAVS1* locus by electroporation of K562 cells with 1,000 ng of reporter donor plasmid and 500 ng of each TALEN-L (Addgene no. 35431) and TALEN-R (Addgene no. 35432) plasmid (targeting upstream and downstream the intended DNA cleavage site, respectively). After 7 days, the cells were treated with 1,000 ng ml^{-1} puromycin antibiotic for 5 days to select for a population where the donor was stably integrated in the intended locus. Fluorescent reporter expression was measured by microscopy and by flow cytometry. The PGK reporter cell line was generated by electroporation of K562 cells with 0.5 µg each of plasmids encoding the *AAVS1* TALENs and 1 µg of donor reporter plasmid using program T-016 on the Nucleofector 2b (Lonza, AAB-1001). Cells were treated with 0.5 µg ml^{-1} puromycin for 1 week to enrich for successful integrants. The PGK reporter donor plasmid generated in this study is available from Addgene (Addgene no. 196545). These cell lines were not authenticated. All cell lines tested negative for mycoplasma.

TF tiling library design

To construct the TF tiling library, 1,294 human TFs were selected from ref. 1. To make this library's size feasible for high-throughput measurements, we excluded 476 proteins that we have previously characterized with HT-recruit[4]: a set of 132 CRs and 344 KRAB-containing TFs. The canonical transcript of each gene was retrieved from Ensembl and chosen using the APPRIS (annotation of principal and alternative splice isoforms) principal transcript[56]. If no APPRIS tag was found, the transcript was chosen using the TSL principal transcript. If no TSL tag was found, the longest transcript with a protein coding CDS was retrieved. The coding sequences were divided into 80 aa tiles with a 10 aa sliding window. For each gene, a final tile was included spanning from 80 aa upstream of the last residue to that last residue, such that the C-terminal region would be included in the library. Duplicate sequences were removed, sequences were codon matched for human codon use, 7xC homopolymers were removed, BsmBI restriction sites were removed, rare codons (less than 10% frequency) were avoided, and the GC content was constrained to be between 20 and 75% in every 50 nucleotide window (performed with DNA chisel[57]). To improve the coverage of this large library, we subdivided it into three smaller sublibraries on the basis of the three main classes of TFs: a 25,032 C2H2 ZF sublibrary including 406 C2H2 ZF TFs, a 9,757 homeodomain and bHLH sublibrary including 304 homeodomain and bHLH TFs and a 31,664-member sublibrary containing the rest of the 583 TFs.

One thousand random controls of 80 aa lacking stop codons were computationally generated as controls using the DNA chisel package's random_dna_sequence function and included in each sublibrary. 473 sequences that were found to be non-activators and 42 sequences that were found to be activators in our laboratory's previous minCMV Nuclear Pfam screen[4] were included as negative and positive controls. We made use of alternative codon usage (match_codon_usage and use_best_codon functions) to recode the controls in each sublibrary to give ourselves the option of pooling the three sublibraries and running the library as one 73,288 element screen.

One hundred extra controls were added to each sublibrary to serve as fiduciary markers to aid comparing separately run screens. These controls were not recoded in each sublibrary and thus were repeated when pooling sublibraries.

Fifty activation domains from 45 proteins involved in transcriptional activation were curated from UniProt[3]. We queried the UniProt database for human proteins whose regions, motifs or annotations included the term 'transcriptional activation'. We then filtered for ADs that ranged in length from 30 to 95 aa. For ADs shorter than 95 aa, we extended the protein sequence equally on either side until it reached 95 aa. The protein sequences were reverse translated and further divided into 95 aa sequences with 15 aa deletions positioned with a 2-aa sliding window. Duplicate sequences were removed, sequences were codon matched for human codon usage, 7xC homopolymers were removed, BsmBI restriction sites were removed, rare codons (less than 10% frequency) were avoided and the GC content was constrained to be between 20 and 75% in every 50 nucleotide window, performed with DNA chisel[57]. Fifity yeast Gcn4 controls were added, which included previously studied deletions[27]. Then 2,024 library elements in total were added to the 31,664 element TF tiling sublibrary.

CR tiling library design

Candidate CR genes were initially chosen by including all members of the EpiFactors database, genes with gene name prefixes that matched any genes in the EpiFactors database and genes with any of the following gene ontology (GO) terms: GO:000785 (chromatin), GO:0035561 (regulation of chromatin binding), GO:0016569 (covalent chromatin modification), GO:1902275 (regulation of chromatin organization), GO:0003682 (chromatin binding), GO:0042393 (histone binding), GO:0016570 (histone modification) and GO:0006304 (DNA modification). Genes present in previous Silencer tiling screens[4] and genes present in the TF tiling screen were then filtered out. Biomart was used to identify and retrieve the canonical transcript, and chosen by (in order of priority) the APPRIS principal transcript, the TSL principal transcript or the longest transcript with a protein coding CDS. Tiles for each of these DNA sequences were generated using the same 80-aa tile/10-aa sliding window approach as the TF tiling library. Duplicate sequences were removed, DNA hairpins and 7xC homopolymers were removed, and sequences were codon matched for human codon usage with GC content being constrained to be between 20 and 75% globally and between 25 and 65% in any 50-bp window. To improve the coverage while performing the screen, this 51,297 element library was split into two sublibraries: a 38,241 element CR Tiling Main sublibrary and an 13,056 element CR Tiling Extended sublibrary. Computationally generated random negative controls, negative control tiles from the DMD protein screened in previous Nuclear Pfam screens[4] and fiduciary marker controls were added to each sublibrary: 1,700 elements to the Main sublibrary and 3,700 elements to the Extended sublibrary. These controls were not recoded, and thus were repeated when pooling sublibraries.

Library filtering

As we pooled the sublibraries and screened them as one large pool, several of the control sublibraries, that were not recoded, wound up being repeated in the pool several times. Sequences that were repeated upwards of five times had systematically lower enrichment scores than what was expected from previous screens, probably due to PCR bias. Therefore, we removed all repeated control elements and instead relied on individual validations to confirm our screens worked. Furthermore, there was a computational error in removing BsmBI sites from the CR tiling library, resulting in some sequences having accidental restriction cut sites in the middle of the open reading frame. We removed these sequences from further analysis and supplementary tables.

Activating hits validation library design

Here, 1,055 putative hit tiles were chosen by selecting all tiles where both biological replicates were recovered and had activation enrichment scores above 5.365 (determined by two standard deviations above the mean of poorly expressed random controls). We included 200 randomly selected random negative controls that were poorly expressed (expression threshold of −1.427) and 100 randomly selected non-hit tiles that had no activity in both the minCMV and the pEF CRTF tiling screens. There were 1,355 total library elements.

Repressing hits validation library design

Here, 9,438 putative hit tiles were chosen by selecting all tiles where both biological replicates were recovered and had pEF repression enrichment scores above 1.433 or had a PGK repression enrichment score above 0.880 (determined from three standard deviations above the mean of poorly expressed random controls). We included 500 randomly selected random negative controls that were poorly expressed (expression threshold of −1.427) and 100 randomly selected non-hit tiles that had no activity in the minCMV, pEF nor PGK CRTF tiling screens. There were 10,038 total library elements.

AD mutants library design

We defined compositional bias as any residue that represented more than 15% of the sequence (more than 12 residues). We took 424 compositionally biased tiles and replaced all residues with alanine. We took 1,055 aromatic or leucine-containing tiles and replaced all Ws, Fs, Ys and Ls with alanines. We took 1,052 acidic residue-containing tiles and replaced all Ds and Es with alanines. In 51 tiles that contained the 'LxxLL' motif (ELM accession ELME000045, regex pattern = [^P]L[^P][^P]LL[^P]), we replaced the motif with alanines. In 22 tiles that contained the 'WW' motif (ELM accession ELME000003, regex pattern = PP.Y), we replaced the motif with alanines. 8,205 deletions were designed by systematically removing 10 aa chunks, with a sliding window of 5 aa across 547 maximum activating tiles. All mutated sequences were reverse translated into DNA sequences using a probabilistic codon optimization algorithm, such that each DNA sequence contained some variation beyond the substituted residues, which improved the ability to unambiguously align sequencing reads to unique library members. The 1,055 putative hit tiles were included as positive controls (slightly more activating tiles than we report in the main text because these libraries were designed before we screened the validation library). We included 500 randomly selected random negative controls that were poorly expressed (expression threshold −1.427). There were 12,364 total library elements.

RD mutants library design

Twelve thousand deletions were designed by systematically removing 10 aa chunks, with a sliding window of 5 aa of the maximum tile across 800 putative RDs that were hits in both PGK and pEF CRTF tiling screens (slightly more RDs than we report in the main text because these libraries were designed before we screened the validation library). All mutated sequences were reverse translated into DNA using the method described above. The 1,593 putative hit tiles were included as positive controls. We took 644 compositionally biased tiles and replaced all residues with alanine. We replaced with alanines all the following motifs: 104 CtBP interaction motif-containing tiles (ELM accession: ELME0000098); 18 HP1 interaction motif-containing tiles (ELM accession: ELME000141); nine 'ARKS' motif-containing tiles (ELM accession DRAFT - LIG_CHROMO); 180 SIM-containing tiles (ELM accession: ELME000335) and seven WRPW motif-containing tiles (ELM accession ELME000104). We included 500 randomly selected random negative controls that were poorly expressed (expression threshold −1.427). There were 15,055 total library elements.

Bifunctional deletion scan library design

Three thousand, three hundred and one deletions were created by systematically removing 10 aa chunks, with a sliding window of 2 aa across 96 bifunctional activating and repressing tiles. All mutated sequences were reverse translated into DNA sequences using the method described above. We included the wild-type (WT) bifunctional tiles and 250 randomly selected random negative controls that were poorly expressed (expression threshold of −1.427). There were 3,674 total library elements.

Library cloning

Oligonucleotides with lengths up to 300 nucleotides were synthesized as pooled libraries (Twist Biosciences) and then PCR amplified. 6 × 50 µl reactions were set up in a clean PCR hood to avoid contamination with plasmid DNA from individual validations. For each reaction, we used either 5 or 10 ng of template, 1 µl of each 1 mM primer, 1 µl of Herculase II polymerase (Agilent), 1 µl of DMSO, 1 µl of 10 mM dNTPs and 10 µl of 5× Herculase buffer. The thermocycling protocol was 3 min at 98 °C, then cycles of 98 °C for 20 s, 61 °C for 20 s, 72 °C for 30 s and then a final step of 72 °C for 3 min. The default cycle number was 20×, and this was optimized for each library to find the lowest cycle that resulted in a clean visible product for gel extraction (23 cycles was the maximum when small libraries were represented in large pools). After PCR, the resulting double-stranded DNA libraries were gel extracted by loading a 2% TAE gel, excising the band at the expected length (around 300 bp), and using a Qiagen gel extraction kit. The libraries were cloned into a lentiviral recruitment vector pJT126 (Addgene no. 161926) with 4–16× 10 µl GoldenGate reactions (75 ng of predigested and gel-extracted backbone plasmid, 5 ng of library (2:1 molar ratio of insert:backbone), 2 µl of 10× T4 Ligase Buffer and 1 µl of NEB GoldenGate Assembly Kit (BsmBI-V2)) with 65 cycles of digestion at 42 °C and ligation at 16 °C for 5 min each, followed by a final 5 min digestion at 42 °C and then 20 min of heat inactivation at 70 °C. The reactions were then pooled and purified with MinElute columns (Qiagen), eluting in 6 µl of ddH_2O. Then 2 µl per tube was transformed into two tubes of 50 ml of Endura electrocompetent cells (Lucigen, catalogue no. 60242-2) following the manufacturer's instructions. After recovery, the cells were plated on 1–8 large 10 × 10 inch Luria-Bertani plates with carbenicillin. After overnight growth in a warm room (37 °C), the bacterial colonies were scraped into a collection bottle and plasmid pools were extracted with a Hi-Speed Plasmid Maxiprep kit (Qiagen). Two to three small plates were prepared in parallel with diluted transformed cells to count colonies and confirm the transformation efficiency was sufficient to maintain at least 20× library coverage. To determine the quality of the libraries, the putative EDs were amplified from the plasmid pool by PCR with primers with extensions that include Illumina adapters and sequenced. The PCR and sequencing protocol were the same as described below for sequencing from genomic DNA, except these PCRs use 10 ng of input DNA and 17 cycles. These sequencing datasets were analysed as described below to determine the uniformity of coverage and synthesis quality of the libraries. In addition, 20–30 colonies from the transformations were Sanger sequenced (Quintara) to estimate the cloning efficiency and the proportion of empty backbone plasmids in the pools.

Pooled delivery of libraries in human cells using lentivirus

Large scale lentivirus production and spinfection of K562 cells were performed as follows. To generate sufficient lentivirus to infect the libraries into K562 cells, we plated HEK293T cells on 1–12 15-cm tissue culture plates. On each plate, 8.8×10^6 HEK293T cells were plated in 30 ml of DMEM, grown overnight and then transfected with 8 µg of an equimolar mixture of the three third-generation packaging plasmids (pMD2.G, psPAX2, pMDLg/pRRE) and 8 µg of rTetR-domain library vectors using 50 ml of polyethylenimine (Polysciences no. 23966). pMD2.G (Addgene plasmid no. 12259; http://addgene.org/12259),

psPAX2 (Addgene plasmid no. 12260; http://addgene.org/12260) and pMDLg/pRRE (Addgene plasmid no. 12251; http://addgene.org/12251) were gifts from D. Trono. After 48 and 72 h of incubation, lentivirus was harvested. We filtered the pooled lentivirus through a 0.45-mm polyvinyl difluoride filter (Millipore) to remove any cellular debris. K562 reporter cells were infected with the lentiviral library by spinfection for 2 h, with two separate biological replicates infected. Infected cells grew for 2 days and then the cells were selected with blasticidin (10 mg ml^{-1}, Gibco). Infection and selection efficiency were monitored each day using flow cytometry to measure mCherry (Bio-rad ZE5). Cells were maintained in spinner flasks in log growth conditions each day by diluting cell concentrations back to a 5×10^5 cells per ml. Because lentiviral particles integrate randomly across accessible regions of the genome, we aimed for 600× infection coverage, and our lowest infection coverage was 130× (that is, 130 cells per library element during infection). We aimed to have 2,000–10,000× maintenance coverage (that is, 2,000–10,000 cells per library element postinfection). On day 8 postinfection, recruitment was induced by treating the cells with 1,000 ng ml^{-1} doxycycline (Fisher Scientific) for either 2 days for activation or 5 days for repression.

Magnetic separation

At each time point, cells were spun down at 300*g* for 5 min and media was aspirated. Cells were then resuspended in the same volume of phosphate buffered saline (PBS) (GIBCO) and the spin down and aspiration was repeated, to wash the cells and remove any IgG from serum. Dynabeads M-280 Protein G (ThermoFisher, 10003D) were resuspended by vortexing for 30 s. 50 ml of blocking buffer was prepared per 2×10^8 cells by adding 1 g of biotin-free bovine serum albumin (Sigma-Aldrich) and 200 ml of 0.5 M pH 8.0 EDTA into DPBS (GIBCO), vacuum filtering with a 0.22-mm filter (Millipore) and then kept on ice. For all activation screens, 30 µl of beads was prepared for every 1×10^7 cells, 60 µl of beads per 10 million cells for the pEF CRTF tiling, PGK CRTF tiling and minCMV bifunctional deletion scan screens, 120 µl of beads per 10 million cells for the pEF validation, 90 µl of beads per 10 million cells for the RD Mutants and pEF bifunctional deletion scan screens. Magnetic separation was performed as previously described in ref. 4.

FLAG staining for protein expression

The expression level measurements for the CRTF tiling library were made in K562 minCMV cells (with citrine OFF). 4×10^8 cells per biological replicate were used after 7 days of blasticidin selection (10 mg ml^{-1}, Gibco), which was 9 days postinfection. Citrine positive cells (K562 cells with the pEF-citrine reporter, no lentiviral infection with rTetR-FLAG constructs) were used to measure the background level of staining in cells known to lack the 3XFLAG tag (to help gate for sorting); 4×10^7 of these cells were spiked into each replicate. Fix Buffer I (BD Biosciences, BDB557870) was preheated to 37 °C for 15 min and Permeabilization Buffer III (BD Biosciences, BDB558050) and PBS (GIBCO) with 10% FBS (Omega) were chilled on ice. The library of cells expressing domains was collected and cell density was counted by flow cytometry (Bio-rad ZE5). To fix, cells were resuspended in a volume of Fix Buffer I (BD Biosciences, BDB557870) corresponding to pellet volume, with 20 ml per 1 million cells, at 37 °C for 10–15 min. Cells were washed with 1 ml of cold PBS containing 10% FBS, spun down at 500 × *g* for 5 min and then supernatant was aspirated. Cells were permeabilized for 30 min on ice using cold BD Permeabilization Buffer III (BD Biosciences, BDB558050), with 20 ml per 1 million cells, which was added slowly and mixed by vortexing. Cells were then washed twice in 1 ml of PBS+10% FBS, as before, and then supernatant was aspirated. Antibody staining was performed for 1 h at room temperature, protected from light, using 5 µl per 1×10^6 cells of a-FLAG-Alexa647 (RNDsystems, IC8529R). We then washed the cells and resuspended them at a concentration of 3×10^7 cells per ml in PBS + 10% FBS. Cells were sorted into two bins on the basis of the level of APC-A and mCherry fluorescence (Sony SH800S) after gating for viable cells (Supplementary Fig. 3). A small number of unstained control cells was also analysed on the sorter to confirm staining was above background. The spike-in citrine positive cells were used to measure the background level of staining in cells known to lack the 3XFLAG tag, and the gate for sorting was drawn above that level. After sorting, the cellular coverage was roughly 2,000×. The sorted cells were spun down at 500 × *g* for 5 min and then resuspended in PBS. Genomic DNA extraction was performed following the manufacturer's instructions (Qiagen Blood Midi kit was used for samples with fewer than 1×10^7 cells) with one modification: the Proteinase K+AL buffer incubation was performed overnight at 56 °C.

Library preparation and sequencing

Genomic DNA was extracted with the Qiagen Blood Maxi Kit following the manufacturer's instructions with up to 1×10^8 cells per column. DNA was eluted in elution buffer and not AE buffer to avoid subsequent PCR inhibition. The domain sequences were amplified by PCR with primers containing Illumina adapters as extensions. A test PCR was performed using 5 µg of genomic DNA in a 50 ml (half-size) reaction to verify if the PCR conditions would result in a visible band at the expected size for each sample. Then, 3–48× 100 µl reactions were set up on ice (in a clean PCR hood to avoid amplifying contaminating DNA), with the number of reactions depending on the amount of genomic DNA available in each experiment. Next, 10 µg of genomic DNA, 0.5 ml of each 100 mM primer and 50 ml of NEBnext Ultra 2× Master Mix (NEB) was used in each reaction. The thermocycling protocol was to preheat the thermocycler to 98C, then to add the samples for 3 min at 98 °C, then an optimized number of cycles of 98 °C for 10 s, 63 °C for 30 s, 72 °C for 30 s and then a final step of 72 °C for 2 min. All subsequent steps were performed outside the PCR hood. The PCR reactions were pooled and 145 µl were run on a 2% TAE gel, the library band around 395 bp was cut out and DNA was purified using the QIAquick Gel Extraction kit (Qiagen) with a 30 µl elution into non-stick tubes (Ambion). A confirmatory gel was run to verify that small products were removed. These libraries were then quantified with a Qubit HS kit (ThermoFisher) and sequenced on an Illumina HiSeq (2x150).

Computing enrichments and hits thresholds

Sequencing reads were demultiplexed using bcl2fastq (Illumina). A Bowtie reference (v.1.2.3) was generated using the designed library sequences with the script 'makeIndices.py' (HT-Recruit Analyse package) and reads were aligned with 0 mismatch allowance using the script 'makeCounts.py'. The enrichments for each domain between OFF and ON (or FLAGhigh and FLAGlow) samples were computed using the script 'makeRhos.py'. Domains with fewer than five reads in both samples for a given replicate were dropped from that replicate (assigned 0 counts), whereas domains with fewer than five reads in one sample would have those reads adjusted to five to avoid the inflation of enrichment values from low depth.

For all of the screens, domains with fewer than 20 counts in both conditions of a given replicate were filtered out of downstream analyses. Hit thresholds varied across screens, depending on coverage, separation purity and bio-replicate reproducibility, and were set based on: (1) the scores of negative controls, and (2) the validation curves relating screen scores to fractions of cells with the reporter ON or OFF as measured by flow cytometry for individual points. These validation curves are plotted for each screen (Fig. 1g,i for the CRTF tiling screens, Extended Data Fig. 3e,f for the hit validations screens and Extended Data Fig. 5f and 7d for the mutant screens). We chose the threshold to be 1–3 standard deviations away from the mean of poorly expressed random controls, with the exact number of standard deviations chosen to maximize the number of true positives and minimize the number of false positives across the validations. Noisier screens, with lower reproducibility, had higher hit thresholds to avoid false positives. For the expression screens, well-expressed tiles were those

with a $\log_2$(FLAGhigh:FLAGlow) one standard deviation above the median of the random controls. For the CRTF tiling repressor screens, hits were tiles with enrichment scores three standard deviations above the mean of the poorly expressed random controls. For the minCMV CRTF tiling, pEF bifunctional deletion scan and minCMV bifunctional deletion scan screens, hits were proteins with enrichment scores two standard deviations above the mean of the poorly expressed random controls. For the validation and mutant screens, hits were proteins with enrichment scores one standard deviation above the mean of the poorly expressed random controls.

Annotation of domains from tiles

Tiles must have been hits in both the CRTF tiling and validation screens to have been considered potential EDs. A domain started anywhere the previous tile was not a hit. If the previous tile was not a hit because it was not expressed, and if the antepenultimate (previous, previous) tile was a hit, then that tile was not considered the start and instead it was recovered into the middle of the domain. A domain ended anywhere the next successive tile was not a hit. If the next tile was not a hit because it was not expressed, and the following tile was a hit, then the tile that was not expressed was not considered the end. Domains started at the first residue of the first tile and extended until the last residue of the last tile within the domain. Single tiles that were hits in both the CRTF tiling and validation screens were considered EDs. For example, AKAP8's single activation tile (Supplementary Fig. 2), had activity when recruited individually (Supplementary Table 3), and its corresponding tile in the Mutant AD screen (Supplementary Table 4) contains deletions of unnecessary regions that maintained activation.

Individual recruitment assays and flow cytometry measurements

Protein fragments were cloned as a fusion with rTetR upstream of a T2A-mCherry-BSD marker, using GoldenGate cloning in the backbone pJT126 (Addgene no. 161926). K562 citrine reporter cells were then transduced with each lentiviral vector and, 3 days later, selected with blasticidin (10 mg ml^{-1}) until more than 80% of the cells were mCherry positive (6–9 days). Cells were split into separate wells of a 24-well plate and either treated with doxycycline (Fisher Scientific) or left untreated. Time points were measured by flow cytometry analysis of more than 10,000 cells (Bio-rad ZE5, Everest v.2.3-3.0). Doxycycline was assumed to be degraded each day, so fresh doxycycline media was added each day of the time course.

Flow cytometry analysis

Data were analysed using Cytoflow (v.1.1, https://github.com/bpteague/cytoflow) and custom Python scripts. Events were gated for viability and mCherry as a delivery marker. To compute a fraction of ON cells during doxycycline treatment, we fit a Gaussian model to the untreated rTetR-only negative control cells that fits the OFF peak and then set a threshold that was two standard deviations above the mean of the OFF peak to label cells that have activated as ON. We do the same for computing the fraction of OFF cells in repressor validations but fit a two component Gaussian and set a threshold that was two standard deviations below the mean of the ON peak. A logistic model, including a scale parameter, was fit to the validation and screen data using SciPy's curve fit function.

Western blots

Twenty million cells were pelleted and washed once with 5 ml of PBS. Pelleted cells were resuspended in 500 µl of ice cold lysis buffer (1× RIPA (EMD Millipore 20-188), 1% Triton X-100, 0.1% SDS, Roche cOmplete protease inhibitor cocktail mini tablet) and were put on a rotator at 4 °C for 30 min. Next, the lysates were sonicated with a COVARIS ultra-sonicator for 15 min (peak power of 140–175, duty factor of 10, cycles per burst 200). Lysates were spun down at 20,000 × *g* for 5 min. Protein amounts were quantified using the Qubit protein broad range assay kit (Thermo Scientific, no. A50668) and 30 µg were denatured in 1× laemmli sample buffer (Bio-rad no. 1610747) + 10% 2-mercaptoethanol for 10 min at 70 °C and subsequently loaded onto a gel and transferred to a polyvinyl difluoride membrane. Membrane was first blocked with 7% nonfat dry milk (Bio-rad no. 1706404) for 1 h at room temperature, then probed using FLAG M2 monoclonal antibody (1:1,000, mouse, Sigma-Aldrich, F1804) and Histone 3 antibody (1:2,000, rabbit, Abcam, AB1791) as primary antibodies overnight. Next, the membrane was washed with TBS-T 3×, 5 min each before being blotted again with goat anti-mouse IRDye 680 RD (1:20,000) and goat anti-rabbit IRDye 800CW (1:40,000, LICOR Biosciences, catalogue nos. 926-68070 and 926-32211, respectively) secondary antibodies for 1 h at room temperature. Blots were imaged on a Licor Odyssey CLx imager. Band intensities were quantified using ImageJ's gel analysis routine (see Supplementary Fig. 1 for regions of interest used).

Data analysis and statistics

All statistical analyses and graphical displays were performed in Python[58] (v.3.8.5). Enrichment scores shown in all figures (apart from replicate plots) are the average across two separately transduced biological replicates. The *P* values, statistical tests used and *n* are indicated in the figure legends.

Protein sequence analysis

Compositional bias was defined as an aa that appeared at least 12 times in 80 aa (that is, 15% of the sequence). In Fig. 2b, for each aa, a ratio was computed by counting the abundance of each aa in the tile and normalizing by the length and total number of sequences. Randomly sampled 10,000 non-hit 80 aa sequences were similarly calculated and the enrichment ratio was calculated by dividing the hits by non-hits. For the few activation tiles that contained glycine- and glutamine-rich sequences, there were fewer than five mutants that expressed well as measured by FLAG (Supplementary Table 4), so we excluded these from further statistical analyses.

Biological materials availability

Oligonucleotide libraries are available upon request.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The Illumina sequencing datasets generated in this study are available from the Sequencing Read Archive (BioProject PRJNA916593).

Code availability

The HT-recruit Analyze software for processing high-throughput recruitment assay and high-throughput protein expression assays are available on GitHub (https://github.com/bintulab/HT-recruit-Analyze). All custom codes used for data processing and computational analyses are available from the authors upon request.

56. Rodriguez, J. M. et al. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–117 (2013).
57. Zulkower, V. & Rosser, S. DNA Chisel, a versatile sequence optimizer. *Bioinformatics* **36**, 4508–4509 (2020).
58. Van Rossum, G. & Drake, F. L. Python 3 Reference Manual (CreateSpace, 2009).
59. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.* **112**, 16–21 (2017).
60. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

Acknowledgements We thank M. Hinks and members of our laboratories for helpful conversations and assistance. This work was supported by grant nos. NIH-NIGMS R35M128947 (L.B.), NIH-NHGRI R01HG011866 (L.B. and M.C.B.), NSF GRFP DGE-1656518 (N.D.), NIH-NIDDK F99/K00 F99DK126120 (J.T.), Stanford Bio-X Bowes Fellowship (P.S.), Stanford School of Medicine Dean's Fund (C.A.), NIH-NIGMS 5T32GM007365-45 (A.M.), Stanford Interdisciplinary Graduate Fellowship affiliated with Stanford Bio-X (A.M.), NIH Director's New Innovator Award 1DP2HD08406901 (M.C.B.), NSF CAREER 2142336 (P.F.) and the BWF-CASI Award (L.B.). P.F. is a Chan Zuckerberg Biohub Investigator.

Author contributions N.D. and L.B. designed the study, with significant intellectual contributions from P.S. and A.M. P.S. and N.D. designed the TF tiling libraries. A.M. designed the CR tiling libraries, with contributions from J.T., M.C.B. and L.B. N.D. designed all other libraries with contributions from J.T., A.M., P.S., M.C.B. and L.B. N.D. screened the CRTF minCMV and FLAG libraries with assistance from P.S. and J.T. A. and K.S. screened the CRTF pEF and PGK promoter libraries. N.D. performed all other screens. N.D. analysed the data, with assistance from L.B. I.L., C.A. and N.D. performed individual recruitment assay experiments. N.D. performed western blot experiments. C.L. generated the PGK cell line. N.D., P.F. and L.B. wrote the manuscript, with significant contributions from J.T. and C.L., along with contributions from all authors. L.B. supervised the project, with contributions from M.C.B. and P.F.

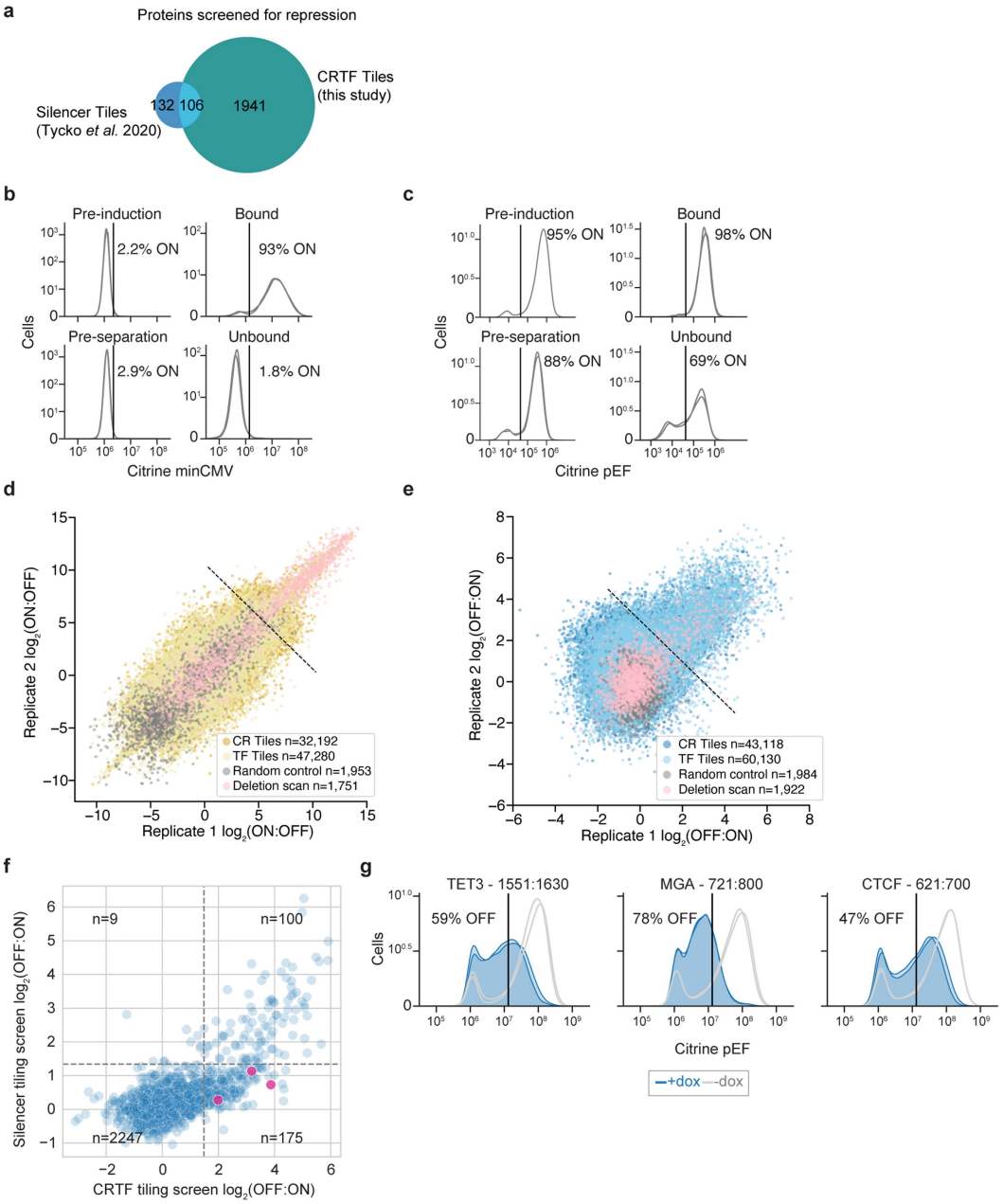
Competing interests N.D., A.M., P.S., J.T. and M.C.B. have filed a provisional patent (U.S. Provisional Application No. 63/318,144) related to this work. All remaining authors declare no competing interests.

Additional information

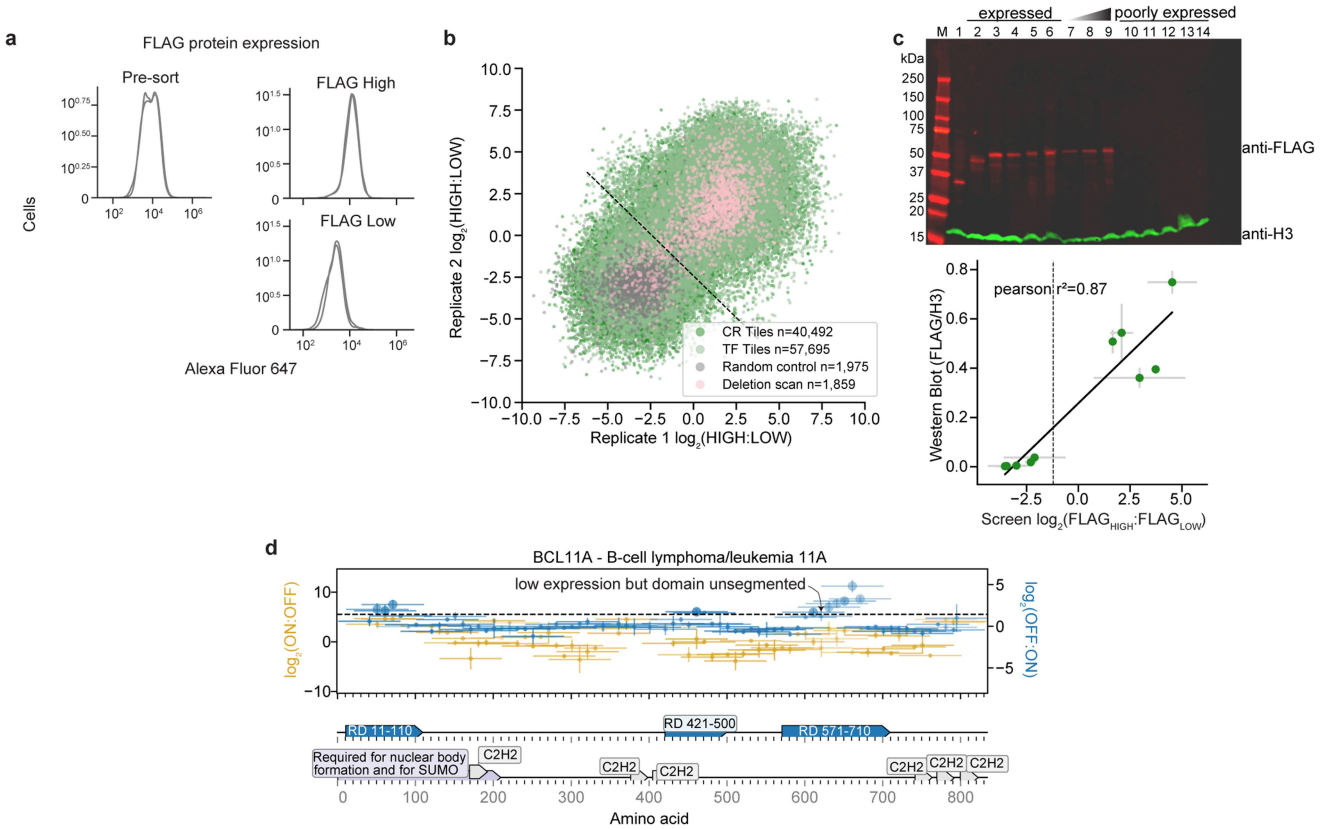
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05906-y.

Correspondence and requests for materials should be addressed to Lacramioara Bintu.

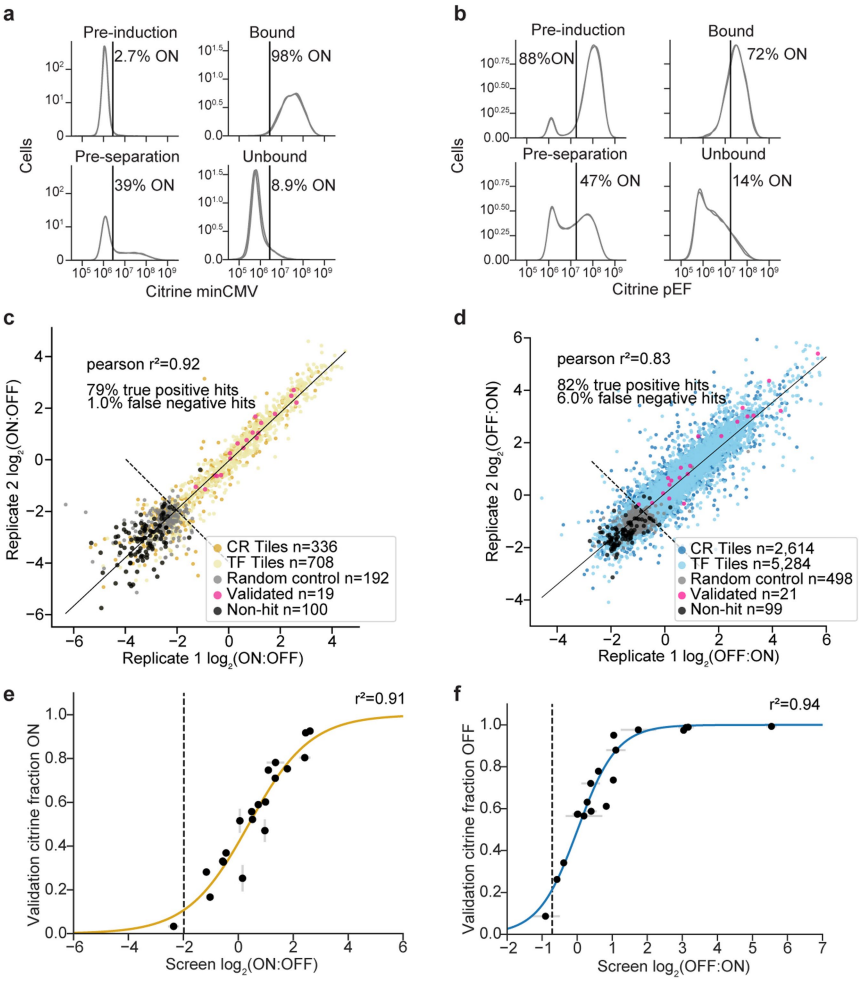
Peer review information *Nature* thanks Martha Bulyk, Steven Hahn and Matthew Welrauch for their contribution to the peer review of this work. Peer reviewer reports are available.



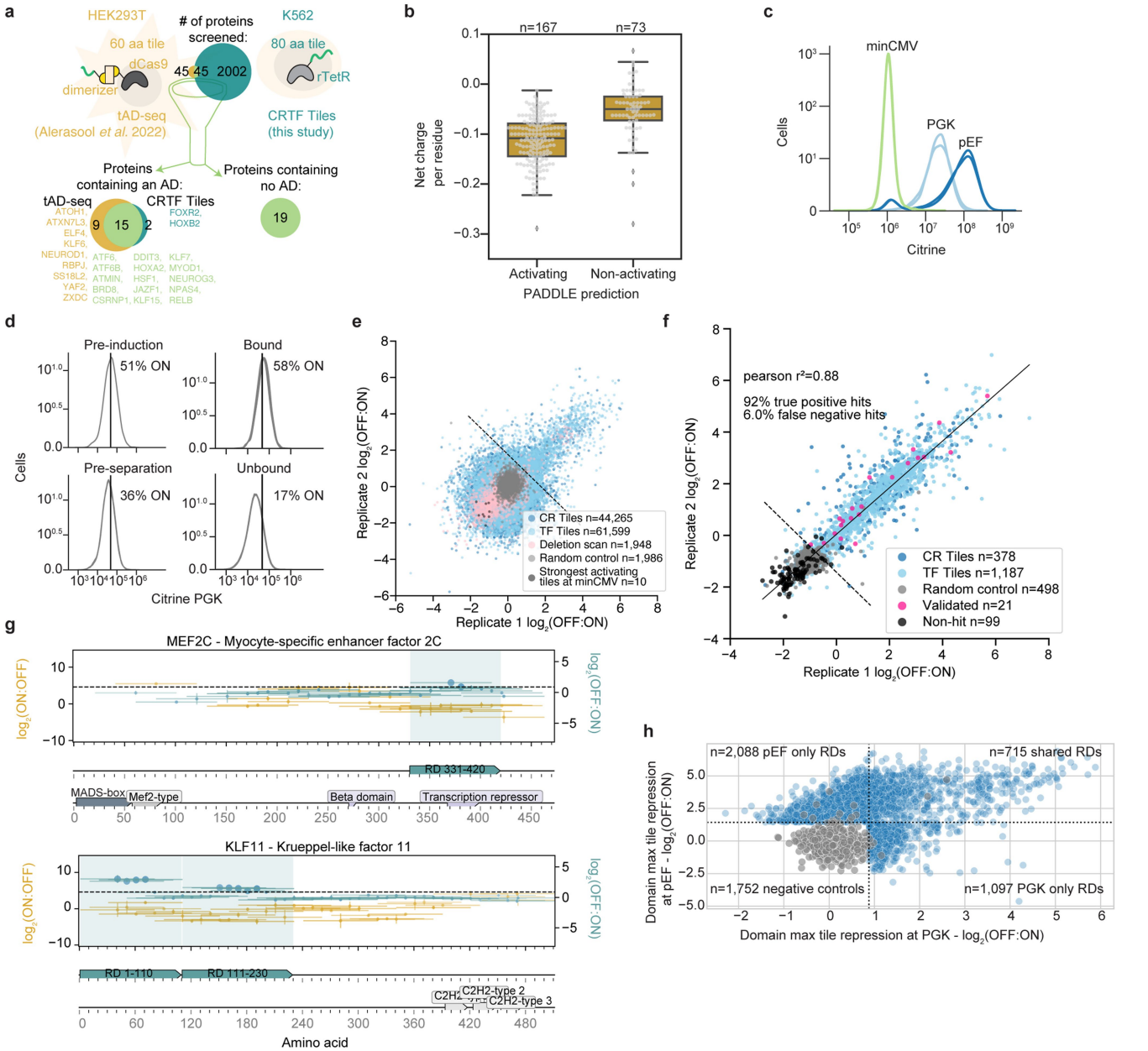
Extended Data Fig. 1 | CRTF tiling screens' separation purity, reproducibility, and validation. a, Comparison between the set of proteins tiled in Tycko et al., 2020 and this study. **b**, Flow cytometry data showing citrine reporter distributions for the minCMV promoter screen on the day we induced localization with dox (Pre-induction), on the day of magnetic separation (Pre-separation), and after separation (Bound). Overlapping histograms are shown for 2 separately transduced biological replicates. The average percentage of cells ON is shown to the right of the vertical line showing the citrine level gate. **c**, Citrine reporter distributions for the pEF promoter screen (n = 2). **d-e**, Biological replicate screen reproducibility (for hits above the threshold: pearson $r^2 = 0.78$ for minCMV and $r^2 = 0.19$ for pEF; for all data, including noise under the hit threshold: pearson $r^2 = 0.66$ for minCMV and $r^2 = 0.16$ for pEF). **f**, Comparison between average repression enrichment scores of tiles that were screened in the CRTF tiling pEF screen (x-axis) and previous Silencer tiling screen (y-axis)[4]. Dashed lines are the hits thresholds for each screen. Tiles were identical with a 1 aa register shift (as Silencer library tiles included an initial methionine absent from the CRTF tiling library). Pink dots are tiles that were individually validated in **g**. **g**, Citrine reporter distributions of individually validated CRTF tiling pEF screen hits that were not identified within the Silencer tiling screen (n = 2).



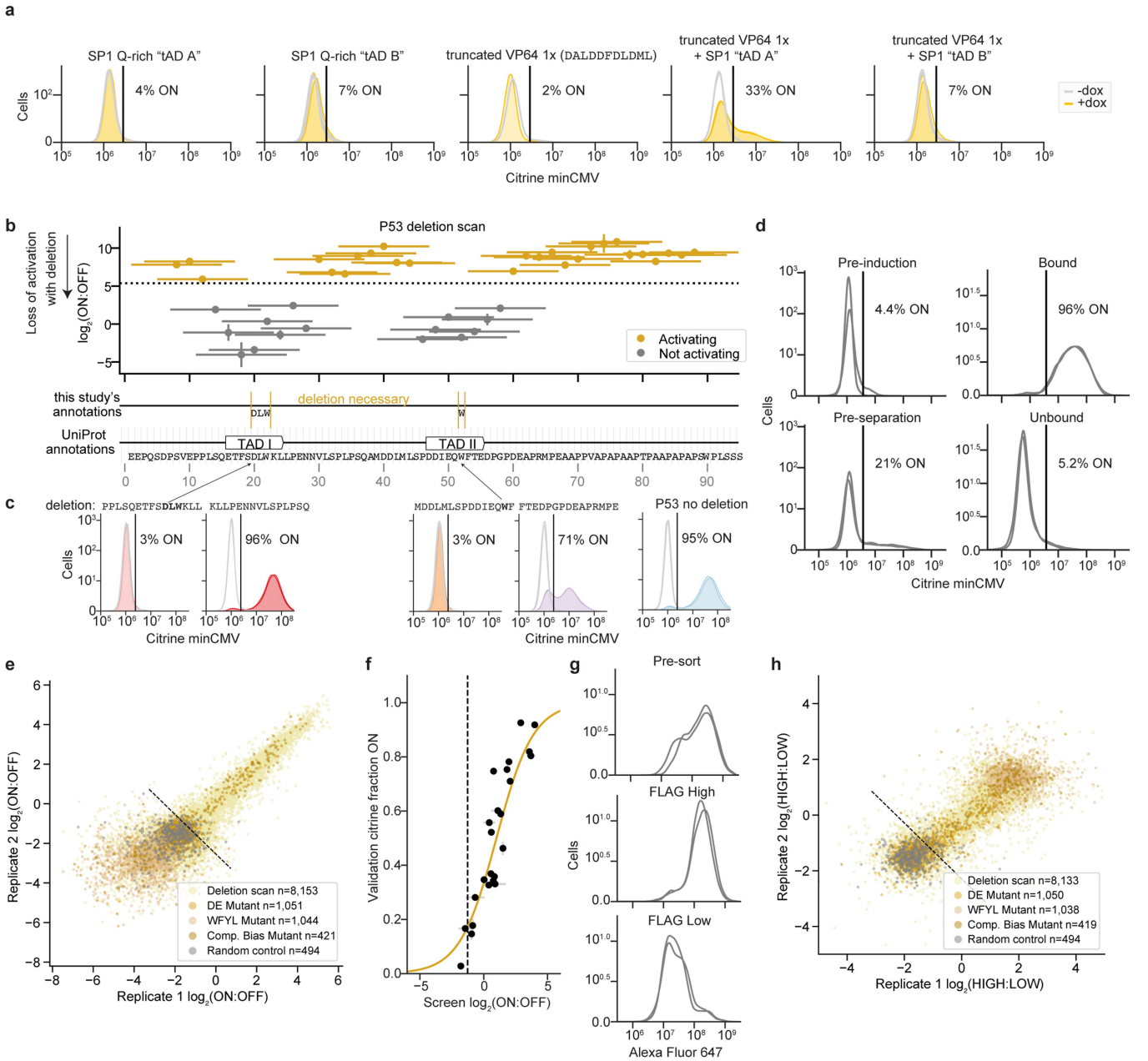
Extended Data Fig. 2 | CRTF tiling FLAG protein expression screen separation purity, reproducibility, validation, and example of how the data were used. a, Alexa Fluor 647 distributions from anti-FLAG staining of the CRTF tiling library in minCMV promoter reporter cells (n = 2). **b**, Biological replicate screen reproducibility (pearson $r^2 = 0.49$). **c**, Validations of FLAG protein expression screen. Expression levels were measured by Western blot with an anti-FLAG antibody. Anti-histone H3 was used as a loading control for normalization (see Supplementary Fig. 1 for regions of interest that were selected for quantification using ImageJ's gel analysis routine). Lane 1: rTetR-3xFLAG (no tile) theoretical molecular weight of 29 kDa; lanes 2-6: rTetR-3xFLAG-screened P53 deletions, theoretical molecular weight of 39 kDa; lanes 7-9: rTetR-3xFLAG-P53's AD loaded at increasing amounts; lanes 10-14: rTetR-3xFLAG-screened random control (see Supplementary Table 3 for protein sequences). Shift from expected molecular weight of the expressed P53 proteins is likely due to post-translational modifications P53's AD undergoes[28]. Comparison between high-throughput measurements of expression and Western blot protein levels ($r^2 = 0.87$, n = 10 proteins, n = 2 blot replicates, dots are the mean, bars the range). **d**, Tiling plot for BCL11A (n = 2, dots are the mean, bars the range). Example of a domain that was annotated at position 571-710. This domain had a low expression tile in the middle but the domain was left unsegmented. See more about how domains were called in Methods.



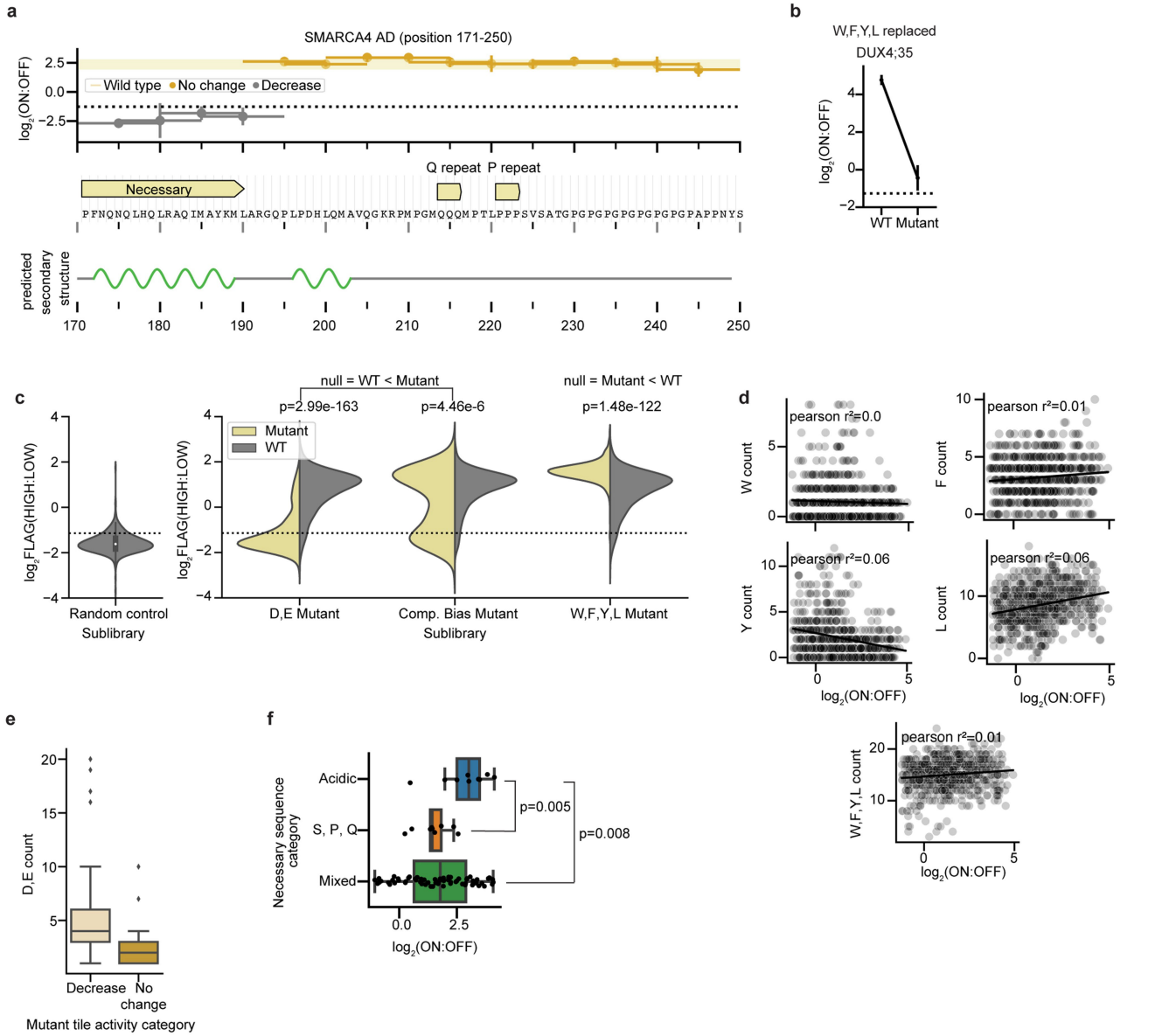
Extended Data Fig. 3 | CRTF tile hits validation screens' separation purity, reproducibility, and validation. a, Flow cytometry data showing citrine reporter distributions for the minCMV promoter screen on the day we induced localization with dox (Pre-induction), on the day of magnetic separation (Pre-separation), and after separation (Bound). Overlapping histograms are shown for 2 biological replicates. The average percentage of cells ON is shown to the right of the vertical line showing the citrine level gate. **b**, Citrine reporter distributions for the pEF promoter validation screen (n = 2). **c**-**d**, Biological replicate screen reproducibility. **e**, Comparison between individually recruited measurements and minCMV promoter validation screen measurements (n = 2, dots are the mean, bars the range) with logistic model fit plotted as solid line ($r^2 = 0.91$, n = 20). Dashed line is the hits threshold. Note, both screen thresholds are below 0, with several validated screen measurements below 0 (**Methods**). **f**, Comparison between individually recruited measurements and pEF promoter validation screen measurements (n = 2, dots are the mean, bars the range) with logistic model fit plotted as solid line ($r^2 = 0.94$, n = 19).



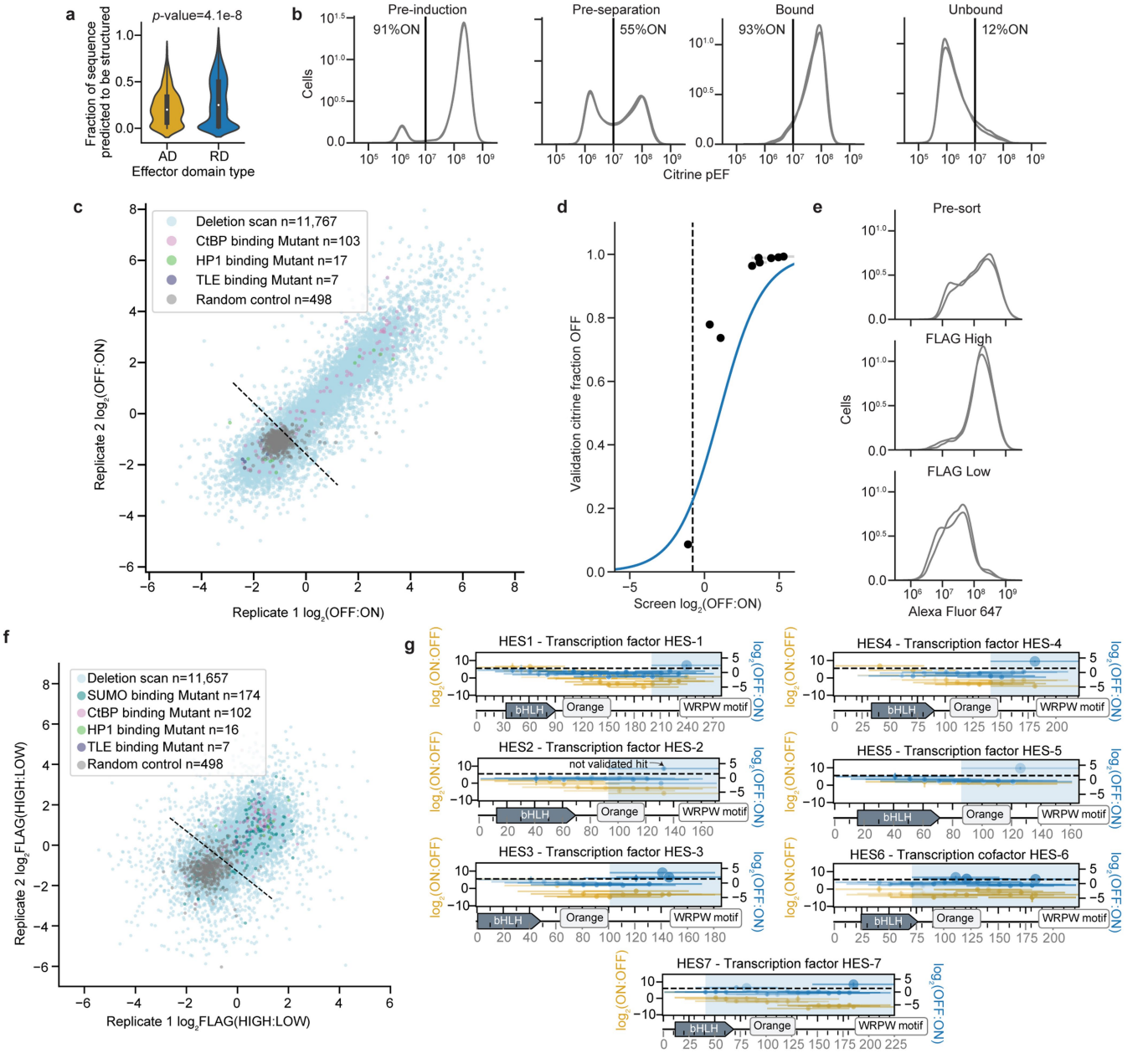
Extended Data Fig. 4 | Validations of CR & TF EDs. a, Comparison between set of proteins screened in Alerasool et al., 2022's tAD-seq and CRTF tiles (this study). **b**, Net charge per residue distributions (calculated by CIDER[59]) of activation domains identified by HT-recruit compared to their PADDLE-predicted function[12] (Mann-Whitney p-value = 1.4e-15, boxes: median and interquartile range (IQR); whiskers: Q1- 1.5*IQR and + Q3). **c**, CRTF tiling library screened at three different promoters with distinct expression levels. minCMV is a minimal promoter with all cells off. PGK is a low expression, medium strength promoter, and pEF is a high expression, strong promoter. **d**, Flow cytometry data showing citrine reporter distributions for the PGK promoter screen on the day we induced localization with dox (Pre-induction), 5 days later on the day of magnetic separation (Pre-separation), and after separation (Bound). Overlapping histograms are shown for 2 biological replicates. The average percentage of cells ON is shown to the right of the vertical line showing the citrine level gate. **e**, Biological replicate PGK promoter screen reproducibility (for hits above the threshold: pearson $r^2 = 0.27$ for repression hits; for all data, including noise under the hit threshold: pearson $r^2 = 0.11$ for all data). Although it is possible to detect activators at the PGK promoter, the dynamic range is very small (ten of the strongest activating tiles at the minCMV promoter (black dots) are very close to the random controls (grey dots)). **f**, Validation screen biological replicate reproducibility of tiles that were hits in both the PGK and pEF promoter screens. **g**, Tiling plots for MEF2C and KLF11 (n = 2, dots are the mean, bars the range). PGK repression domains annotated in teal. **h**, Comparison of each repression domain's max tile average repression scores in PGK (x-axis) and pEF promoter screen (y-axis). Dashed lines are the hits thresholds for each screen.



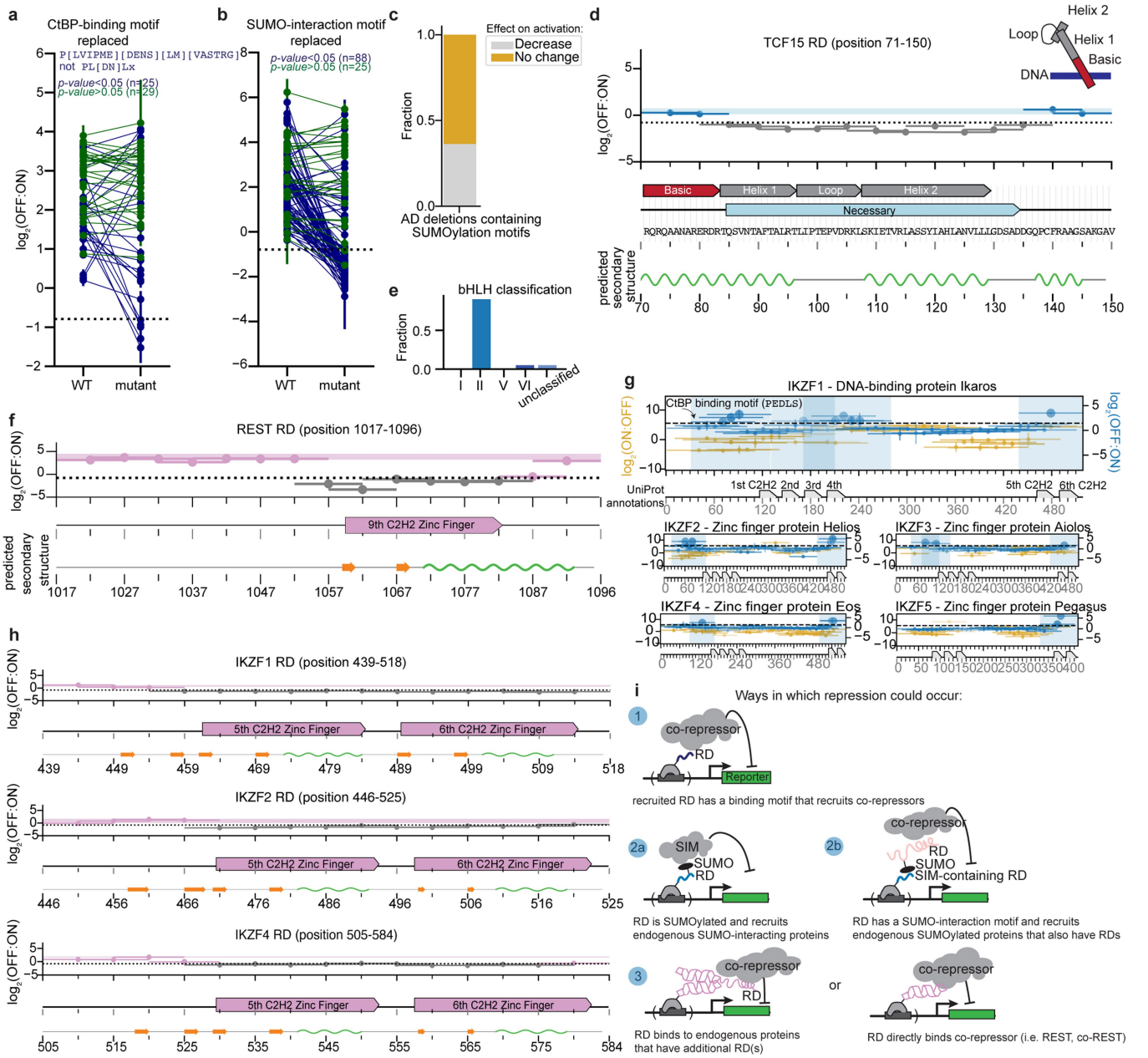
Extended Data Fig. 5 | Mutant AD screen's separation purity, reproducibility, and validation. a, Citrine distributions after 2 days recruitment to minCMV of UniProt-annotated Q-rich ADs with or without an 11 aa acidic sequence from VP64 (n = 2). **b**, Deletion scan across P53's AD: Deletions that caused a complete loss of activation, meaning they are below the experimentally validated activation threshold (dotted line, determined in Fig. 1g for the screen that included these constructs), are coloured in gray, and deletions that retained some activation are colored in yellow (n = 2, dots are the mean, bars the range). **c**, Individual validations of tiles including 15 aa deletions (deleted sequences shown above each panel). Untreated cells (gray) and dox-treated cells (colors) shown with two biological replicates in each condition. Vertical line is the citrine gate used to determine the fraction of cells ON (written above each distribution). **d**, Flow cytometry data showing citrine reporter distributions for the Mutant AD transcriptional activity screen on the day we induced localization with dox (Pre-induction), on the day of magnetic separation (Pre-separation), and after separation (Bound). Overlapping histograms are shown for 2 separately transduced biological replicates. The average percentage of cells ON is shown to the right of the vertical line showing the citrine level gate. **e**, Biological replicate Mutant AD transcriptional activity screen reproducibility. **f**, Comparison between individually recruited measurements and Mutant AD screen measurements (n = 2, dots are the mean, bars the range) with logistic model fit plotted as solid line ($r^2 = 0.95$, n = 23). **g**, Alexa Fluor 647 distributions from anti-FLAG staining. **h**, Biological replicate Mutant AD protein expression screen reproducibility.



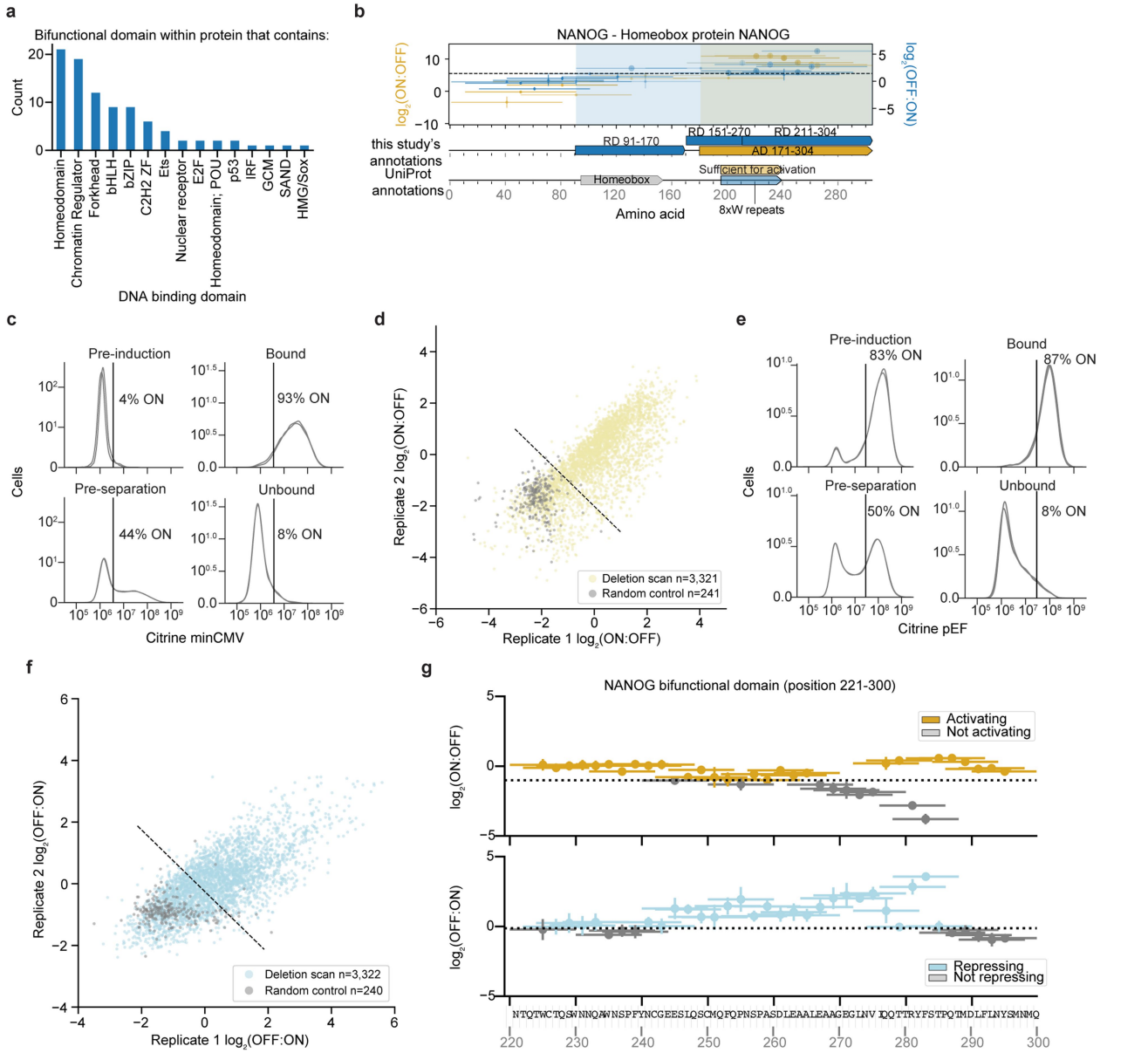
Extended Data Fig. 6 | Mutant AD screen follow-up. a, Deletion scan across SMARCA4's AD (n = 2, dots are the mean, bars the range). Predicted secondary structure (prediction from whole protein sequence using AlphaFold)[60] shown below, where green regions are alpha helices. Deletions that are significantly different from WT are colored in gray ($p < 0.05$, one-tailed z test). **b**, Enrichment scores comparing WT versus the W, F, Y, L mutant of DUX4 tile 35 (p-value = 3.3e-13, one-tailed z-test, n = 2, dots are the mean, bars the range). **c**, Violin plots of average FLAG enrichment scores from 2 biological replicates binned by each sublibrary. Dashed line represents the hit threshold for this screen. P-values computed from Mann-Whitney one-sided U tests. Boxes: median and interquartile range (IQR); whiskers: Q1- 1.5*IQR and + Q3. **d**, Correlations between each tile's activation strength in the minCMV validation screen and the count of indicated aa. **e**, Boxplot of acidic count for each mutant's activation category (Decrease n = 33, No change n = 18). Mann-Whitney one-sided U test, p-value = 2.25e-3. Boxes: median and interquartile range (IQR); whiskers: Q1- 1.5*IQR and + Q3. **f**, Boxplot of average activation enrichment scores with interquartile range shown for tiles that contain a single necessary sequence across each category (Acidic n = 9 S, P, Q n = 9, Mixed n = 64). P-values computed from Mann-Whitney one-sided U tests. Boxes: median and interquartile range (IQR); whiskers: Q1- 1.5*IQR and + Q3.



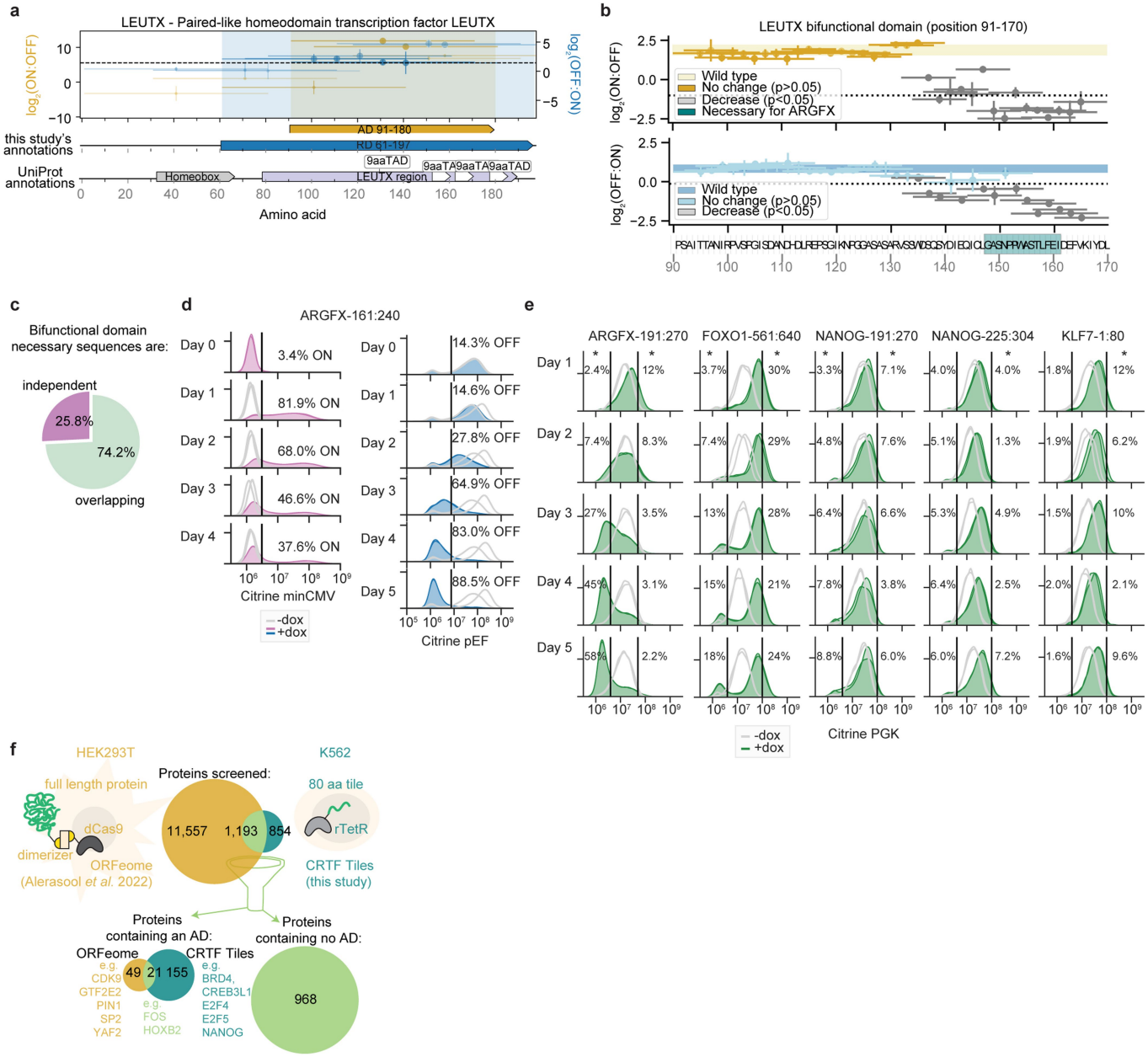
Extended Data Fig. 7 | Distribution of tile's predicted secondary structure, mutant RD screen's separation purity and reproducibility, and HES family tiling plot examples. a, Distributions of activating and repressing tile's fraction of the sequence predicted to be structured from AlphaFold's[60] predictions on the full length protein sequence. p-value = 4.1e-8 (Mann Whitney U test, one-sided, boxes: median and interquartile range (IQR); whiskers: Q1-1.5*IQR and + Q3). **b**, Flow cytometry data showing citrine reporter distributions for the Mutant RD transcriptional activity screen on the day we induced localization with dox (Pre-induction), on the day of magnetic separation (Pre-separation), and after separation (Bound). Overlapping histograms are shown for 2 separately transduced biological replicates. The average percentage of cells ON is shown to the right of the vertical line showing the citrine level gate. **c**, Biological replicate Mutant RD transcriptional activity screen reproducibility. **d**, Comparison between individually recruited measurements and Mutant RD screen measurements (n = 2, dots are the mean, bars the range) with logistic model fit plotted as solid line (r^2 = 0.91, n = 9). There are significantly fewer points for this plot compared to others because unlike the Mutant AD screen which included all hits that contained a W, F, Y or L, the Mutant RD screen had much fewer hits that overlapped our set of validations since only the strongest tiles within domains or hits that contained co-repressor binding motifs were included in the library design **e**, Alexa Fluor 647 staining distributions for the Mutant RD FLAG protein expression screen. **f**, Biological replicate Mutant RD protein expression screen reproducibility. **g**, Tiling plots for all 7 HES family members (n = 2, dots are the mean, bars the range).



Extended Data Fig. 8 | Mutant RD screen follow-up. a, Repression enrichment scores for a subset of repressing tiles (n indicated in figure) that contain a relatively more flexible CtBP-binding motif (regex shown above), excluding the more refined CtBP-binding motif (regex shown on second line). Mutants have their binding motifs replaced with alanines (p-values computed from one-tailed z-test). **b**, Repression enrichment scores for repressing tiles that contain a flexible SUMO-binding motif (fraction of non-hit sequences containing motif = 0.155). (n = 2, dots are the mean, bars the range, p-values computed from one-tailed z-test). **c**, Fraction of AD deletion sequences containing a SUMOylation motif binned according to their effect on activity (yellow=no change on activation relative to WT, gray=decreased activation). 11 total ADs. **d**, Deletion scan across TCF15's RD (n = 2, dots are the mean, bars the range). Deletions are colored by whether they were above (blue) or below (gray) the experimentally validated detection threshold for repression (dotted line). AlphaFold's[60] predicted secondary structure (prediction from whole protein sequence) shown below where green regions are alpha helices. Annotations shown from protein accession NP_004600.3 **e**, Distribution of bHLH classifications of RDs overlapping bHLH UniProt annotations. Classifications taken from ref. 34. **f**, Deletion scan across REST's RD (n = 2, dots are the mean, bars the range). Deletions are colored by whether they were above (pink) or below (gray) the validated threshold. AlphaFold's[60] predicted secondary structure (prediction from whole protein sequence) shown below where green regions are alpha helices and orange arrows are beta sheets. **g**, Tiling plots for IKZF family members (n = 2, dots are the mean, bars the range. **h**, Deletion scan across IKZF1, 2 and 4's RDs (n = 2, dots are the mean, bars the range). Deletions are colored by whether they were above (pink) or below (gray) the validated threshold. **i**, Cartoon model of potential mechanisms corresponding to the RD categories in Fig. 3f.



Extended Data Fig. 9 | Bifunctional domain deletion scan screen's separation purity, reproducibility, and examples. a, Counts of bifunctional domains from proteins that contain the indicated DNA binding domains. Homeodomains are enriched among TFs containing bifunctional domains compared to the frequency of homeodomains among all TFs ($p = 2.5e\text{-}4$, Fisher's exact test, two-sided). **b**, Tiling plot for NANOG (n = 2, dots are the mean, bars the range). **c**, Flow cytometry data showing citrine reporter distributions for the bifunctional deletion scan minCMV promoter screen on the day we induced localization with dox (Pre-induction), on the day of magnetic separation (Pre-separation), and after separation marker (Bound). Overlapping histograms are shown for 2 separately transduced biological replicates. The average percentage of cells ON is shown to the right of the vertical line showing the citrine level gate. **d**, Biological replicate bifunctional deletion scan minCMV promoter screen reproducibility. **e**, Citrine reporter distributions for the bifunctional deletion scan pEF promoter screen (n = 2). **f**, Biological replicate bifunctional deletion scan pEF promoter screen reproducibility. **g**, Example of a bifunctional domain from NANOG with independent activating and repressing regions (n = 2, dots are the mean, bars the range). Note, deletion of the necessary sequence for activation, caused an increase in repression, and vice-versa.



Extended Data Fig. 10 | Examples of bifunctional domain sequences at three different promoters. a, Tiling plot for LEUTX (n = 2, dots are the mean, bars the range). **b**, Deletion scan across one of LEUTX's bifunctional tiles (n = 2, dots are the mean, bars the range). Deletions were binned by their statistical significance into those that decreased activity (gray lines) compared to the WT tile and those that did not (one-tailed z-test). The necessary sequence for another gene family member, ARGFX, is highlighted in teal. **c**, Bifunctional domain necessary region location categories. Overlapping regions were defined as any tile that contained a deletion that was both necessary (below activity threshold) for activation and necessary for repression. **d**, Citrine distributions of ARGFX-161:240 recruited to minCMV (n = 2, left), and recruited to pEF (n = 2, right). **e**, Citrine distributions of bifunctional tiles identified from minCMV and pEF CRTF tiling screens recruited to PGK promoter (n = 2). Asterisks denote p-values < 0.05 for the percentage of cells on (right) and off (left) in the dox population (one-sided Welch's t-test, unequal variance). ARGFX-191:270 off p = 0.0003, on p = 0.02; FOXO1-561:640 off p = 0.017, on p = 2.44e-5; NANOG 191:270 off p = 2.12e-5, on p = 0.0002; NANOG 225:304 off p = 0.202, on p = 0.0004; KLF7 1:80 off p = 0.99, on p = 0.0005. **f**, Comparison between set of proteins screened in Alerasool et al., 2022's ORFeome and this study.

nature portfolio

Corresponding author(s): Lacra Bintu

Last updated by author(s): 01/28/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Flow cytometry data was collected using Everest version 2.3-3.0.
Data analysis	High-throughput recruitment assay and high-throughput protein expression assays were processed by the HT-recruit Analyze software (version 1). Bowtie version 1.2.3. All statistical analyses and graphical displays were performed in Python version 3.8.5 All flow cytometry data were analyzed using Cytoflow version 1.1 All DNA sequences were optimized for cloning/expression with DNA Chisel (version 3.2.2.) All western blot quantifications were done in ImageJ (version 1.53q)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The Illumina sequencing datasets generated in this study are available from the Sequencing Read Archive (SRA BioProject PRJNA916593 https://www.ncbi.nlm.nih.gov/sra/PRJNA916593).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No calculations were performed to predetermine sample size. Two biological replicates were performed for every screen and validation, which is standard in the field. Exact n is listed in each figure legend panel.
Data exclusions	No data were excluded
Replication	Every experiment had 2 independently transduced replicates. All attempts at replication were successful.
Randomization	Not relevant since data were quantitative and did not require subjective grouping.
Blinding	Not relevant since data were quantitative and did not require subjective grouping.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: FLAG M2 monoclonal antibody (1:1000, mouse, Sigma-Aldrich, F1804) and Histone 3 antibody (1:2000, mouse, Abcam, AB1791). Goat anti-mouse IRDye 680 RD (1:20,000) and goat anti-rabbit IRDye 800CW (1:40,000 dilution, LICOR Biosciences, cat nos. 926-68070 and 926-32211, respectively). FLAG-Alexa647 (5 uL / 10M cells, RNDsystems, IC8529R)

Validation: F1804: This is commercially available Western blot antibody against a common tag epitope and validated by the supplier. It has been referenced in 7215 publications. It has been referenced in Tycko et al., 2020
AB1791: This is commercially available Western blot antibody against a common epitope and validated by the supplier. It has been referenced in 3819 publications. It has been referenced in Tycko et al., 2020
IC8529R: This antibody has been optimized for Intracellular staining by flow cytometry. It has been referenced in Tycko et al., 2020

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s): K562 cells (ATCC, CCL-243, female)
HEK293T-LentiX (Takara Bio, 632180, female)

Authentication: These cell lines were not authenticated.

Mycoplasma contamination: All cell lines tested negative for mycoplasma.

Commonly misidentified lines (See ICLAC register): No commonly misidentified cell lines were used in the study.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation: After staining, we washed the cells and resuspended them at a concentration of 3 x 10^7 cells/ml in PBS+10%FBS.

Instrument: Sony SH800, and Biorad ZE5

Software: Sony SH800S software, Everest version 2.3-3.0.

Cell population abundance: FLAG expression screens: purity of each post-sort fractions were confirmed by flow cytometry.

Gating strategy: Cells were sorted into two bins based on the level of APC-A and mCherry fluorescence after gating for viable cells. A small number of unstained control cells was also analyzed on the sorter to confirm staining was above background. The spike-in citrine positive cells were used to assess the background level of staining in cells known to lack the 3XFLAG tag, and the gate for sorting was drawn above that level.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.