

1 Large-scale mapping and systematic mutagenesis of human transcriptional effector 2 domains

3
4 Nicole DelRosso¹, Josh Tycko², Peter Suzuki³, Cecelia Andrews⁴, Aradhana², Adi Mukund¹, Ivan
5 Liongson⁵, Connor Ludwig³, Kaitlyn Spees², Polly Fordyce^{2,3,6,7}, Michael C. Bassik², Lacramioara Bintu^{3,8}

6
7 ¹Biophysics Program, Stanford University, Stanford, CA 94305, USA

8 ²Department of Genetics, Stanford University, Stanford, CA 94305, USA

9 ³Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

10 ⁴Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA

11 ⁵Department of Biology, Stanford University, Stanford, CA 94305, USA

12 ⁶ChEM-H Institute, Stanford University, Stanford, CA 94305, USA

13 ⁷Chan Zuckerberg Biohub, San Francisco, CA 94110, USA

14 ⁸Correspondance (lbintu@stanford.edu)

15 16 Summary

17 Human gene expression is regulated by over two thousand transcription factors and chromatin
18 regulators^{1,2}. Effector domains within these proteins can activate or repress transcription. However, for
19 many of these regulators we do not know what type of transcriptional effector domains they contain, their
20 location in the protein, their activation and repression strengths, and the amino acids that are necessary
21 for their functions. Here, we systematically measure the transcriptional effector activity of >100,000
22 protein fragments (each 80 amino acids long) tiling across most chromatin regulators and transcription
23 factors in human cells (2,047 proteins). By testing the effect they have when recruited at reporter genes,
24 we annotate 307 new activation domains and 592 new repression domains, a ~5-fold increase over the
25 number of previously annotated effectors^{3,4}. Complementary rational mutagenesis and deletion scans
26 across all the effector domains reveal aromatic and/or leucine residues interspersed with acidic, proline,
27 serine, and/or glutamine residues are necessary for activation domain activity. Additionally, the majority
28 of repression domain sequences contain either sites for SUMOylation, short interaction motifs for
29 recruiting co-repressors, or are structured binding domains for recruiting other repressive proteins.
30 Surprisingly, we discover bifunctional domains that can both activate and repress and can dynamically
31 split a cell population into high- and low-expression subpopulations. Our systematic annotation and
32 characterization of transcriptional effector domains provides a rich resource for understanding the
33 function of human transcription factors and chromatin regulators, engineering compact tools for
34 controlling gene expression, and refining predictive computational models of effector domain function.

35 36 Introduction

37 Human gene expression is regulated by a constellation of over two thousand proteins known as
38 transcription factors (TFs) and chromatin regulators (CRs). TFs bind genomic DNA site-specifically¹, and
39 CRs recognize DNA and histone modifications⁵. Both classes of proteins contain effector domains that
40 recruit other macromolecules to activate or repress transcription. Consequently, mapping the location of
41 effector domains within these thousands of proteins is an essential resource for decoding the functional
42 properties of the human genome. Large scale efforts have mapped where in the genome TFs and CRs
43 bind^{6,7}. However, equivalent maps of human transcriptional effector domains are incomplete: we are
44 currently missing effector domain annotations for about 60% of the human TFs⁸.

45 Moreover, the sequence characteristics of what makes a good human activation or repression
46 domain are still under current investigation. Activation domains (ADs) are often disordered and are
47 typically characterized by their amino acid compositions^{9,10}. Most AD sequence characteristics have been
48 elucidated from yeast, where all ADs consist of a mix of acidic and hydrophobic residues^{11–13}. The acidic
49 residues are thought to keep the hydrophobic residues exposed to contact co-activators, an idea known
50 as the acidic exposure model^{14,15}. In addition to acidic activators, some human ADs have non-acidic
51 compositional biases, such as glutamine-, proline-, serine-, glycine-, and alanine-rich sequences^{16–19}. It
52 remains unclear how many non-acidic human ADs there are and how they work.

53 Repression domains (RDs), on the other hand, are less disordered⁸. As a result, RDs are not
54 typically described by their sequence compositions. Instead, a more useful description of RDs has been
55 categorization of their structural domains and repression mechanisms. For example, the family of
56 hundreds of KRAB domains generally recruit the scaffold protein KAP1, which represses transcription
57 and creates H3K9me3-associated heterochromatin by further recruitment of histone deacetylases,
58 histone methylase SETDB1, and heterochromatin protein HP1^{20,21}. Systematic categorization of human
59 RD sequences remains incomplete.

60 One useful assay for characterizing individual protein effector domains and mutants that test
61 specific sequence requirements consists of recruitment of domains to reporter genes (reviewed in⁸). This
62 approach has been extended from recruiting single domains to high-throughput assays in yeast^{11,12,14,22},
63 *Drosophila*^{23–25}, and human cells with a subset of transcriptional domains⁴ or a subset of full length
64 transcription factors²⁶. These works have extended our list of effector domains and have set the stage
65 for systematically mapping the effector domains across the thousands of human transcriptional proteins.

66 In order to map the human effector domains at unprecedented scale and resolution, here, we use
67 a high-throughput reporter assay to measure the transcriptional activity of 113,528 protein fragments
68 tiling across 2,047 chromatin regulators and transcription factors, the largest high-throughput assay for
69 protein function performed in human cells to date. Using rational mutagenesis and deletion scanning at
70 scale, we elucidate necessary sequence properties for both activation and repression domains. We find
71 that AD sequences that are glutamine-, proline-, and serine-rich behave similarly to acidic sequences:
72 the glutamines, prolines, and serines that are necessary for activity are the ones that are interspersed
73 with hydrophobic residues. Additionally, our data suggest the pervasive role that SUMOylation and zinc
74 finger domains play in causing repression for hundreds of RDs. Finally, we uncover a new set of
75 bifunctional domains, some of which are capable of simultaneously enhancing and silencing expression
76 from a single promoter.

77

78 **High-throughput mapping of effector domains**

79 To map the effector domain locations in human TFs and CRs, we selected 1,292 human TFs from the
80 Lambert 2018 dataset¹, 735 CRs from the EpiFactors Database², and added 20 genes with GO terms
81 matching chromatin and histone regulation (**Supplementary Table 1, Methods**). To make this library's
82 size feasible for high throughput measurements, we excluded 473 proteins that we have previously
83 characterized with HT-recruit⁴: a set of 129 CRs⁴ and 344 KRAB-containing TFs. For each TF and CR in
84 our list, we synthesized DNA sequences encoding 80 amino acid (aa) segments that tile across the full-
85 length protein (hereafter CRTF tiling library) with a 10 aa step size between segments (**Fig. 1a**). In
86 addition, we included 2,000 random 80 aa protein sequences as negative controls, 10 previously
87 validated effector domains⁴ as positive controls, and a deletion scan across 50 UniProt annotated ADs
88 as a pilot test of sequence perturbations (**Methods**).

89 This library, consisting of 128,565 sequences, was cloned into a lentiviral vector, where each
90 protein tile is expressed as a fusion protein with rTetR (a doxycycline inducible DNA binding domain),
91 and delivered to K562 cells containing a reporter with binding sites for rTetR⁴ (**Fig. 1a, Methods**). The
92 reporter gene is driven by either a minimally active minCMV promoter for identifying activators, or
93 constitutively active pEF promoter for finding repressors. To simultaneously measure the transcriptional
94 effector function of these sequences, we used a high-throughput recruitment assay we recently
95 developed: HT-recruit⁴. Briefly, the library was cloned and delivered as a pool at a low lentiviral infection
96 rate, such that each cell contains a single rTetR-tile. After treating the cells with doxycycline, which
97 recruits each CRTF tiling library member to the reporter, we magnetically separated the cells into ON and
98 OFF populations, extracted genomic DNA, and sequenced the tiles to identify sequences that were
99 enriched in the ON or OFF cell populations (**Extended Data Fig. 1a-b**). Activating tiles are enriched in
100 the ON population, while repressing tiles are enriched in the OFF population. Each screen was
101 reproducible across two biological replicates (**Extended Data Fig. 1c-d**). Using the random negative
102 controls, we drew thresholds for calling hits (**Extended Data Fig. 1c-d, Methods**). 90% and 92% of the
103 positive control domains for activation and repression⁴, respectively, were hits above this threshold, as
104 expected (**Supplementary Table 1**). We identified an additional subset of shared tiles (n=175) that were
105 only hits in this repression screen (**Extended Data Fig. 1e**) and whose activity validated in individual flow
106 cytometry experiments (**Extended Data Fig. 1f**). Overall, these results demonstrated HT-recruit reliably
107 identified transcriptional effectors while using an order-of-magnitude larger library than our previous
108 experiments⁴.

109 As measured transcriptional strength depends not only on the intrinsic potential of the sequence
110 but also on the levels at which individual tiles are expressed, we measured expression of all protein tiles.
111 All our library members contain a 3xFLAG tag, allowing us to measure each fusion protein's expression
112 levels by staining with an anti-FLAG antibody, FACS sorting the cells into FLAG HIGH and LOW
113 populations (**Extended Data Fig. 2a**), and measuring the abundance of each member in the two
114 populations by sequencing the domains (**Extended Data Fig. 2b**). Our FLAG screen scores correlate
115 well with individual validations (**Extended Data Fig. 2c**). These expression measurements were used
116 when annotating effector domains, for example allowing us to identify and filter out false negative library
117 members that have lower activation or repression scores due to low expression (**Extended Data Fig. 2d,**
118 **Methods**).

119 To further confirm all the hits and help remove false positives, we screened a smaller library
120 containing only the activating and repressive hit tiles (referred to as the validation screen,
121 **Supplementary Table 1, Methods**). Because of their small size (1,055 activating tiles and 7,939
122 repressive tiles), these screens had better magnetic separation purity (**Extended Data Fig. 3a-b**), and
123 the libraries could be screened at 10-fold higher coverage, which resulted in higher reproducibility than
124 the original, larger screens (**Extended Data Fig. 3c-d**), and even better correlation between screen
125 scores and individual validations (**Extended Data Fig. 3e-f**). Encouragingly, about 80% of the original
126 hits also were confirmed as hits in these validation screens (**Supplementary Table 1, Extended Data**
127 **Fig. 3c-d**). We only considered these confirmed sequences in subsequent analyses: 830 activation and
128 6,755 repression tiles (**Supplementary Table 1**).

129 Using these filtered tiling data, we annotated repression and activation domains from contiguous
130 hit tiles (**Extended Data Fig. 2d, Methods, Fig. 1b, Supplementary Table 2**). Doing so can accurately
131 identify effector domains previously annotated in UniProt, for example the activation and repression
132 domains in MYB (**Fig. 1b**). Some of the strongest ADs come from gene families with some family

133 members already annotated as activators, such as MYB, ATF, and NCOA, making us more confident our
134 screens returned reliable results. Similarly, some of the strongest RDs come from gene families with
135 some family members already annotated as repressors, such as MBD, KLF, and ZNF gene families (**Fig.**
136 **1c-d**). TFs from some gene families, like KLF, ETV, and KMT, contain both strong ADs and RDs, which
137 highlights our results can identify bifunctional transcriptional regulators. In total, 12% of the proteins
138 screened are bifunctional (having both ADs and RDs) and 76% of proteins have at least one effector
139 domain (**Supplementary Table 2**).

140 In addition, this method allows us to discover previously unannotated effector domains (**Fig. 1e**).
141 For example, we found both a new AD and four new RDs within the DNA demethylating protein, TET2.
142 We validated tens of these new effector domains by individually cloning them, creating stable cell lines,
143 and measuring their effect using flow cytometry after dox-induced recruitment at the minCMV reporter for
144 activation (**Fig. 1f, Supplementary Table 3**) or pEF reporter for repression (**Fig. 1h, Supplementary**
145 **Table 3**). Doing so, we validated the screen thresholds: all tiles above the thresholds had activity and no
146 tiles below the thresholds had activity (**Fig. 1g,i**). In total, 307 of the ADs and 592 of these RDs are new
147 compared to UniProt and previous HT-recruit screen⁴ annotations (**Fig. 1j, Extended Data Fig. 1g**).

148 Prior screens in yeast have led to the development of a machine learning model (PADDLE¹¹)
149 capable of predicting activation levels from sequence alone with an area under the precision-recall curve
150 of 81%. If the sequence properties that drive activation in humans are similar to those in yeast, we would
151 expect PADDLE to predict human ADs with similar accuracy. While PADDLE was able to predict many
152 ADs (70%), the domains that PADDLE predicted to be activating (like the C-terminal AD in CSRNP1)
153 were more negatively charged than the ADs it missed (**Extended Data Fig. 4a**), suggesting that in human
154 cells there are additional non-acidic activator classes compared to yeast.

155 Observed repression at the pEF promoter could reveal sequences that universally repress gene
156 expression; alternatively, this behavior may depend on the promoter to which tiles are recruited. To
157 distinguish between these scenarios and because there are no other comprehensive studies to reference
158 our results to, we decided to determine how many of these tiles are repressor hits at a different
159 constitutive promoter. We performed a new screen of the CRTF tiling library at the PGK promoter. While
160 this promoter is weaker, we were able to separate the silent and active cells by magnetic separation
161 (**Extended Data Fig. 4b**) and observed good reproducibility across two replicates (**Extended Data Fig.**
162 **4c**). 92% of the hit tiles that showed up in the pEF and PGK screens also showed up as hits in the pEF
163 validation screen (**Extended Data Fig. 4d**), suggesting higher confidence results when combining both
164 screens. We called RDs from contiguous hit tiles in the PGK data (**Extended Data Fig. 4e**). Across the
165 two repressive screens (at pEF and PGK), we found a total of 3,900 repressor domains, noting that some
166 of these domain boundaries are overlapping. Taking the maximum tile's enrichment scores within each
167 RD revealed 715 RDs were shared across both screens (**Extended Data Fig. 4f**). Together, these results
168 suggest that at the 80 aa scale there are more sequences across the CRs and TFs that can work as
169 repressors versus activators.

170 171 **Activation domain sequence characteristics**

172 ADs have been classified by the abundance of particular amino acids such as acidic (D, E), glutamine-
173 rich (Q), and proline-rich (P) sequences^{10,27}. Acidic residues have been shown to be essential for function
174 in all yeast activation domains¹¹ and some human ADs¹⁵. Certain human ADs have compositional biases
175 that are not present in other organisms, often containing stretches of single amino acid homotypic
176 repeats²⁸ (i.e. QQQQ). Additionally, some human ADs are enriched in particular hydrophobic residues -

177 aromatics (W, F, Y) and leucines (L), that are important for function in that context¹⁵. It remains unclear
178 how many human ADs fall under each of these categories - acidic, compositionally biased, or
179 hydrophobic, if these categories are indeed distinct from one another, and if these amino acids that are
180 enriched have functional significance for the ADs in those categories. Specifically, would activation be
181 lost if we deleted or mutated these amino acids?

182 Our new large set of activating tiles provides a great opportunity to systematically quantify the
183 prevalence of each of these sequence properties within human ADs. Every activating tile contained at
184 least one aromatic or leucine residue, and nearly every tile contained at least one acidic residue (**Fig.**
185 **2a**). Moreover, 45% of activating tiles contained a compositional bias (**Fig. 2a**), where serine (sequences
186 from NR4A and NFE2 families) and proline (sequences from FOX family and P53) were the most
187 abundant. Given that several ADs have been categorized by their glutamine content, we were surprised
188 to find very few glutamine-rich ADs across all the CRs and TFs (sequences from SMARCA family and
189 TRERF1, **Supplementary Table 4**). Consistent with these observations, when we normalize the amino
190 acid frequencies in the AD sequences by the amino acid counts in the non-hit sequences, we observe
191 an enrichment in certain hydrophobic, acidic, serine, and proline residues (**Fig. 2b**).

192 To determine which amino acid types among these enrichments are necessary for activation and
193 to find the necessary sequences within each activating tile, we took a deletion scanning approach, which
194 others have used to identify necessary sequences in yeast ADs²⁹. First, by performing scanning deletions
195 (15 aa each) across 24 UniProt annotated ADs that had activity at the minCMV promoter in K562, we
196 found that while most (61%) deletions do not affect activation, in the majority of these ADs (20/24) we
197 found at least one deletion that was well-expressed and could abolish activator function (**Supplementary**
198 **Table 1**). In order to validate that this deletion scanning approach returned residues that were necessary
199 for activity, we compared our deletion scan data from P53 to UniProt annotations and found the minimized
200 necessary sequences to be residues 20-22 (DLW) within one region and residue W52 within another
201 region, corresponding to UniProt-annotated TAD I and TAD II, respectively (**Extended Data Fig. 5a**).
202 Furthermore, individual validations confirmed the complete loss of activity when deletions including these
203 residues were tested (**Extended Data Fig. 5b**).

204 Confident in our deletion scan approach, we designed a second library of 10 aa deletions across
205 the maximum activating tile from each AD, resulting in 304 total deletion scans (**Supplementary Table**
206 **4**). We measured activation scores for all 12,320 members of this library using the minCMV reporter and
207 HT-recruit workflow described in **Fig. 1a** (**Extended Data Fig. 5c-d**). We FLAG-stained for protein
208 expression (**Extended Data Fig. 5e-f**) and filtered out mutants that were poorly expressed. Across each
209 of these expression-filtered deletion scans we first binned deletions into those that had an effect on
210 activation and those that did not (**Fig. 2c**). Using these data, we can identify which of the amino acids
211 that contributed to the compositional bias are important for function: for example, while NFAT5's AD has
212 a patch of 4 serines near the C-terminus, deleting those residues had no effect on activation (**Fig. 2c**).
213 We highlight similar examples for stretches of prolines and glutamines that are not essential for activation
214 (**Extended Data Fig. 6a**). Applying this analysis to all ADs containing a homotypic repeat, and after
215 removing all poorly-expressed deletions, we find homotypic repeats of certain hydrophobic residues like
216 glycine, alanine and leucine were equally found in deletions that had no effect on activation and in
217 deletions that decreased activation (**Fig. 2d**). However, serine, proline, acidic and glutamine homotypic
218 repeats were more often found in deletions that had no effect on activation than in deletions that
219 decreased activation (**Fig. 2d**). Therefore, homotypic repeats of these amino acids are generally not
220 necessary for activation.

221 The deletion scans also allow us to identify the necessary sequence for activation of each tile:
222 sequences that, once removed, completely abolished activation (**Fig. 2c**). We were able to annotate at
223 least one necessary sequence (median length=10 aa) in the majority (69%) of our screened ADs, and
224 most (61%) ADs have multiple necessary sequences, supporting the idea that ADs are composed of
225 multiple small linear binding motifs (**Fig. 2c, Supplementary Table 4**). Nearly every necessary sequence
226 (96%) contained a W, F, Y or L.

227 In order to validate this enrichment of specific hydrophobic residues, we rationally designed
228 mutant libraries where we systematically replaced every amino acid of a particular type within the
229 sequence with alanines (**Supplementary Table 4**). Replacement of all W, F, Y or Ls with alanine (range:
230 3-24 aa replaced/80 aa tile, median=10 aa) in all our activating tiles resulted in a total loss of activation
231 (**Fig. 2e**). The one exception that remained active was within DUX4, and the mutation did in fact make it
232 weaker (**Extended Data Fig. 6b**). This systematic loss of activation was not due to a decrease in protein
233 expression, as measured by FLAG staining (**Extended Data Fig. 6c**). This means all 732 tested tiles
234 from 258 proteins with ADs require some aromatic or leucine residues to activate.

235 We next wanted to follow up more on the acidic sequences, so we replaced all acidic residues
236 with alanine in the entire set of activating tiles (not just the few that had a compositional acidic bias).
237 Surprisingly, more than half of the acidic mutants had reduced expression (**Extended Data Fig. 6c**).
238 These results suggest acidic residues increase protein levels, at least in the context of transcriptional
239 activators. Of the remaining 247 well-expressed activating tile mutants, the majority of mutants lost the
240 ability to activate (**Fig. 2f**, n=196). 33 mutants decreased their activities upon mutation, and only 18
241 mutants had no change in activation, where some in fact increased (**Supplementary Table 4**). The
242 activator tiles that depended on acidic residues came from a wide range of TF families, including E2Fs
243 and GRHLs and the classical example acidic activator ATF4 (GCN4's mammalian homolog). Some of
244 the sequences that do not require acidic residues came from SMARCA5, TET2, PLAG1s, and every
245 paralog from the EYA family that had an AD. These mutants with no change in activity had significantly
246 fewer acidic residues than the tiles whose mutants had a decreasing effect (**Extended Data Fig. 6d**),
247 supporting the idea that acidic ADs are not the only class of human ADs.

248 Intrigued by what other compositional biases could be functional in human ADs, we next tested
249 the necessity of other frequently-appearing amino acids. We replaced compositionally biased amino
250 acids with alanine. For the few activation tiles that contained glycine-rich and glutamine-rich sequences,
251 there were fewer than 5 mutants that expressed well as measured by FLAG (**Supplementary Table 4**),
252 so we excluded these from further statistical analyses. Consistent with the results above, all tiles with
253 leucine compositional biases lost activity once mutated, and the few tiles with acidic biases lost activity
254 once mutated (**Fig. 2g**). Removal of serine and proline compositional biases had more mild effects: the
255 vast majority of mutants still had activity (**Fig. 2g**, top), even though the strength of activation decreased
256 for a subset of them (**Fig. 2g**, bottom).

257 Wanting to follow up more on the compositionally biased tiles that decreased activity upon
258 compositional bias removal (**Fig. 2g**), we next wondered if it was the homotypic repeats themselves that
259 explained this loss in activity or if a subset of compositionally biased residues overlapped important co-
260 activator binding motifs. To answer whether the placement of serines, prolines, and acidic residues within
261 the sequence were more important than their overall abundance, we analyzed the set of deletion
262 necessary sequences from the compositionally biased activating tiles that lost activity upon bias removal
263 (**Fig. 2g**, bottom). For each compositional bias type, the majority of necessary sequences also contain a
264 W, F, Y, or L (**Fig. 2h**).

265 In summary, sequences that are necessary for activation consist of certain hydrophobic amino
266 acids (W, F, Y, and/or L) that are interspersed with either acidic, proline, serine, and/or glutamine residues
267 (Fig. 2i, Extended Data Fig. 6e).

268

269 Repression domain sequence characteristics

270 Repressing tile sequences have significantly more secondary structure than activating tile's (Extended
271 Data Fig. 7a). Therefore, we needed to take a different approach for understanding the sequence
272 characteristics of RDs. Instead of looking at RD sequence compositions, we first set out to classify the
273 RDs by their potential mechanism. We used the ELM database to search for co-repressor interaction
274 motifs (Methods), and UniProt to search for domain annotations. We observe 72% of the RDs overlap
275 diverse annotations, such as sites for SUMOylation, zinc fingers (C2H2, PHD, CXXC, MYM), SUMO-
276 interacting motifs, co-repressor binding motifs (CtBP-, HP1-, TLE-binding), DNA binding domains
277 (Homeodomain DBDs, consistent with previous results⁴), and dimerization domains (bHLH, Leucine-
278 zipper) (Fig. 3a). To address whether these annotated sequences are necessary for repression, we
279 rationally designed mutant libraries that systematically replaced sections of 1,313 RDs (Supplementary
280 Table 5, Methods) and screened this RD mutant library using the pEF reporter and workflow described
281 in Fig. 1a (Extended Data Fig. 7b-c). We stained for protein expression (Extended Data Fig. 7d-e) and
282 filtered out mutants that had low FLAG enrichment scores.

283 First, we systematically searched and replaced the co-repressor interaction motifs with alanine to
284 test their contribution to activity (Fig. 3b). The TLE-binding motif, WRPW, appears exclusively in the C-
285 terminal RDs of the HES family and all tiles containing this motif were repressive (Extended Data Fig.
286 7f). All tested motifs were necessary for repression (Fig. 3b, left). The HP1-binding motif, PxVxL, was
287 necessary or contributed to repression in the majority of the tiles containing it (12/13 tiles with decreasing
288 effects Fig. 3b, middle). CtBP's binding motif, Px[DENS][LM]x, and the SUMO interaction motif, - ϕ x ϕ ϕ -
289 (non-covalent binding site to SUMOylated proteins, found in co-repressors that promote heterochromatin
290 formation such as SETD1a), are both relatively more flexible than the former two motifs and therefore
291 appeared in more RDs. However, in many RDs, they are not essential for function, as their deletion does
292 not decrease repression (Extended Data Fig. 8a-b). We found that a more refined CtBP motif of
293 PL[DN]Lx explained the majority of tiles that lost activity upon mutation (16/17 tiles Fig. 3b, right).
294 Altogether, 94% of the 36 repressing tiles with a co-repressor associated motif (TLE-, HP1-, or CtBP-
295 binding) decreased in repression strength when the motif was mutated, while 72% of 113 SUMO
296 interaction motif-containing repressing tiles were similarly sensitive to mutation (Extended Data Fig. 8b).

297 We were intrigued by the many RDs that contain a SUMOylation site (site for covalent conjugation
298 of a SUMO domain) (Fig. 3a). The ELM database classifies SUMOylation sites with the search pattern
299 ϕ KxE. Because this motif is short and relatively flexible, some non-hit sequences (12.3%) also contain
300 SUMOylation motifs. In order to investigate whether SUMOylation sites within non-hit sequences are
301 functional, we first used the AD deletion scan data. Deleting a SUMOylation motif within ADs rarely
302 decreased activation (Extended Data Fig 8c). Next, we asked if these motifs are functional in RDs using
303 the same deletion scanning approach (Supplementary Table 5, Fig. 3c). For example, residue K550 in
304 the SP3 protein is a SUMOylation site and has been shown before to be important for repression³⁰;
305 indeed we also find the SUMOylation site to overlap with the region essential for repression for this RD
306 of SP3 (Figure 3c). In a similar manner, we find SUMOylation motifs are important for the repression of
307 at least 149 out of the 166 RDs where they are found (Fig. 3d, Supplementary Table 5). This result is
308 concordant with our previous finding that a short 10 amino acid tile from the TF MGA, which contains this

309 SUMOylation motif, IKEE, is itself sufficient to be a repressor⁴. While the role of this modification in
310 repression still needs to be better understood, SUMOylation of certain TFs, such as FOXP1 (which also
311 shows up as a necessary region in our measurements, **Supplementary Table 5**), has been shown to
312 promote repression via CtBP recruitment^{31,32}. Our results suggest the pervasive role, across over a
313 hundred TFs, that SUMOylation plays in repression.

314 We next used our deletion scan data to gain better resolution of the region within RDs overlapping
315 dimerization domains, such as basic helix-loop-helix domains (bHLHs). Within bHLHs, the basic region
316 binds DNA, and mutations in the HLH region are known to impact dimerization³³. Deletion scans across
317 tiles that overlap HLH domains reveal part of helix 1, the loop region, and helix 2 are necessary for
318 repression (**Extended Data Fig. 8d**). The majority of RDs that overlap HLHs can be classified as Class
319 II, tissue specific dimerization domains that can either be activating or repressing depending on the
320 context³³ (**Extended Data Fig. 8e**). Our data suggests many Class II bHLHs can function as RDs. This
321 does not exclude the possibility bHLHs can also function as ADs, but we only observe NEUROG3's bHLH
322 activate the minCMV promoter, suggesting there is promoter specificity to activation of HLH domains.

323 Many RDs overlap annotated zinc fingers (n=124), and some specifically overlap C2H2 zinc
324 fingers (n=50, compared to only 3 ADs that overlap C2H2 zinc fingers) (**Fig. 3a**). We wondered if the
325 C2H2 domain itself or the protein sequence flanking it was responsible for repression. For example,
326 REST's zinc finger directly recruits the co-repressor coREST³⁴, and indeed REST deletions that had no
327 effect on repression (pink) corresponded to the disordered region just outside of the zinc finger, and
328 deletions necessary for repression (gray) corresponded to the zinc finger structural fold (**Extended Data**
329 **Fig. 8f**).

330 In addition to binding DNA and directly binding co-repressors, zinc fingers dimerize with other zinc
331 fingers³⁵. We reasoned some zinc fingers could cause repression by binding to other zinc finger domains
332 within endogenous repressive proteins. Support for this indirect recruitment of repressive TFs via zinc
333 fingers comes from the IKZF family where the N-terminus of some members, such as IKZF1, directly
334 recruits CtBP³⁶, while the C-terminal zinc fingers bind other IKZF family members³⁷. Indeed, we recover
335 the N-terminal repressive domains in IKZF1, and the associated sequence contains a CtBP binding motif
336 (**Extended Data Fig 8g**). In addition, all IKZF family members show C-terminal RDs that overlap the last
337 two zinc fingers (**Extended Data Fig 8g**). These two zinc fingers are both necessary for repression in
338 IKZF5 (**Fig. 3e**) and in all tested family members (**Extended Data Fig 8h**), and therefore likely dimerizes
339 with the IKZFs that recruit CtBP. While in general zinc fingers are well-known DNA binding domains, our
340 data expands the list of zinc finger sequences that are likely protein binding domains to other repressive
341 TFs (**Supplementary Table 5**).

342 In summary, repression domains can be categorized by their sequence properties in the following
343 way: (1) domains that contain short, linear motifs that directly recruit co-repressors, (2) domains that
344 contain SUMO interaction motifs or can be SUMOylated and most likely recruit co-repressors through
345 the conjugated SUMO domain, or (3) structured repressive protein binding domains that can recruit co-
346 repressors or other repressive TFs (**Fig. 3f, Extended Data Fig 8i**).

347

348 **Bifunctional activating and repressing domains**

349 Transcriptional proteins are often categorized as activating, repressing, or bifunctional⁸. Bifunctionality is
350 when the protein activates some promoters but represses others³⁸. There are 248 bifunctional CRs &
351 TFs that have both an AD and RD (such as in **Fig. 1b, Supplementary Table 2**). Additionally, we observe
352 bifunctionality at the domain level, wherein the same 80 aa tile both activated a minimal promoter and

353 repressed a constitutive promoter (**Fig. 4a-c, Supplementary Table 6**). We wondered if the 92
354 bifunctional domains we discovered appear in specific TF families and found many are within
355 homeodomain TFs (**Extended Data Fig. 9a**).

356 We individually validated bifunctional domains by flow cytometry, and confirmed doxycycline-
357 dependent activation of the minCMV and repression of the pEF reporter genes for all tested domains
358 (**Fig. 4b, Supplementary Table 3**). Some domains have both weak repression and activation, like the
359 tile from NANOG (**Fig. 4b, Extended Data Fig. 9b**). Some domains are stronger activators than
360 repressors, some stronger repressors than activators, and other domains show both strong activation
361 and repression (**Fig. 4b**). Together, the screen and validations demonstrate the CRTF tiling library can
362 be screened at multiple promoters to uncover bifunctional domains.

363 We hypothesized most bifunctional domains are similar to bifunctional proteins (as in **Fig. 1b**),
364 composed of smaller activating and repressing regions at independent locations. To address whether the
365 same exact sequence could be responsible for activation and repression we did a deletion scan across
366 all 92 bifunctional domains at the minCMV and pEF reporters (**Supplementary Table 6, Extended Data**
367 **Fig 9c-f**). These deletion scans revealed that some bifunctional tiles, including ones in NANOG
368 (**Extended Data Fig 9g**), have independent activating and repressing regions. In contrast, in other tiles,
369 the same amino acids are necessary for both activation and repression: for example, a single 14 aa
370 region mediated both the strong activation and repression for ARGFX tile 16 (**Fig. 4d**). Similarly, the
371 same 14 aa region is necessary for both activities for LEUTX, a TF in the same gene family as ARGFX
372 (**Extended Data Fig 10a-b**). In summary, a region as small as 14 amino acids can be necessary for both
373 activating and repressing activities, and as many as 69 other bifunctional domains similarly contain single
374 regions that are necessary for both activities (**Extended Data Fig 10c**).

375 In individual validations that measured activation over a time course, bifunctional ARGFX tile 16
376 (**Figure 4b**) was stronger at activating the minCMV promoter at the day 1 time point compared to day 2
377 (when the screen was measured) (**Extended Data Fig. 10d**), and in fact, the activated promoter slowly
378 silenced upon further recruitment to day 4. Intrigued by these dynamics, we tested several bifunctional
379 tiles at a promoter that has intermediate levels between minCMV and pEF, to see which direction they
380 would tune transcription. Surprisingly, when we recruited ARGFX tile 16 to the intermediate promoter, we
381 observed both a highly activated population of cells and a repressed population of cells after 5 days of
382 recruitment (**Fig. 4e**). Similar to the minCMV promoter, most cells initially increased in expression at day
383 1, then a subpopulation of cells silenced while another remained high (**Fig. 4e-f**). Other bifunctional tiles
384 recruited to the PGK promoter, from FOXO1 and ARGFX, led to similar dynamics that start with activation
385 and eventually end in a split of the cell population into silenced cells and cells that continue to express
386 (**Extended Data Fig. 10e**). These tiles, like ARGFX tile 16, had overlapping regions that are necessary
387 for both activities (**Supplementary Table 6**). However, not every bifunctional tile that activates minCMV
388 and represses pEF has bifunctional activity at the PGK promoter (**Extended Data Fig. 10e**): for example,
389 NANOG and KLF7 tiles do not significantly change expression of the PGK promoter. These tiles, in
390 contrast, have independent activating and repressing regions (**Extended Data Fig. 9g, Supplementary**
391 **Table 6**). In summary, some bifunctional tiles that independently activate and repress different promoters
392 are bifunctional even at a single promoter and can dynamically split a cell population into high- and low-
393 expressing cells.

394

395 Discussion

396 A systematic understanding of how transcriptional proteins function in human cells is needed to make
397 medical advances. When a new transcriptional protein is sequenced, homology models robustly identify
398 the DNA binding domain locations, but are unable to predict where the effector domains are³⁹. Compared
399 to DNA binding domains, many effector domain sequences are poorly conserved and do not align well
400 with one another in a multiple sequence alignment. As a result, we do not have nearly as robust nor as
401 comprehensive of predictors or sequence patterns for finding effector domains within protein sequences,
402 and thus need high-throughput experimental approaches for discovering them.

403 Here, we report the most comprehensive measurements to date of human transcriptional effector
404 domains. Via high-throughput tiling screens combined with systematic deletion scans and rational
405 mutagenesis, we collectively assigned transcriptional effector domains to 76% of the CRs and TFs
406 screened and comprehensively dissected the sequence properties that are necessary for activation and
407 repression.

408 The sequences that are necessary for function in ADs consist of certain hydrophobic amino acids
409 (W, F, Y, or L) that are interspersed with either acidic, proline, serine or glutamine residues (**Fig. 2i**).
410 Although prior work has shown homopolymeric stretches of glutamine and proline are sufficient to
411 activate a weak synthetic reporter²⁸, we find only OCT4's AD has proline repeats that are necessary for
412 activation. In fact, the majority of glutamine and proline repeats within ADs of the human CRs and TFs
413 are not part of the sequence necessary for activation. While these homotypic repeats might still be
414 important for other effects within the full-length TFs, such as solubility or nuclear localization, our data
415 suggests they are generally dispensable parts of AD sequences. In addition to the acidic exposure model,
416 our data suggests additional ways human ADs promote hydrophobic exposure, where serines could
417 functionally mimic acidic residues when phosphorylated, and prolines could promote exposure by their
418 intrinsic disorder. Furthermore, ADs contain certain hydrophobic amino acids, but our data suggest those
419 residues can be arranged in many ways, interspersed with serine, proline, and/or acidic residues. Unlike
420 RDs, we did not find any AD motifs, other than the previously reported LxxLL which appeared in 41
421 activating tiles (**Supplementary Table 4**). AD grammar flexibility might be related to their promiscuity in
422 binding, where many ADs have been shown to bind to more than one co-activator target⁴⁰, likely because
423 co-activators are a scarce resource in the cell⁴¹. ADs lacking motifs, or having little grammar, might impart
424 their flexibility by binding several different co-activators. In order to improve our understanding of ADs, it
425 will be important to take the next step and dissect how their sequence composition relates to their binding
426 to co-activators.

427 Strong sequence enrichment patterns across families have never been observed for RDs⁸. This
428 observation is likely due to the fact that there are many distinct functional categories of RDs. Indeed, 514
429 of our RDs overlap diverse functions, including co-repressor binding motifs, SUMO interaction motifs,
430 and structured binding domains. It has been shown before that the presence of the SUMO-1 domain
431 alone is sufficient to cause repression³⁰, and some well-characterized RDs contain SUMOylation sites³².
432 Here, we find hundreds of RDs that contain SUMOylation sites and show that repression activity is lost
433 upon deletion of these sites for the majority (>90%) of these RDs (**Figure 3d**). One mechanism by which
434 SUMOylation leads to repression is by recruitment of SUMO interaction motif-containing co-repressors
435 (**Extended Data Fig. 8i**). Supporting this, mutagenesis of SUMO interaction motifs within our set of RDs,
436 for example in SETD1a, led to a reduction of repression in 81/113 tested tiles (**Extended Data Fig. 8b**).
437 An alternative hypothesis for SUMOylation-mediated repression is that SUMOylation affects the TF's

438 localization within the nucleus towards regions associated with heterochromatin^{30,32}, but more
439 investigation into the mechanism of each SUMOylated RD will be needed.

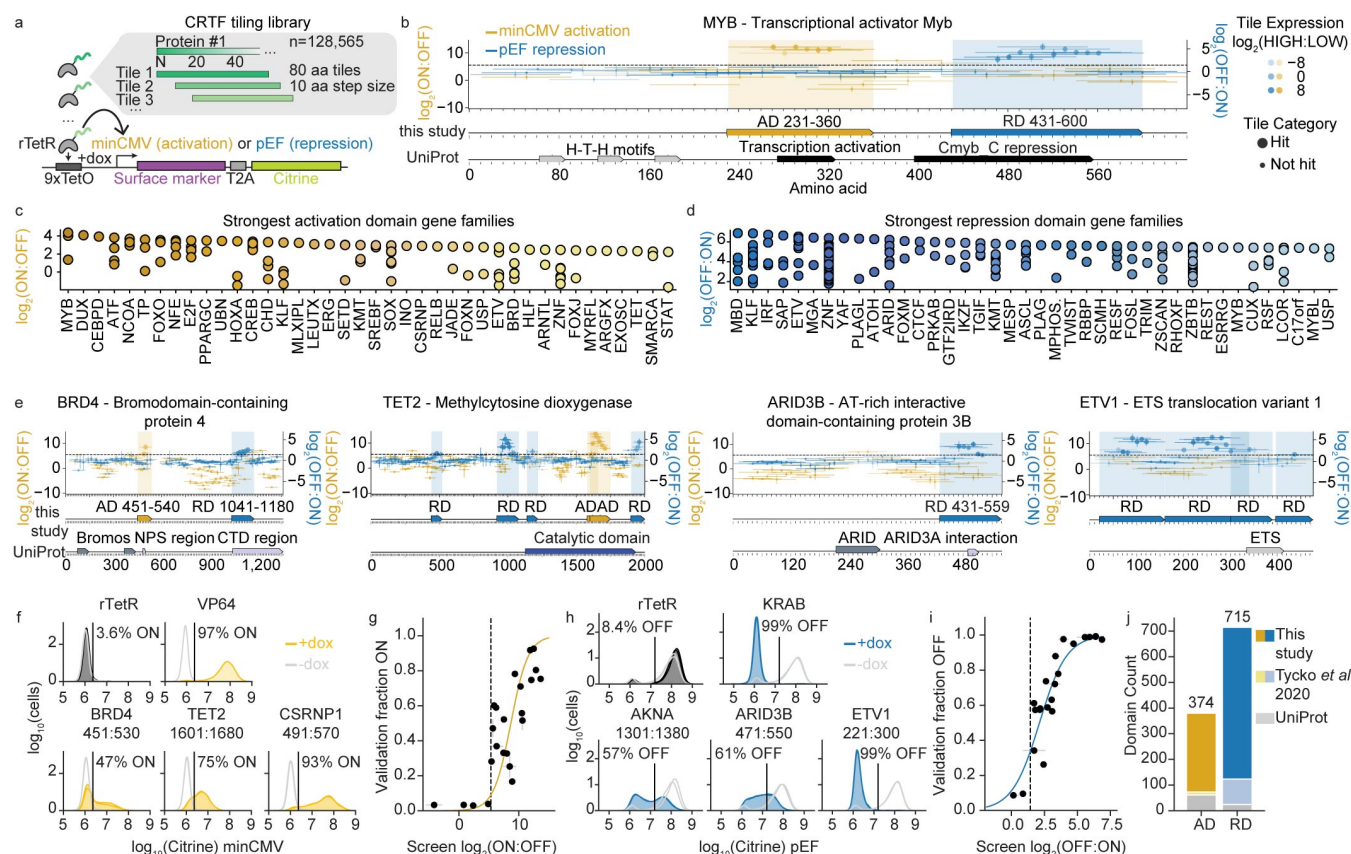
440 Zinc finger domains were originally identified as DNA binding domains, yet many of these domains
441 are also protein binding domains³⁵. Very few examples of transcriptionally repressive zinc finger domains
442 existed before this study. Here, our data suggest zinc finger domains are a prevalent repression
443 sequence, which are necessary for the repression of hundreds of domains. While a handful of zinc fingers
444 have been shown to interact with corepressors or repressive partner TFs, for most of them, their partners
445 remain to be found.

446 By systematically measuring both activation and repression of the same library, we were able to
447 find effector domains that can perform both roles. While bifunctional TFs that contain separate activator
448 and repressor domains have previously been observed⁴²⁻⁴⁴, to our knowledge, this is the first observation
449 of bifunctional domains that are capable of simultaneously enhancing and silencing expression from a
450 single promoter. Deletion scan data revealed activating and repressing regions within these bifunctional
451 domains can be very close to one another (less than 80 amino acids apart), and even overlapping in the
452 majority of domains. Previous observations of master transcriptional regulators activating some genes
453 and repressing others, such as NANOG^{45,46}, might be explained by this protein's bifunctional domain. We
454 find it interesting that bifunctional domains are most commonly found in the homeodomain family of TFs
455 (**Extended Data Fig. 9a**). Many homeodomain DBDs are not sufficiently specific to bind DNA on their
456 own and thus compensate by either having multiple motifs or multiple proteins helping the homeodomain
457 bind its enhancer^{47,48}. Therefore, the direction of the bifunctionality (whether the gene gets activated or
458 whether it becomes repressed) might be tuned by the DBD's motif within an enhancer/silencer. Evidence
459 for this hypothesis has been shown for the bifunctionality of the homeodomain TF CRX where observing
460 repression, in addition to activation, depends on the number of CRX binding motifs in a synthetic context
461 or the presence of other TF binding motifs in a genomic context^{49,50}. The functional difference between a
462 bifunctional protein's silencer or enhancer sequence might simply be explained by the bifunctionality of
463 the effector domain and the ratio of recruited activating and repressing complexes, where more binding
464 motifs will lead to repression. It will be interesting to determine how the CRs and TFs with bifunctional
465 domains that lead to a pulse of activation followed by repression of the gene affect development and
466 patterning.

467 Although we have acquired quite an extensive dataset, there is still more to be discovered by
468 using the same approach and libraries, and performing these high-throughput measurements in other
469 cell types and under different signaling conditions. Nevertheless, this is one of the largest high-throughput
470 assays for protein function performed in human cells to date, where we followed up with smaller high-
471 throughput validations and protein expression measurements in order to produce a high quality and
472 comprehensive dataset, moving one step closer to proteome-wide functional screening of protein
473 domains.

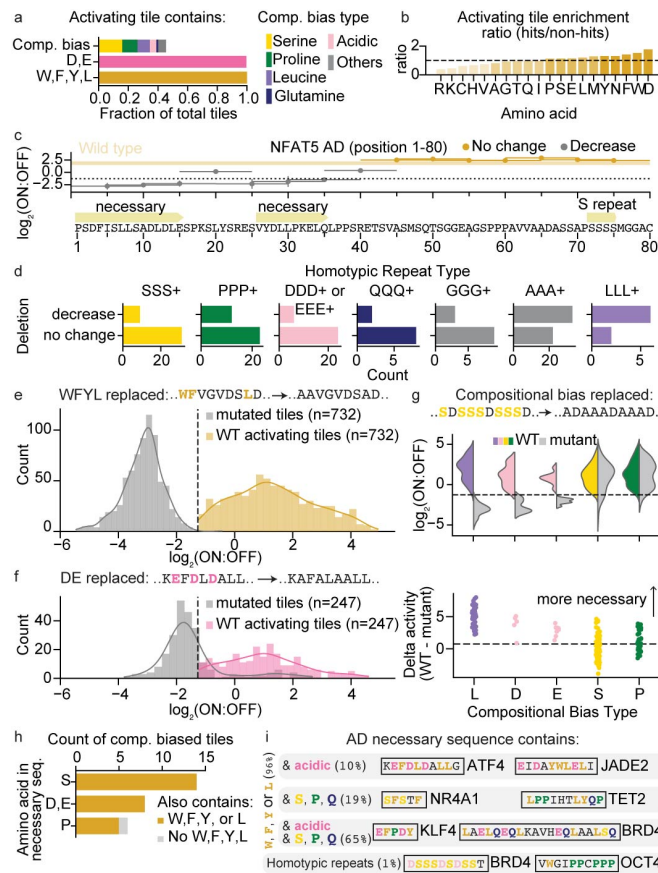
474 Now, this catalog can be used for improving sequence prediction models of transcriptional effector
475 domains, understanding molecular principles and the possible effects of CR and TF disease mutants,
476 and engineering better synthetic transcription factors and CRISPR systems⁵¹. We anticipate this resource
477 will enable exploration of uncharted functional genomic studies.

478 Figures



479

480 **Fig 1. | High-throughput tiling screen across 2,047 human TFs and CRs finds hundreds of undiscovered**
 481 **effector domains.** **a**, Each protein's activating and repressing regions are identified by partitioning the full-length sequence
 482 into 80 aa tiles. Each tile is fused to rTetR-3xFLAG, and the library is delivered with lentivirus to K562 cells containing a magnetic-
 483 fluorescent reporter (surface marker and citrine) stably integrated in the safe harbor AAVS1 locus. The reporter contains 9 TetO
 484 binding sites for recruiting a rTetR-tile fusion upon dox addition (+dox). Activation is measured after recruitment upstream of a
 485 minimal promoter (minCMV) for 2 days, and repression after recruitment upstream of a constitutive promoter (pEF) for 5 days.
 486 **b**, Activation (at minCMV, yellow) and repression (at pEF, blue) tiling enrichment scores overlap previously annotated MYB
 487 effector domains (bottom, black, sourced from UniProt). Each horizontal yellow and blue line represents an 80 aa tile, and each
 488 vertical error bar is the standard error from two biological replicates. The dashed horizontal line represents the hit calling
 489 threshold based on random controls (Methods). Points with larger marker sizes were hits in the second (validation) screen.
 490 FLAG-stained expression levels are plotted as the hue, with higher expressing-tiles in darker hues. Effector domains identified
 491 in this study are annotated as contiguous regions at the bottom: yellow bars for ADs and blue bars for RDs (Methods). **c-d**,
 492 Distribution of the strongest effector domains from the top 40 gene families. Enrichment scores are from the validation screen
 493 (Extended Data Fig. 3), measured for the maximum activating/repressing tile within each domain. **e**, Tiling results for BRD4,
 494 TET2, ARID3B, and ETV1. **f**, Individual validations of activating tiles after 2 days of recruitment (+dox). Untreated cells (gray)
 495 and dox-treated cells (colors) shown with two biological replicates in each condition. Vertical line is the citrine gate used to
 496 determine the fraction of cells ON (written above each distribution). rTetR alone is a negative control and VP64 is a positive
 497 control. **g**, Comparison between individually recruited and screen measurements with logistic model fit plotted as solid line
 498 ($r^2=0.67$, $N=23$). Error bars are the standard error for 2 biological replicates (screens and validations) and dashed line is the hits
 499 threshold. **h**, Individual validations of repressing tiles after 5 days of recruitment ($n=2$). KRAB is a positive control. **i**, Comparison
 500 between individually recruited and pEF promoter screen measurements ($r^2=0.84$, $N=22$). **j**, Domain counts that are new (dark
 501 gold and blue), overlap UniProt annotations (gray), or overlap prior HT-recruit screen results⁴ (light gold and blue). Total is shown
 502 above each bar. RDs are annotated from tiles that were hits in both pEF and PGK promoter screens (Extended Data Fig. 4).



503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

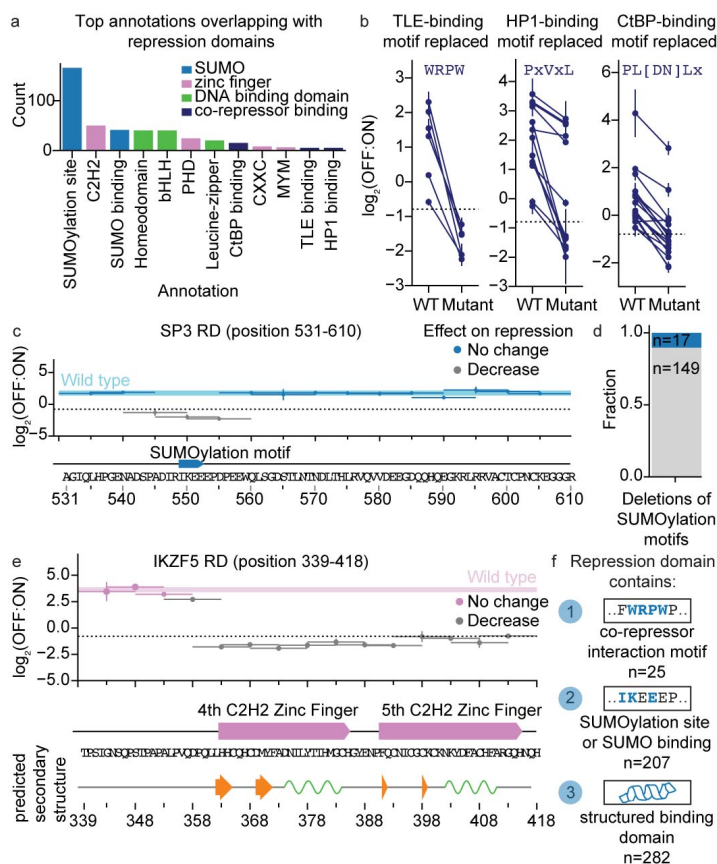
527

528

529

530

Fig. 2 | Hydrophobic amino acids that are interspersed with acidic, serine, proline or glutamine residues are necessary for AD activity. **a**, Fraction of activating tiles that contain a compositional bias (amino acids that appeared at least 12 times in the 80 aa, i.e. 15% of the sequence). Number of tiles in each compositional bias type: serine=132, proline=85, leucine=69, acidic=36, glutamine=10, others are alanine=21, glycine=13, asparagine=6, and methionine=3. **b**, Enrichment ratio for each amino acid across all activating tile sequences. Ratios were computed by counting the abundance of each amino acid in the hit sequences, 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. Horizontal dashed line is at a ratio of 1. **c**, Deletion scan across NFAT5's AD. Yellow rectangle represents WT enrichment score, with the standard error between two biological replicates represented as the height. If the deletion's score is lower than 2 times the average standard error for measuring a deletion, it's binned as "decrease." Otherwise it's binned as "no change." **d**, Counts of deletion sequences containing a homotypic repeat of 3 or more amino acids of the indicated type binned according to their effect on activity compared to the WT sequence: decrease or no change upon deletion. Probability we would observe serine ratio $p=7.58e-3$, proline= $5.17e-2$, acidic= $6.57e-3$, glutamine= $1.73e-3$, glycine= $6.48e-2$, alanine=0.699 (Fisher's exact test compared with LLL+ distribution, two-sided) **e**, Distribution of average activation enrichment scores (2 biological screen replicates) for WT (yellow) and W,F,Y,L mutant tiles (gray) for all well-expressed W,F,Y,L-containing activating tiles. Dashed line represents the hit threshold. **f**, Distribution of average activation enrichment scores (2 biological screen replicates) for WT (yellow) and D,E mutant tiles (gray) for all well-expressed D,E-containing activating tiles. **g**, (Top) Distributions of average activation enrichment scores (2 biological screen replicates) for WT (colored) and compositional bias mutants (gray). Dashed line represents the hit threshold. (Bottom) Mutant enrichment scores subtracted from WT enrichment scores plotted for each compositional bias that was replaced with alanine. Dashed line drawn 2 times the average standard error (across all mutants) above 0. **h**, Count of all compositionally biased tiles that lost activity upon mutation that contain the compositionally biased amino acid in at least one of its necessary sequences and whether an aromatic/leucine was also present (yellow). Probability we would observe this for serine: $p=3.75e-4$, acidic residues: $p=3.0e-3$, proline residues: $p=5.45e-1$ (Fisher's exact test comparing counts of tiles that had W,F,Y, or L present with a size-matched, randomly selected distribution of sequences that had no change upon deletion, two-sided). Deletion scans were only performed on the max activating tile from each AD, so only the max tile from a compositionally biased AD has a corresponding necessary sequence. **i**, Summary of findings: AD sequences that are necessary for function consist of hydrophobic amino acids that are interspersed with acidic, prolines, serines and/or glutamine residues.



531

532

533

534

535

536

537

538

539

540

541

542

543

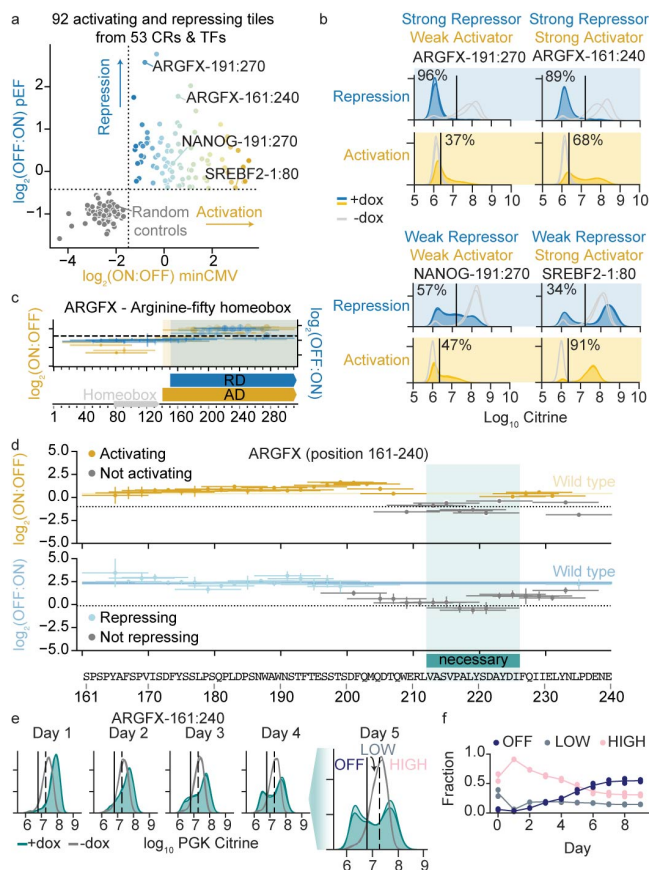
544

545

546

547

Fig. 3 | The majority of RD sequences contain either sites for SUMOylation, short interaction motifs for recruiting co-repressors, or are structured binding domains for recruiting other repressive proteins. a, Count of repression domains (repressive in both pEF and PGK promoter screens) that overlap annotations from UniProt and ELM (Eukaryotic Linear Motifs). Annotations that had at least 6 counts are shown. **b**, Repression enrichment scores for tiles that contain a co-repressor binding motif (WT) and the co-repressor binding motif replaced with alanines (mutant). TLE-binding: 6 lost all repressive activity upon motif removal. Fraction of non-hit sequences containing motif=0. HP1-binding: 7 lost all repressive activity, 5 decreased, 1 had little change. Fraction of non-hit sequences containing motif=0.002. CtBP-binding: 10 lost all repressive activity, 6 decreased, 1 had little change. Fraction of non-hit sequences containing motif=0.002. 2 biological replicates shown with standard error. **c**, Deletion scan across SP3's RD. SUMOylation motif is "IKEE" (indicated on the bottom). Blue shaded bar spanning the entire domain length above the threshold represents the WT enrichment score, where the standard error between two biological replicates is represented as the height. Deletions were binned into those that had an effect on repression (gray lines) and those that did not (dark blue lines). **d**, Fraction of deletion sequences containing a SUMOylation motif binned according to their effect on activity (blue=no change on repression relative to WT, gray=decreased repression relative to WT, n=166 total RDs). **e**, Deletion scan across IKZF5's RD. AlphaFold's predicted secondary structure (prediction from whole protein sequence) shown below where green regions are alpha helices and orange arrows are beta sheets. **f**, Summary of repression domain functional sequence categories (n indicated in Figure).



548

549

550

551

552

553

554

555

556

557

558

559

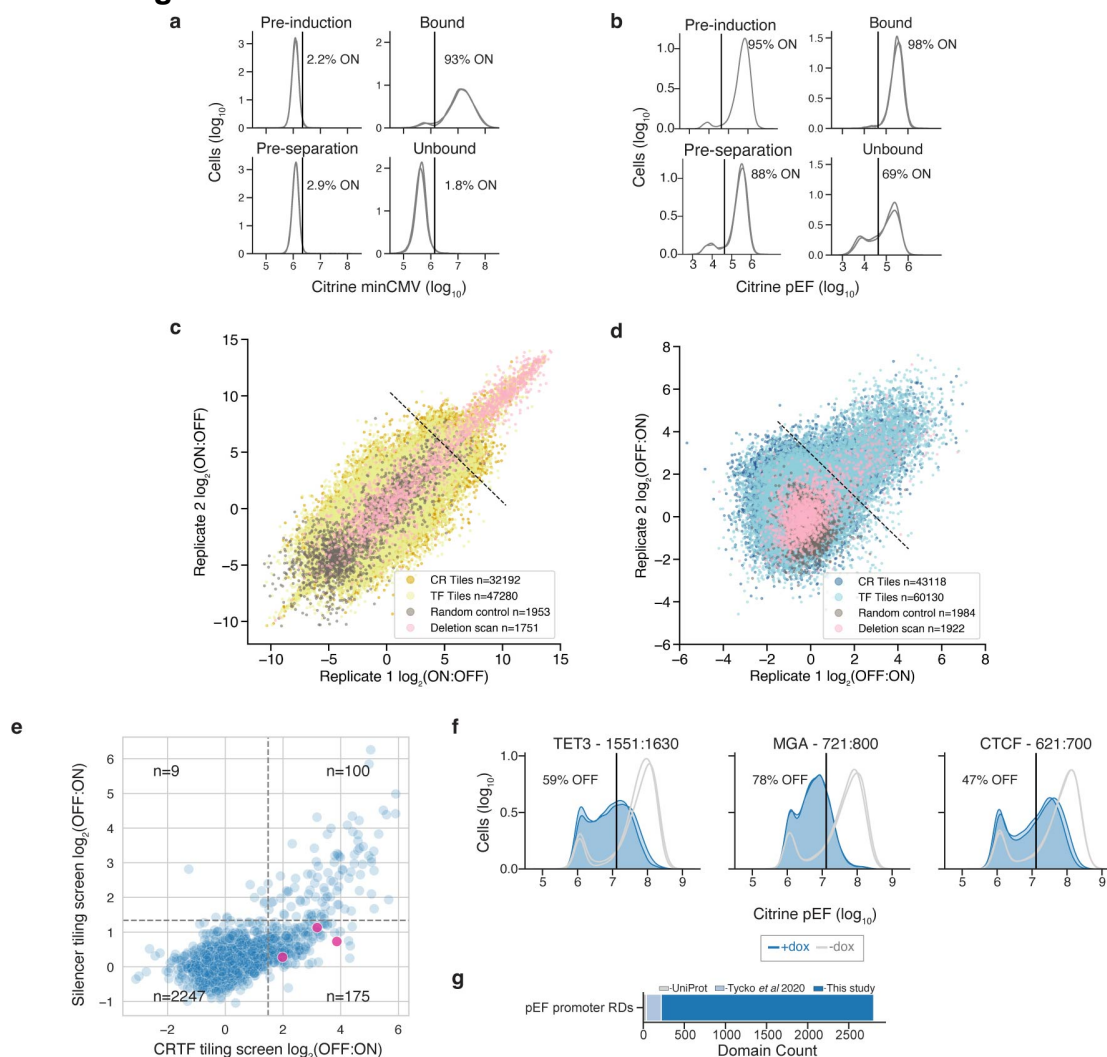
560

561

562

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). Average across two biological replicates shown for each point. **b**, Individual validations of bifunctional tiles. 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 for activation and OFF for repression. **c**, Tiling plot for ARGFX. 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, with the standard error between two biological replicates represented as the heights. The 3 deletions that caused no activation and no repression across both screens are highlighted in teal and the sequence annotated as necessary. **e**, Citrine distributions of bifunctional tile ARGFX-161:240 recruited 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 were measured as the cells in between both gates. **f**, Fraction of ARGFX-161:240 cells OFF (navy), LOW (gray), and HIGH (pink) over time (2 biological replicates plotted with the average plotted as a line).

563 Extended Data Figures



564

565

566

567

568

569

570

571

572

573

574

575

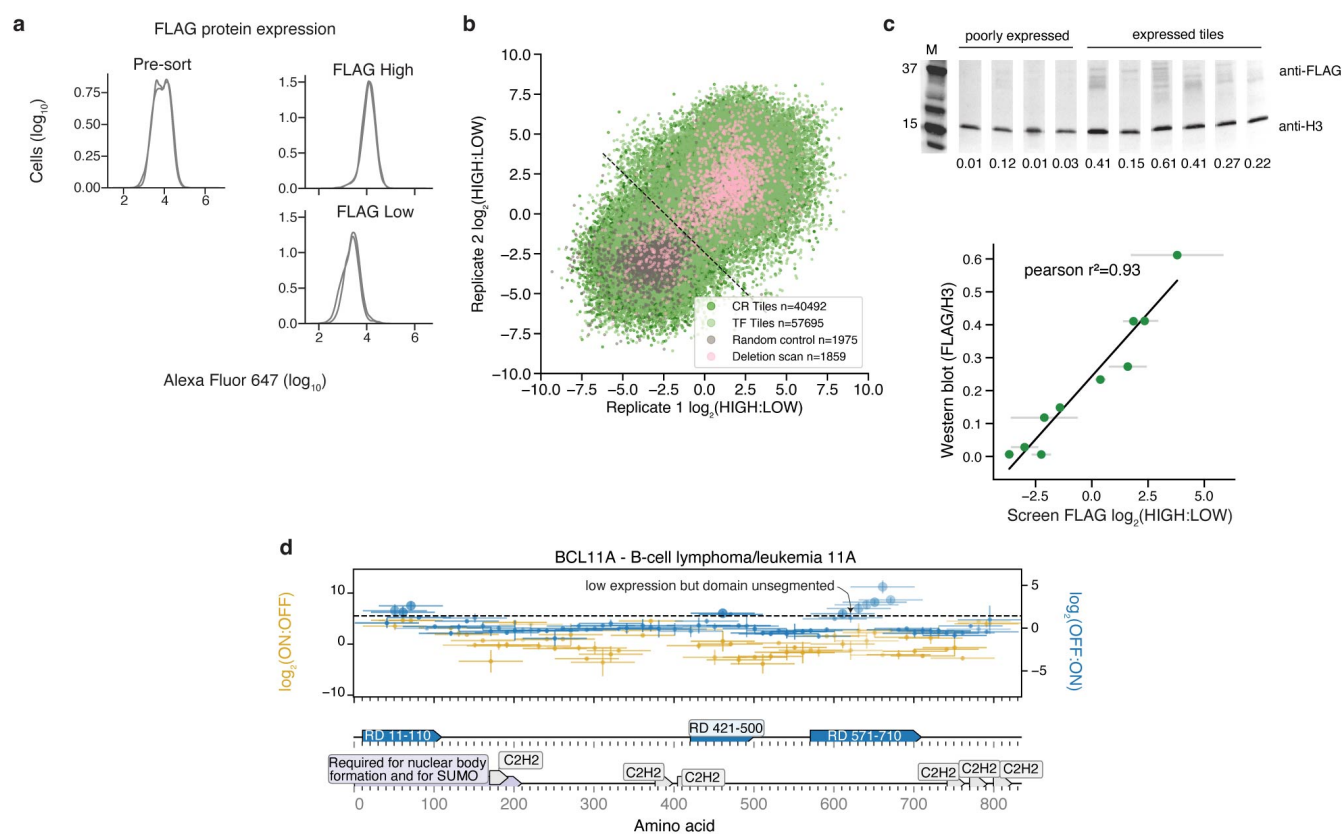
576

577

578

Extended Data Fig. 1 | CRTF tiling screen's 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), 2 days later on the day of magnetic separation (Pre-separation), and after separation using ProG DynaBeads that bind to the surface synthetic 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. A total of 1,000 ng/mL dox was added each day of dox treatment. **b**, Citrine reporter distributions for the pEF promoter screen (n=2). Pre-separation was after 5 days of dox treatment. **c-d**, Biological replicate screen reproducibility (pearson $r^2=0.78$ for minCMV and $r^2=0.19$ for pEF hits). **e**, Comparison between 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 amino acid 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 **f**. **f**, Citrine reporter distributions of individually validated CRTF tiling pEF screen hits that were not identified within the Silencer tiling screen. **g**, Counts of RDs annotated from tiles that were hits in the pEF promoter screen. Domain counts that are new (dark blue), overlap UniProt annotations (gray), or overlap prior HT-recruit screen results⁴ (light blue). Total of 2,803 domains, where 2,585 are new when recruited at the pEF promoter.

579



580

581

582

583

584

585

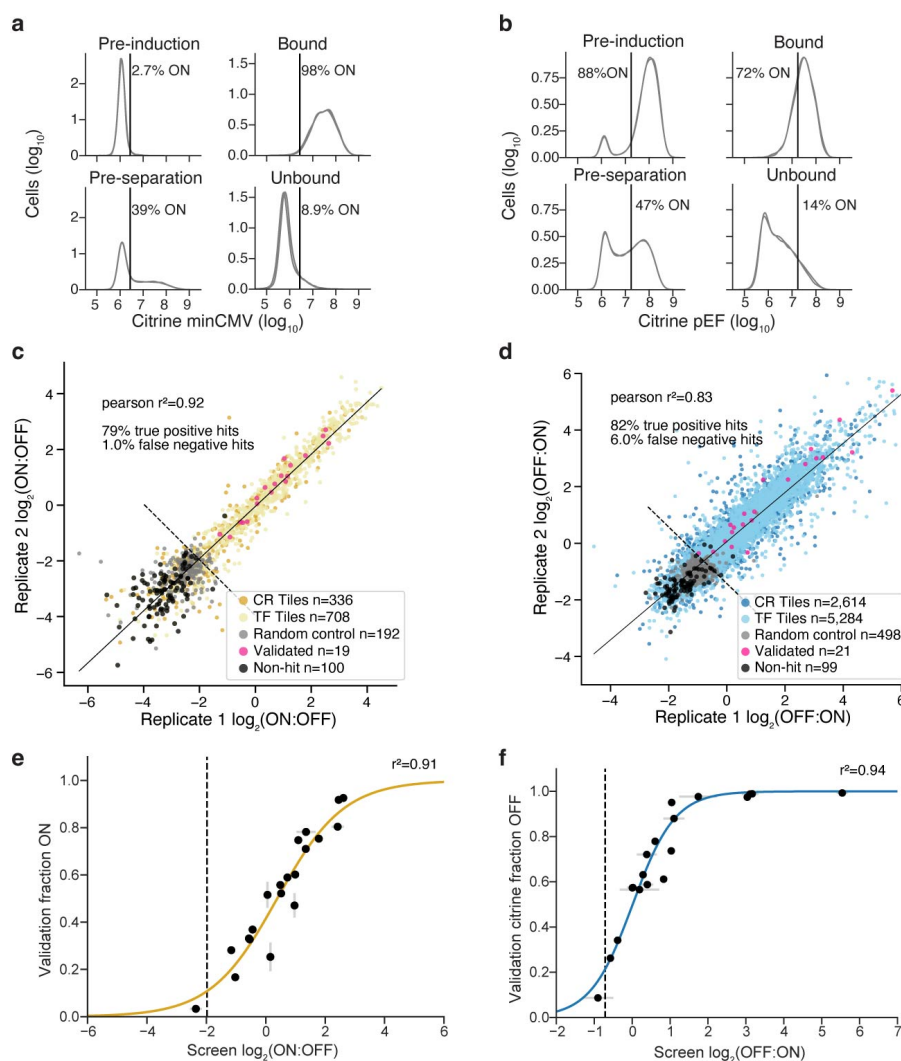
586

587

588

589

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. **c**, Validation of expression level for a panel of tiles. Expression level was measured by western blot with an anti-FLAG antibody. Anti-histone H3 was used as a loading control for normalization. Levels were quantified from all bands in each lane using ImageJ. Superfluous lanes from the gel are cropped out and the relevant lanes are shown consecutively with white lines between each lane. Comparison between high-throughput measurements of expression and western blot protein levels ($r^2=0.93$). **d**, Tiling plot for BCL11A. 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.



590

591

Extended Data Fig. 3 | CRTF tile hits validation screen's separation purity, reproducibility, and validation.

592

a, Flow cytometry data showing citrine reporter distributions for the minCMV promoter screen on the day we induced localization

593

with dox (Pre-induction), 2 days later on the day of magnetic separation (Pre-separation), and after separation using ProG

594

DynaBeads that bind to the surface synthetic marker (Bound). Overlapping histograms are shown for 2 biological replicates. The

595

average percentage of cells ON is shown to the right of the vertical line showing the citrine level gate. A total of 1,000 ng/mL dox

596

was added each day of dox treatment. **b**, Citrine reporter distributions for the pEF promoter validation screen (n=2). Pre-

597

separation was after 5 days of dox treatment. **c-d**, Biological replicate screen reproducibility. **e**, Comparison between individually

598

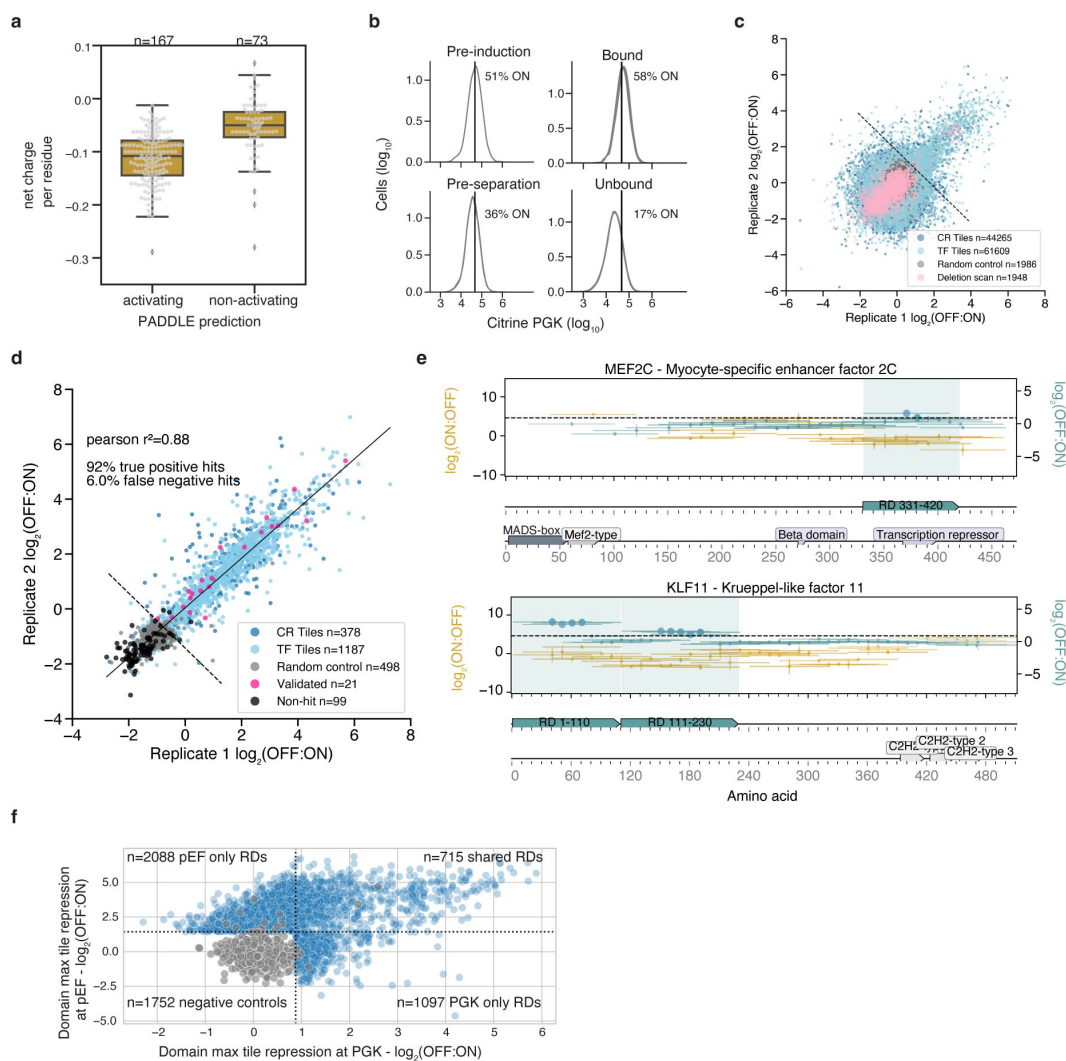
recruited measurements and minCMV promoter validation screen measurements with logistic model fit plotted as solid line

599

($r^2=0.91$, $N=20$). **f**, Comparison between individually recruited measurements and pEF promoter validation screen

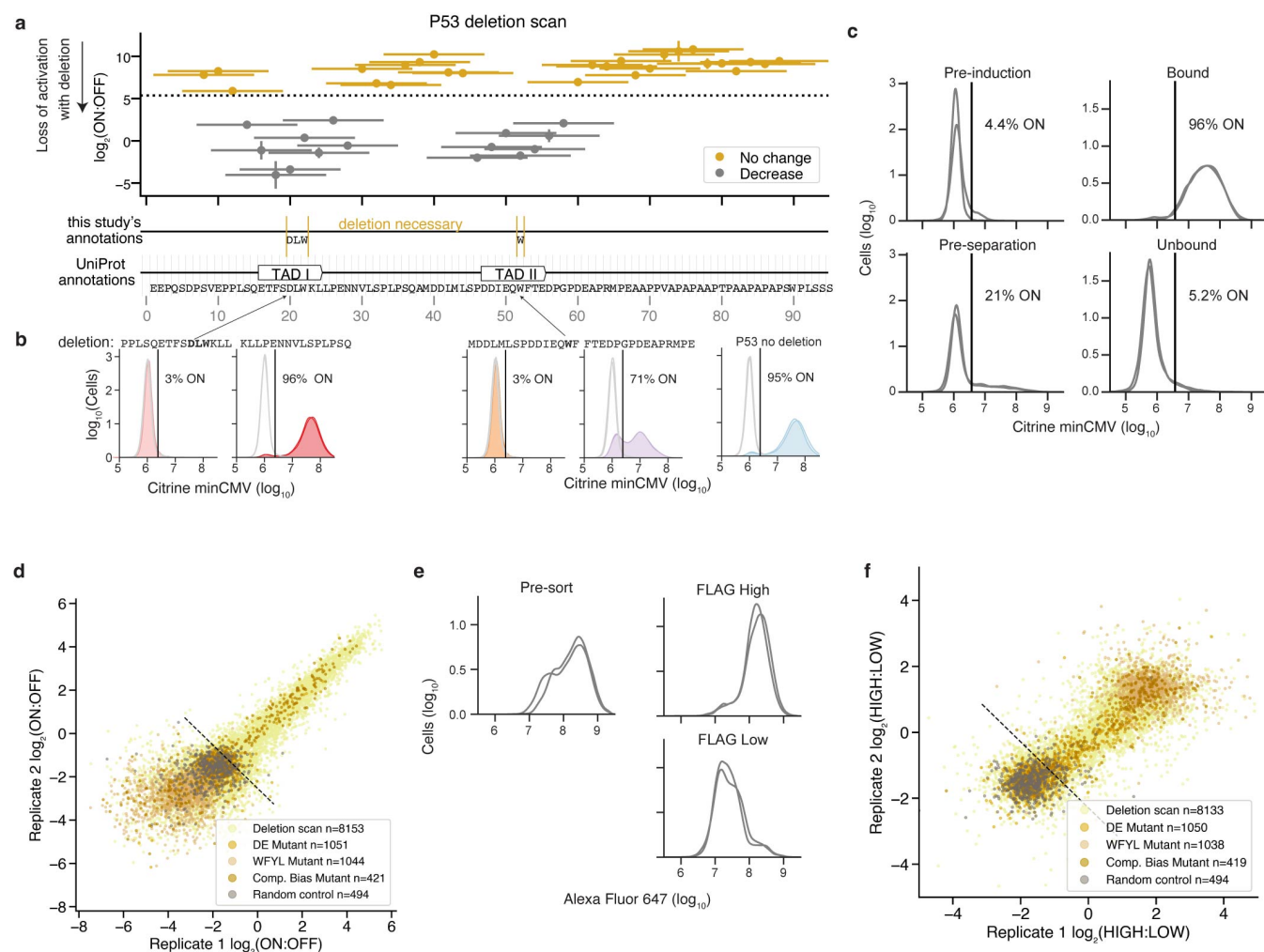
600

measurements with logistic model fit plotted as solid line ($r^2=0.94$, $N=19$).



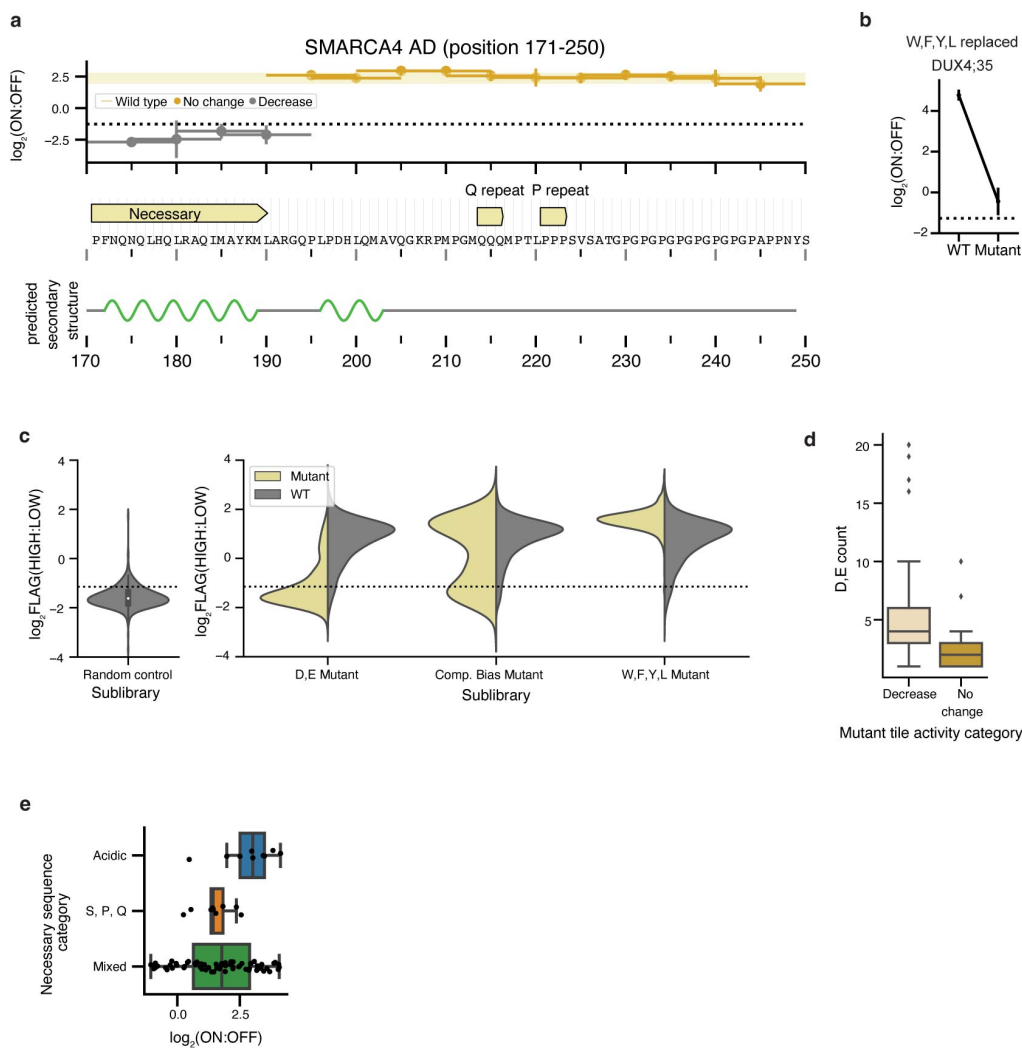
601
602
603
604
605
606
607
608
609
610
611
612

Extended Data Fig. 4 | Validations of CR & TF effector domains. **a**, Net charge per residue distributions (calculated by CIDER⁵²) of activation domains identified by HT-recruit compared to their PADDLE-predicted function¹¹ (Mann-Whitney p-value=1.39e-15). **b**, 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 using ProG DynaBeads that bind to the surface synthetic marker (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. A total of 1,000 ng/mL dox was added each day of dox treatment. **c**, Biological replicate PGK promoter screen reproducibility (pearson $r^2=0.27$ for hits). **d**, Validation screen biological replicate reproducibility of tiles that were hits in both the PGK and pEF promoter screens. **e**, Tiling plots for MEF2C and KLF11. PGK repression domains annotated in teal. **f**, Comparison of each repression domain's max tile repression scores in PGK (x-axis) and pEF promoter screen (y-axis). Dashed lines are the hits thresholds for each screen.



613
614
615
616
617
618
619
620
621
622
623
624
625

Extended Data Fig. 5 | Mutant AD screen's separation purity, reproducibility, and validation. **a**, Deletion scan across P53's AD. If the deletion's score is lower than 2 times the average standard error for measuring a deletion, it's binned as "decrease". Otherwise it's binned as "no change." **b**, Individual validations of 80 aa sequences 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). **c**, Flow cytometry data showing citrine reporter distributions for the Mutant AD transcriptional activity screen on the day we induced localization with dox (Pre-induction), 2 days later on the day of magnetic separation (Pre-separation), and after separation using ProG DynaBeads that bind to the surface synthetic 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. A total of 1,000 ng/mL dox was added each day of dox treatment. **d**, Biological replicate Mutant AD transcriptional activity screen reproducibility. **e**, Alexa Fluor 647 distributions from anti-FLAG staining. **f**, Biological replicate Mutant AD protein expression screen reproducibility.



626

627

628 **Extended Data Fig. 6 | Mutant AD screen follow-up.** **a**, Deletion scan across SMARCA4's AD. AlphaFold's predicted

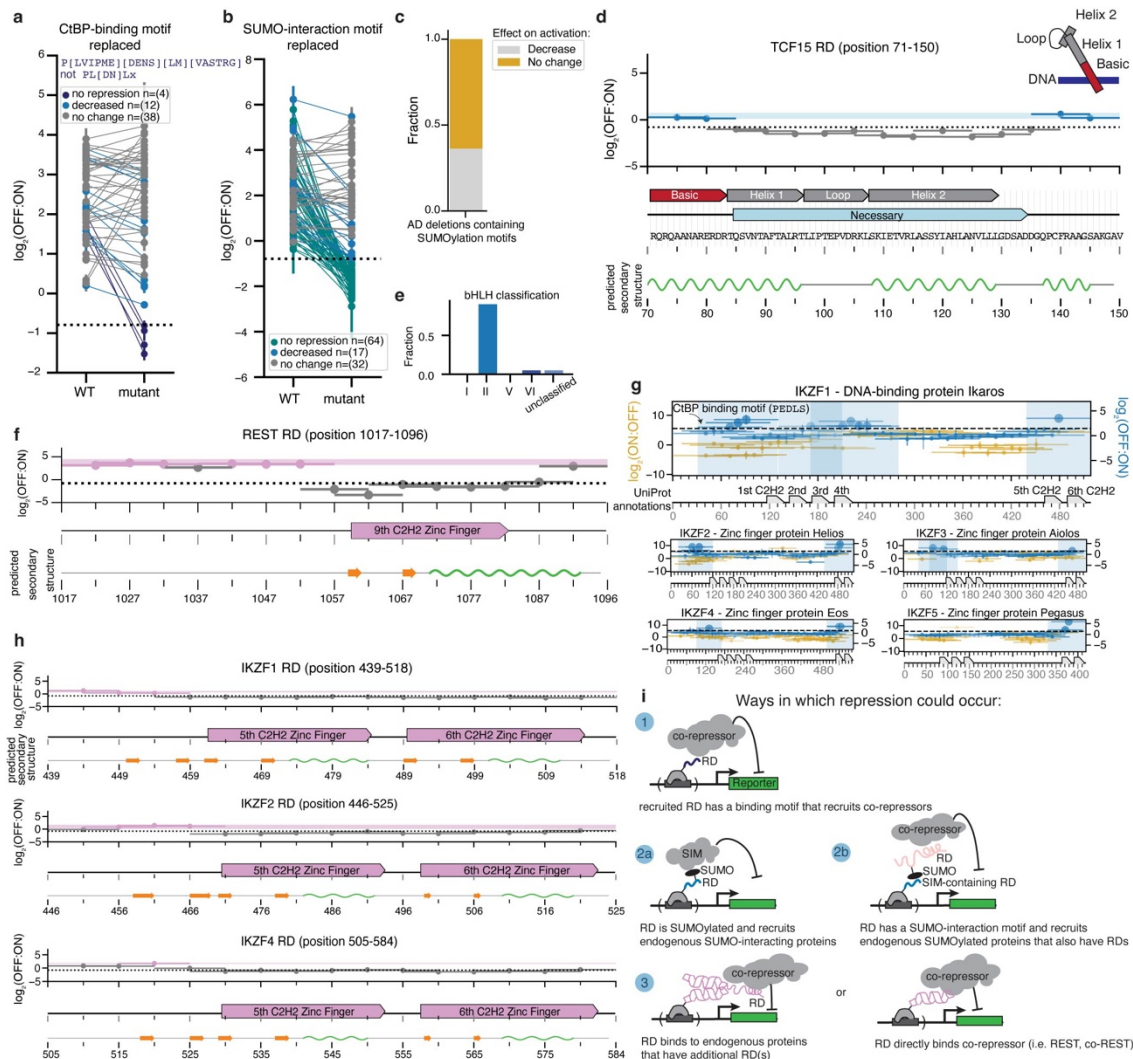
629 secondary structure (prediction from whole protein sequence) shown below where green regions are alpha helices. **b**, Line plot

630 of average enrichment scores from two biological replicates. **c**, Violin plots of average FLAG enrichment scores from 2 biological

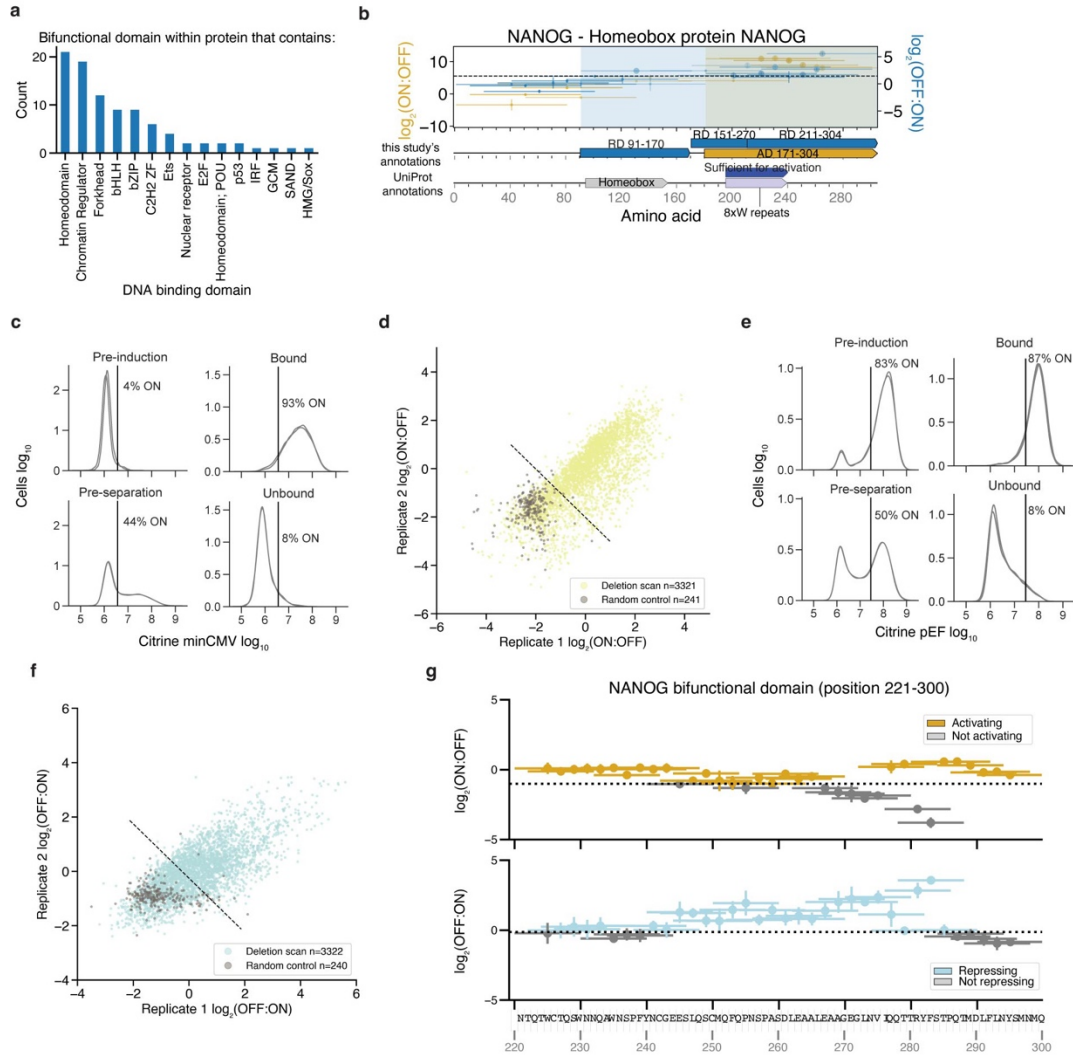
631 replicates binned by each sublibrary. Dashed line represents the hit threshold. **d**, Boxplot of acidic count for each mutant's

632 activation category. Mann-Whitney one-sided U test, p -value= $2.25e-3$. **e**, Boxplot of average activation enrichment scores with

IQR shown for tiles that contain a single necessary sequence across each category.



645
 646 **Extended Data Fig. 8 | Mutant RD screen follow-up.** **a**, Repression enrichment scores for a subset of repressing tiles
 647 that contain a relatively more flexible CtBP-binding motif (regex shown above), excluding the more refined CtBP-binding motif
 648 (regex shown on second line). Mutants have their binding motifs replaced with alanines. **b**, Repression enrichment scores for
 649 repressing tiles that contain a flexible SUMO-binding motif (fraction of non-hit sequences containing motif=0.155). 2 biological
 650 replicates shown with standard error bars. **c**, Fraction of AD deletion sequences containing a SUMOylation motif binned
 651 according to their effect on activity (yellow=no change on activation relative to WT, gray=decreased activation). 11 total ADs. **d**,
 652 Deletion scan across TCF15's RD. AlphaFold's predicted secondary structure (prediction from whole protein sequence) shown
 653 below where green regions are alpha helices. Annotations shown from protein accession NP_004600.3 **e**, Distribution of bHLH
 654 classifications of RDs overlapping bHLH UniProt annotations. Classifications taken from ref³³. **f**, Deletion scan across REST's
 655 RD. AlphaFold's predicted secondary structure (prediction from whole protein sequence) shown below where green regions are
 656 alpha helices and orange arrows are beta sheets. **g**, Tiling plots for IKZF family members. **h**, Deletion scan across IKZF1, 2
 657 4's RDs. **i**, Cartoon model of potential mechanisms corresponding to the RD categories in Fig. 7f.



658

659

660

661

662

663

664

665

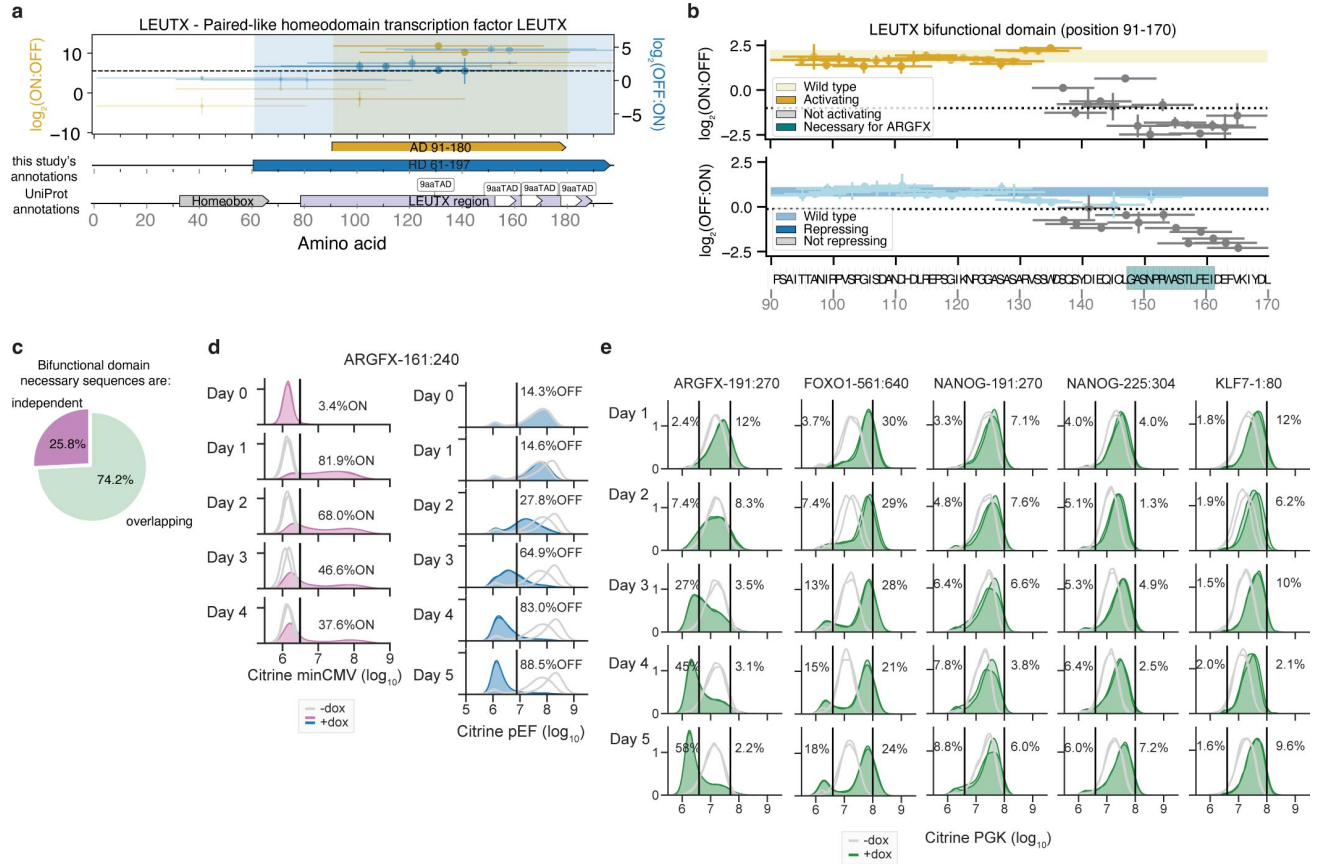
666

667

668

669

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 following DNA binding domains. **b**, Tiling plot for NANOG. **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), 2 days later on the day of magnetic separation (Pre-separation), and after separation using ProG DynaBeads that bind to the surface synthetic 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. A total of 1,000 ng/mL dox was added each day of dox treatment. **d**, Biological replicate bifunctional deletion scan minCMV promoter screen reproducibility. **e**, Citrine reporter distributions for the bifunctional deletion scan pEF promoter screen, 5 days of induction (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). Note, deletion of the necessary sequence for activation, caused an increase in repression, and vice-versa.



670

671

672 **Extended Data Fig. 10 | Examples of bifunctional domain sequences at three different promoters. a,**

673 **b,** Deletion scan across one of LEUTX's bifunctional tiles (n=2). The necessary sequence for gene family

674 **member, ARGFX, is highlighted in teal. c,** Bifunctional domain necessary region location categories. Overlapping regions were

675 **defined as any tile that contained a deletion that was both necessary (below activity threshold) for activation and necessary for**

676 **repression. d,** Citrine distributions of ARGFX-161:240 recruited to minCMV (n=2, left), and recruited to pEF (n=2, right). **e,** Citrine

677 **Supplementary Tables**

678

679 **Supplementary Table 1: CRTF Tiles**

680 CRTF tiling library sequences and enrichment scores from the FLAG protein expression screen,
681 minCMV, pEF, and PGK promoter screens, and the validation screens are attached in an Excel
682 file.

683

684 **Supplementary Table 2: Domains from Tiles**

685 Activation and repression domain sequences and maximum tile enrichment scores are attached
686 in an Excel file.

687

688 **Supplementary Table 3: Validations**

689 Individual validation flow cytometry data are attached in an Excel file.

690

691 **Supplementary Table 4: AD Mutants**

692 AD mutants library sequences and enrichment scores from the FLAG protein expression screen
693 and minCMV promoter screen are attached in an Excel file.

694

695 **Supplementary Table 5: RD Mutants**

696 RD mutants library sequences and enrichment scores from the FLAG protein expression screen
697 and pEF promoter screen are attached in an Excel file.

698

699 **Supplementary Table 6: Bifunctional Domains**

700 Bifunctional domains, deletion scan library sequences and enrichment scores from the FLAG protein
701 expression screen, minCMV, and pEF promoter screens are attached in an Excel file.

702 **Methods**

703 **Cell culture**

704 All experiments presented here were carried out in K562 cells (ATCC, CCL-243, female). Cells were
705 cultured in a controlled humidified incubator at 37C and 5% CO₂, in RPMI 1640 (Gibco, 11-875-119)
706 media supplemented with 10% FBS (Takara, 632180), and 1% Penicillin Streptomycin (Gibco, 15-140-
707 122). HEK293T-LentiX (Takara Bio, 632180, female) cells, used to produce lentivirus, as described
708 below, were grown in DMEM (Gibco, 10569069) media supplemented with 10% FBS (Takara, 632180)
709 and 1% Penicillin Streptomycin Glutamine (Gibco, 10378016). minCMV and pEF reporter cell line
710 generation is described in ref⁴. Briefly, pEF and minCMV promoter reporter cell lines were generated by
711 TALEN-mediated homology-directed repair to integrate donor constructs (pEF promoter: Addgene
712 #161927, minCMV promoter: Addgene #161928) into the *AAVS1* locus by electroporation of K562 cells
713 with 1000 ng of reporter donor plasmid and 500 ng of each TALEN-L (Addgene #35431) and TALEN-R
714 (Addgene #35432) plasmid (targeting upstream and downstream the intended DNA cleavage site,
715 respectively). After 7 days, the cells were treated with 1000 ng/mL puromycin antibiotic for 5 days to
716 select for a population where the donor was stably integrated in the intended locus. Fluorescent reporter
717 expression was measured by microscopy and by flow cytometry. The PGK reporter cell line was
718 generated by electroporation of K562 cells with 0.5 ug each of plasmids encoding the *AAVS1* TALENs
719 and 1 ug of donor reporter plasmid using program T-016 on the Nucleofector 2b (Lonza, AAB-1001).
720 Cells were treated with 0.5 ug/mL puromycin for one week to enrich for successful integrants. These cell
721 lines were not authenticated. All cell lines tested negative for mycoplasma.

722

723 **TF tiling library design**

724 1294 human transcription factors (TFs) were curated from ref¹. We filtered out all KRAB-containing C2H2
725 zinc fingers, as our lab has previously screened the effector domains (KRAB, SCAN, DUF) of these
726 proteins⁴. The canonical transcript of each gene was retrieved from Ensembl and chosen using the
727 APPRIS principle transcript⁵³. If no APPRIS tag was found, the transcript was chosen using the TSL
728 principle transcript. If no TSL tag was found, the longest transcript with a protein coding CDS was
729 retrieved. The coding sequences were divided into 80 amino acid (aa) tiles with a 10 aa sliding window.
730 For each gene, a final tile was included spanning from 80 aa upstream of the last residue to that last
731 residue, such that the C-terminal region would be included in the library. Duplicate sequences were
732 removed, sequences were codon matched for human codon usage, 7xC homopolymers were removed,
733 BsmBI restriction sites were removed, rare codons (less than 10% frequency) were avoided, and the GC
734 content was constrained to be between 20% and 75% in every 50 nucleotide window (performed with
735 DNA chisel⁵⁴). To improve the coverage of this large library, we subdivided into 3 smaller sub-libraries
736 based on the three major classes of TFs: a 25,032 C2H2 ZF sub-library including all 406 C2H2 ZF TFs,
737 a 9,757 Homeodomain and bHLH sub-library including all 304 Homeodomain and bHLH TFs, and a
738 31,664 member sub-library containing the rest of the 583 TFs.

739 1000 random controls of 80 amino acids lacking stop codons were computationally generated as
740 controls using the DNA chisel package's *random_dna_sequence* function and included in each sub-
741 library. 473 sequences that were found to be non-activators and 42 sequences that were found to be
742 activators in our lab's previous minCMV Nuclear Pfam screen⁴ were included as negative and positive
743 controls. We made use of alternative codon usage (*match_codon_usage*, and *use_best_codon*
744 functions) to re-code the controls in each sub-library in order to give ourselves the option of pooling the
745 3 sub-libraries and running the library as one 73,288 element screen.

746 100 additional controls were added to each sub-library to serve as fiduciary markers to aid
747 comparing separately run screens. These controls were not recoded in each sub-library, and thus were
748 repeated when pooling sub-libraries.

749 50 activation domains from 45 proteins involved in transcriptional activation were curated from
750 UniProt³. We queried the UniProt database for human proteins whose regions, motifs or annotations
751 included the term “transcriptional activation.” We then filtered for ADs that ranged in length from 30 to 95
752 aa. For ADs shorter than 95 aa, we extended the protein sequence equally on either side until it reached
753 95 aa. The protein sequences were reverse translated and further divided into 95 aa sequences with 15
754 aa deletions positioned with a 2 aa sliding window. Duplicate sequences were removed, sequences were
755 codon matched for human codon usage, 7xC homopolymers were removed, BsmBI restriction sites were
756 removed, rare codons (less than 10% frequency) were avoided, and the GC content was constrained to
757 be between 20% and 75% in every 50 nucleotide window (performed with DNA chisel⁵⁴. 50 yeast Gcn4
758 controls were added, which included previously studied deletions²⁹. 2,024 library elements in total were
759 added to the 31,664 element TF tiling sub-library.

760

761 **CR tiling library design**

762 Candidate genes were initially chosen by including all members of the EpiFactors database, genes with
763 gene name prefixes that matched any genes in the EpiFactors database, and genes with any of the
764 following GO terms: GO:000785 (chromatin), GO:0035561 (regulation of chromatin binding),
765 GO:0016569 (covalent chromatin modification), GO:1902275 (regulation of chromatin organization),
766 GO:0003682 (chromatin binding), GO:0042393 (histone binding), GO:0016570 (histone modification),
767 and GO:0006304 (DNA modification). Genes present in prior Silencer tiling screens⁴ and genes present
768 in the present TF tiling screen were then filtered out. Biomart was used to identify and retrieve the
769 canonical transcript, and chosen by (in order of priority) the APPRIS principal transcript, the TSL principal
770 transcript, or the longest transcript with a protein coding CDS. Tiles for each of these DNA sequences
771 were generated using the same 80 amino acid tile/10 amino acid sliding window approach as the TF tiling
772 library. Duplicate sequences were removed, DNA hairpins and 7xC homopolymers were removed, and
773 sequences were codon matched for human codon usage with GC content being constrained to be
774 between 20% and 75% globally and between 25% and 65% in any 50-bp window. In order to improve
775 the coverage while performing the screen, this 51,297 element library was split into two sub-libraries: a
776 38,241 element CR Tiling Main sub-library and an 13,056 element CR Tiling Extended sub-library.
777 Computationally generated random negative controls, negative control tiles from the DMD protein
778 screened in prior Nuclear Pfam screens⁴, and fiduciary marker controls were added to each sub-library:
779 1,700 elements to the Main sub-library and 3,700 elements to the Extended sub-library. These controls
780 were not re-coded, and thus were repeated when pooling sub-libraries.

781

782 **Library filtering**

783 Since we pooled the sub-libraries and screened them as one large pool, several of the control sub-
784 libraries, that were not re-coded, wound up being repeated in the pool several times. We noticed
785 sequences that were repeated fewer times had enrichment scores closer to what was observed
786 previously. But sequences that were repeated upwards of five times had systematically lower enrichment
787 scores than what was expected from previous screens, likely due to PCR bias. We removed all repeated
788 control elements and instead relied on individual validations to confirm our screens worked. Additionally,
789 there was a computational error in removing BsmBI sites from the CR tiling library, resulting in some

790 sequences having accidental restriction cut sites in the middle of the ORF. We removed these sequences
791 from further analysis and supplementary tables.

792

793 **Activating hits validation library design**

794 1,055 putative hit tiles were chosen by selecting all tiles where both biological replicates were recovered
795 and had activation enrichment scores above 5.365 (determined by 2 standard deviations above random
796 controls). We included 200 randomly selected random negative controls that were poorly expressed
797 (expression threshold = -1.427) and 100 randomly selected non-hit tiles that had no activity in both the
798 minCMV and the pEF CRTF tiling screens. There were 1,355 total library elements.

799

800 **Repressing hits validation library design**

801 9,438 putative hit tiles were chosen by selecting all tiles where both biological replicates were recovered
802 and had pEF repression enrichment scores above 1.433 or had a PGK repression enrichment scores
803 above 0.880 (determined from 3 standard deviations above random controls). We included 500 randomly
804 selected random negative controls that were poorly expressed (expression threshold = -1.427) and 100
805 randomly selected non-hit tiles that had no activity in the minCMV, pEF nor PGK CRTF tiling screens.
806 There were 10,038 total library elements.

807

808 **AD mutants library design**

809 We defined compositional bias as any residue that represented more than 15% of the sequence (more
810 than 12 residues). We took 424 compositionally biased tiles and replaced all residues with alanine. We
811 took 1055 aromatic or leucine-containing tiles and replaced all Ws, Fs, Ys, and Ls with alanine. We took
812 1052 acidic residue-containing tiles and replaced all Ds and Es with alanine. 51 tiles that contained the
813 “LxxLL” motif (ELM accession: ELME000045, regex pattern = [^P]L[^P][^P]LL[^P]) we replaced with
814 alanine. 22 tiles that contained the “WW” motif (ELM accession: ELME000003, regex pattern = PP.Y) we
815 replaced with alanine. 8205 deletions were designed by systematically removing 10 aa chunks, with a
816 sliding window of 5 aa from 547 max activating tiles. All mutated sequences were reverse translated into
817 DNA sequences using a probabilistic codon optimization algorithm, such that each DNA sequence
818 contains some variation beyond the substituted residues, which improves the ability to unambiguously
819 align sequencing reads to unique library members. The 1055 putative hit tiles were included as positive
820 controls (slightly more activating tiles than we report in the main text because these libraries were
821 designed before we screened the validation library). We included 500 randomly selected random
822 negative controls that were poorly expressed (expression threshold = -1.427). There were 12,364 total
823 library elements.

824

825 **RD mutants library design**

826 12,000 deletions were designed by systematically removing 10 aa chunks, with a sliding window of 5 aa
827 of the maximum tile from 800 putative RDs that were hits in both PGK and pEF CRTF tiling screens
828 (slightly more RDs than we report in the main text because these libraries were designed before we
829 screened the validation library). All mutated sequences were reverse translated into DNA using the
830 method described above. The 1,593 putative hit tiles were included as positive controls. We took 644
831 compositionally biased tiles and replaced all residues with alanine. We replaced with alanines all the
832 following motifs: 104 CtBP interaction motif containing tiles (ELM accession: ELME0000098); 18 HP1
833 interaction motif containing tiles (ELM accession: ELME000141); 9 “ARKS” motif containing tiles (ELM

834 accession: DRAFT - LIG_CHROMO); 180 SUMO interaction motif containing tiles (ELM accession:
835 ELME000335); and 7 WRPW motif containing tiles (ELM accession: ELME000104). We included 500
836 randomly selected random negative controls that were poorly expressed (expression threshold = -1.427).
837 There were 15,055 total library elements.

838

839 **Bifunctional deletion scan library design**

840 3,331 deletions were created by systematically removing 10 aa chunks, with a sliding window of 2 aa
841 from 96 bifunctional activating and repressing tiles. All mutated sequences were reverse translated into
842 DNA sequences using a method described above. We included the WT bifunctional tiles and 250
843 randomly selected random negative controls that were poorly expressed (expression threshold = -1.427).
844 There were 3,674 total library elements.

845

846 **Library cloning**

847 Oligonucleotides with lengths up to 300 nucleotides were synthesized as pooled libraries (Twist
848 Biosciences) and then PCR amplified. 6x 50 ul reactions were set up in a clean PCR hood to avoid
849 amplifying contaminating DNA. For each reaction, we used either 5 or 10 ng of template, 1 ul of each 10
850 mM primer, 1 ul of Herculase II polymerase (Agilent), 1 ul of DMSO, 1 ul of 10 mM dNTPs, and 10 ul of
851 5x Herculase buffer. The thermocycling protocol was 3 minutes at 98C, then cycles of 98C for 20 s, 61C
852 for 20 s, 72C for 30 s, and then a final step of 72C for 3 minutes. The default cycle number was 20x, and
853 this was optimized for each library to find the lowest cycle that resulted in a clean visible product for gel
854 extraction (in practice, 23 cycles was the maximum when small libraries were represented in large pools).
855 After PCR, the resulting dsDNA libraries were gel extracted by loading a 2% TAE gel, excising the band
856 at the expected length (around 300 bp), and using a QIAgen gel extraction kit. The libraries were cloned
857 into a lentiviral recruitment vector pJT126 (Addgene #161926) with 4-16x 10 ul Golden-Gate reactions
858 (75 ng of pre-digested and gel-extracted backbone plasmid, 5 ng of library (2:1 molar ratio of
859 insert:backbone), 2uL of 10x T4 Ligase Buffer, and 1uL of NEB Golden Gate Assembly Kit (BsmBI-V2))
860 with 65 cycles of digestion at 42C and ligation at 16C for 5 minutes each, followed by a final 5 minute
861 digestion at 42C and then 20 minutes of heat inactivation at 70C. The reactions were then pooled and
862 purified with MinElute columns (QIAgen), eluting in 6 ul of ddH₂O. 2 ul per tube was transformed into two
863 tubes of 50 ml of Endura electrocompetent cells (Lucigen, Cat#60242-2) following the manufacturer's
864 instructions. After recovery, the cells were plated on 1-8 large 10"x10" LB plates with carbenicillin. After
865 overnight growth in a warm room, the bacterial colonies were scraped into a collection bottle and plasmid
866 pools were extracted with a Hi-Speed Plasmid Maxiprep kit (QIAgen). 2-3 small plates were prepared in
867 parallel with diluted transformed cells in order to count colonies and confirm the transformation efficiency
868 was sufficient to maintain at least 20x library coverage. To determine the quality of the libraries, the
869 putative effector domains were amplified from the plasmid pool by PCR with primers with extensions that
870 include Illumina adapters and sequenced. The PCR and sequencing protocol were the same as
871 described below for sequencing from genomic DNA, except these PCRs use 10 ng of input DNA and 17
872 cycles. These sequencing datasets were analyzed as described below to determine the uniformity of
873 coverage and synthesis quality of the libraries. In addition, 20-30 colonies from the transformations were
874 Sanger sequenced (Quintara) to estimate the cloning efficiency and the proportion of empty backbone
875 plasmids in the pools.

876

877 **Pooled delivery of library in human cells using lentivirus**

878 Large scale lentivirus production and spinfection of K562 cells were performed as follows: To generate
879 sufficient lentivirus to infect the libraries into K562 cells, we plated HEK293T cells on 1-12 15-cm tissue
880 culture plates. On each plate, 8.8×10^6 HEK293T cells were plated in 30 mL of DMEM, grown overnight,
881 and then transfected with 8 ug of an equimolar mixture of the three third-generation packaging plasmids
882 (pMD2.G, psPAX2, pMDLg/pRRE) and 8 ug of rTetR-domain library vectors using 50 mL of
883 polyethylenimine (PEI, Polysciences #23966). pMD2.G (Addgene plasmid #12259;
884 <http://addgene.org/12259>), psPAX2 (Addgene plasmid #12260; <http://addgene.org/12260>), and
885 pMDLg/pRRE (Addgene plasmid #12251; <http://addgene.org/12251>) were gifts from Didier Trono. After
886 48 hours and 72 hours of incubation, lentivirus was harvested. We filtered the pooled lentivirus through
887 a 0.45-mm PVDF filter (Millipore) to remove any cellular debris. K562 reporter cells were infected with
888 the lentiviral library by spinfection for 2 hours, with two separate biological replicates infected. Infected
889 cells grew for 2 days and then the cells were selected with blasticidin (10 mg/mL, Gibco). Infection and
890 selection efficiency were monitored each day using flow cytometry to measure mCherry (Biorad ZE5).
891 Cells were maintained in spinner flasks in log growth conditions each day by diluting cell concentrations
892 back to a 5×10^5 cells/mL. We aimed for 600x infection coverage and our lowest infection coverage was
893 130x. We aimed to have 2-10,000x maintenance coverage. On day 8 post-infection, recruitment was
894 induced by treating the cells with 1000 ng/ml doxycycline (Fisher Scientific) for either 2 days for activation
895 or 5 days for repression.

896

897 **Magnetic separation**

898 At each time point, cells were spun down at $300 \times g$ for 5 minutes and media was aspirated. Cells were
899 then resuspended in the same volume of PBS (GIBCO) and the spin down and aspiration was repeated,
900 to wash the cells and remove any IgG from serum. Dynabeads M-280 Protein G (ThermoFisher, 10003D)
901 were resuspended by vortexing for 30 s. 50 mL of blocking buffer was prepared per 2×10^8 cells by
902 adding 1 g of biotin-free BSA (Sigma Aldrich) and 200 mL of 0.5 M pH 8.0 EDTA into DPBS (GIBCO),
903 vacuum filtering with a 0.22-mm filter (Millipore), and then kept on ice. For all activation screens, 30 uL
904 of beads was prepared for every 1×10^7 cells, 60 uL of beads/10 million cells for the pEF CRTF tiles,
905 PGK CRTF tiles, and minCMV bifunctional deletion scan screens, 120 uL of beads/10 million cells for the
906 pEF validation, 90 uL of beads/10 million cells for the RD Mutants and pEF bifunctional deletion scan
907 screens. Magnetic separation was performed as previously described in ref⁴.

908

909 **FLAG staining for protein expression**

910 The expression level measurements for the CRTF tiling library were made in K562 minCMV cells (with
911 citrine OFF). 4×10^8 cells per biological replicate were used after 7 days of blasticidin selection (10
912 mg/mL, Gibco), which was 9 days post-infection. 4×10^7 control K562-JT039 cells (citrine ON, no lentiviral
913 infection) were spiked into each replicate. Fix Buffer I (BD Biosciences, BDB557870) was preheated to
914 37C for 15 minutes and Permeabilization Buffer III (BD Biosciences, BDB558050) and PBS (GIBCO) with
915 10% FBS (Hyclone) were chilled on ice. The library of cells expressing domains was collected and cell
916 density was counted by flow cytometry (Biorad ZE5). To fix, cells were resuspended in a volume of Fix
917 Buffer I (BD Biosciences, BDB557870) corresponding to pellet volume, with 20 mL per 1 million cells, at
918 37C for 10 - 15 minutes. Cells were washed with 1 mL of cold PBS containing 10% FBS, spun down at
919 500 3 g for 5 minutes and then supernatant was aspirated. Cells were permeabilized for 30 minutes on
920 ice using cold BD Permeabilization Buffer III (BD Biosciences, BDB558050), with 20 mL per 1 million

921 cells, which was added slowly and mixed by vortexing. Cells were then washed twice in 1 mL PBS+10%
922 FBS, as before, and then supernatant was aspirated. Antibody staining was performed for 1 hour at room
923 temperature, protected from light, using 5 μ L / 1×10^6 cells of α -FLAG-Alexa647 (RNDsystems, IC8529R).
924 We then washed the cells and resuspended them at a concentration of 3×10^7 cells / ml in PBS+10%FBS.
925 Cells were sorted into two bins based on the level of APC-A and mCherry fluorescence (Sony SH800S)
926 after gating for viable cells. A small number of unstained control cells was also analyzed on the sorter to
927 confirm staining was above background. The spike-in citrine positive cells were used to measure the
928 background level of staining in cells known to lack the 3XFLAG tag, and the gate for sorting was drawn
929 above that level. After sorting, the cellular coverage was $\sim 2000\times$. The sorted cells were spun down at
930 $500 \times g$ for 5 minutes and then resuspended in PBS. Genomic DNA extraction was performed following
931 the manufacturer's instructions (QIAGEN Blood Midi kit was used for samples with $> 1 \times 10^7$ cells) with
932 one modification: the Proteinase K + AL buffer incubation was performed overnight at 56C.

933

934 **Library preparation and sequencing**

935 Genomic DNA was extracted with the QIAGEN Blood Maxi Kit following the manufacturer's instructions
936 with up to 1×10^8 cells per column. DNA was eluted in EB and not AE to avoid subsequent PCR inhibition.
937 The domain sequences were amplified by PCR with primers containing Illumina adapters as extensions.
938 A test PCR was performed using 5 μ g of genomic DNA in a 50 μ L (half- size) reaction to verify if the PCR
939 conditions would result in a visible band at the expected size for each sample. Then, 3 - 48x 100 μ L
940 reactions were set up on ice (in a clean PCR hood to avoid amplifying contaminating DNA), with the
941 number of reactions depending on the amount of genomic DNA available in each experiment. 10 μ g of
942 genomic DNA, 0.5 μ L of each 100 mM primer, and 50 μ L of NEBnext Ultra 2x Master Mix (NEB) was
943 used in each reaction. The thermocycling protocol was to preheat the thermocycler to 98C, then add
944 samples for 3 minutes at 98C, then an optimized number of cycles of 98C for 10 s, 63C for 30 s, 72C for
945 30 s, and then a final step of 72C for 2 minutes. All subsequent steps were performed outside the PCR
946 hood. The PCR reactions were pooled and 145 μ L were run on a 2% TAE gel, the library band around
947 395 bp was cut out, and DNA was purified using the QIAquick Gel Extraction kit (QIAGEN) with a 30 μ L
948 elution into non-stick tubes (Ambion). A confirmatory gel was run to verify that small products were
949 removed. These libraries were then quantified with a Qubit HS kit (Thermo Fisher) and sequenced on an
950 Illumina HiSeq (2x150).

951

952 **Computing enrichments and hits thresholds**

953 Sequencing reads were demultiplexed using bcl2fastq (Illumina). A Bowtie reference was generated
954 using the designed library sequences with the script 'makeIndices.py' (HT-Recruit Analyze package) and
955 reads were aligned with 0 mismatch allowance using the script 'makeCounts.py'. The enrichments for
956 each domain between OFF and ON (or FLAGhigh and FLAGlow) samples were computed using the
957 script 'makeRhos.py'. Domains with < 5 reads in both samples for a given replicate were dropped from
958 that replicate (assigned 0 counts), whereas domains with < 5 reads in one sample would have those
959 reads adjusted to 5 in order to avoid the inflation of enrichment values from low depth.

960 For all of the screens, domains with < 20 counts in both conditions of a given replicate were filtered
961 out of downstream analysis. For the expression screens, well-expressed tiles were those with a
962 $\log_2(\text{FLAGhigh}:\text{FLAGlow})$ 1 standard deviation above the median of the random controls. For the CRTF
963 tiling repressor screens, hits were tiles with enrichment scores 3 standard deviations above the mean of
964 the poorly expressed random controls. For the minCMV CRTF tiling, pEF Bifunctional deletion scan, and

965 minCMV bifunctional deletion scan screens, hits were proteins with enrichment scores 2 standard
966 deviations above the mean of the poorly expressed random controls. For the validation and mutant
967 screens, hits were proteins with enrichment scores 1 standard deviation above the mean of the poorly
968 expressed random controls.

969

970 **Annotation of domains from tiles**

971 Tiles must have been hits in both the CRTF tiling and validation screens in order to have been considered
972 potential effector domains. A domain started anywhere the previous tile was not a hit. If the previous tile was
973 not a hit because it was not expressed, and if the antepenultimate (previous, previous) tile was a hit, then that
974 tile was not considered the start, and it was recovered into the middle of the domain. A domain ended
975 anywhere the next successive tile was not a hit. If the next tile was not a hit because it was not expressed,
976 and the following tile was a hit, then that tile was not considered the end, and it was recovered into the middle
977 of the domain. Domains started at the first residue of the first tile and extended until the last residue of the last
978 tile within the domain.

979

980 **Individual recruitment assays**

981 Protein fragments were cloned as a fusion with rTetR upstream of a T2A-mCherry-BSD marker, using
982 GoldenGate cloning in the backbone pJT126 (Addgene #161926). K562 citrine reporter cells were then
983 transduced with each lentiviral vector and, 3 days later, selected with blasticidin (10 mg/mL) until > 80%
984 of the cells were mCherry positive (6-9 days). Cells were split into separate wells of a 24-well plate and
985 either treated with doxycycline (Fisher Scientific) or left untreated. Time points were measured by flow
986 cytometry analysis of >10,000 cells (Biorad ZE5). Doxycycline was assumed to be degraded each day,
987 so fresh doxycycline media was added each day of the timecourse.

988

989 **Western blots**

990 5-10 million cells were lysed in lysis buffer (1% Triton X-100, 150mM NaCl, 50mM Tris pH 7.5, Protease
991 inhibitor cocktail). Protein amounts were quantified using the Pierce BCA Protein Assay kit (Bio-Rad).
992 Equal amounts were loaded onto a gel and transferred to a PVDF membrane. Membrane was probed
993 using FLAG M2 monoclonal antibody (1:1000, mouse, Sigma-Aldrich, F1804) and Histone 3 antibody
994 (1:1000, mouse, Abcam, AB1791) as primary antibodies. Goat anti-mouse IRDye 680 RD and goat anti-
995 rabbit IRDye 800CW (1:20,000 dilution, LICOR Biosciences, cat nos. 926-68070 and 926-32211,
996 respectively) were used as secondary antibodies. Blots were imaged on an iBright (Thermo Scientific).
997 Band intensities were quantified using ImageJ.

998

999 **Data analysis and statistics**

1000 All statistical analyses and graphical displays were performed in Python⁵⁵ (v. 3.8.5). Enrichment scores
1001 shown in all figures (aside from replicate plots) are the average across two separately transduced
1002 biological replicates. The p-values, statistical tests used, and n are indicated in the figure legends.

1003

1004 **Flow cytometry analysis**

1005 Data were analyzed using Cytoflow (<https://github.com/bpteague/cytoflow>) and custom Python scripts.
1006 Events were gated for viability and mCherry as a delivery marker. To compute a fraction of ON cells
1007 during doxycycline treatment, we fit a Gaussian model to the untreated rTetR-only negative control cells
1008 which fits the OFF peak, and then set a threshold that was 2 standard deviations above the mean of the

1009 OFF peak in order to label cells that have activated as ON. We do the same for computing the fraction of
1010 OFF cells in repressor validations but fit a two component Gaussian and set a threshold that was 2
1011 standard deviations below the mean of the ON peak. A logistic model, including a scale parameter, was
1012 fit to the validation and screen data using SciPy's curve fit function.

1013

1014 **Data availability**

1015 All raw NGS data and associated processed data generated in this study will be deposited in the NCBI
1016 GEO database upon publication.

1017

1018 **Code availability**

1019 The HT-recruit Analyze software for processing high-throughput recruitment assay and high-throughput
1020 protein expression assays are available on GitHub
1021 (<https://github.com/bintulab/HT-recruit-Analyze>).

1022 All custom codes used for data processing and computational analyses are available from the authors
1023 upon request.

1024

1025 **Biological materials availability**

1026 Oligonucleotide libraries are available upon request.

1027

1028 **Acknowledgements**

1029 We thank Michaela Hinks and members of our laboratories for helpful conversations and assistance. This
1030 work was supported by NIH-NIGMS R35M128947 (L.B.), NSF GRFP DGE-1656518 (N.D.), NIH-NIDDK
1031 F99/K00 F99DK126120 (J.T.), Stanford Bio-X Bowes Fellowship (P.S.), Stanford School of Medicine
1032 Dean's Fund, (C.A.), NIH-NIGMS 5T32GM007365-45 (A.M.), Stanford Interdisciplinary Graduate
1033 Fellowship affiliated with Stanford Bio-X (A.M.), NIH Director's New Innovator Award (1DP2HD08406901)
1034 (M.B.), and the BWF-CASI Award (L.B.)

1035

1036 **Authorship contributions**

1037 N.D. and L.B. designed the study, with significant intellectual contributions from P.S. and A.M. P.S. and
1038 N.D. designed the TF tiling libraries, A.M. designed the CR tiling libraries, both with contributions from J.T.,
1039 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.
1040 N.D. screened the CRTF minCMV and FLAG libraries with assistance from P.S. and J.T., Aradhana, K.S.
1041 screened the CRTF pEF and PGK promoter libraries. N.D. performed all other screens. N.D. analyzed
1042 the data, with assistance from L.B., I.L., C.A., and N.D. performed individual recruitment assay
1043 experiments. N.D. performed Western blot experiments. C.L. generated the PGK cell line. N.D. and L.B.
1044 wrote the manuscript, with significant contributions from J.T. and C.L., along with contributions from all
1045 authors. P.F., M.C.B. and L.B. supervised the project.

1046

1047 **Competing interests**

1048 Stanford has filed a provisional patent related to this work.

1049 **References**

- 1050 1. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **175**, 598–599 (2018).
- 1051 2. Medvedeva, Y. A. *et al.* EpiFactors: a comprehensive database of human epigenetic
- 1052 factors and complexes. *Database* **2015**, bav067 (2015).
- 1053 3. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic*
- 1054 *Acids Res.* **49**, D480–D489 (2021).
- 1055 4. Tycko, J. *et al.* High-Throughput Discovery and Characterization of Human
- 1056 Transcriptional Effectors. *Cell* **183**, 2020-2035.e16 (2020).
- 1057 5. Berger, S. L. The complex language of chromatin regulation during transcription.
- 1058 *Nature* **447**, 407–412 (2007).
- 1059 6. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints.
- 1060 *Nature* **583**, 729–736 (2020).
- 1061 7. Partridge, E. C. *et al.* Occupancy maps of 208 chromatin-associated proteins in one
- 1062 human cell type. *Nature* **583**, 720–728 (2020).
- 1063 8. Soto, L. F. *et al.* Compendium of human transcription factor effector domains. *Mol.*
- 1064 *Cell* **82**, 514–526 (2022).
- 1065 9. Liu, J. *et al.* Intrinsic Disorder in Transcription Factors. *Biochemistry* **45**, 6873–6888
- 1066 (2006).
- 1067 10. Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by
- 1068 sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).
- 1069 11. Sanborn, A. L. *et al.* Simple biochemical features underlie transcriptional activation
- 1070 domain diversity and dynamic, fuzzy binding to Mediator. *Elife* **10**, (2021).
- 1071 12. Erijman, A. *et al.* A High-Throughput Screen for Transcription Activation Domains
- 1072 Reveals Their Sequence Features and Permits Prediction by Deep Learning. *Mol.*
- 1073 *Cell* **78**, 890-902.e6 (2020).
- 1074 13. Ravarani, C. N. *et al.* High-throughput discovery of functional disordered regions:

- 1075 investigation of transactivation domains. *Mol. Syst. Biol.* **14**, e8190 (2018).
- 1076 14. Staller, M. V. *et al.* A High-Throughput Mutational Scan of an Intrinsically Disordered
1077 Acidic Transcriptional Activation Domain. *Cell Syst* **6**, 444-455.e6 (2018).
- 1078 15. Staller, M. V. *et al.* Directed mutational scanning reveals a balance between acidic
1079 and hydrophobic residues in strong human activation domains. *Cell Syst* **13**, 334-
1080 345.e5 (2022).
- 1081 16. Courey, A. J., Holtzman, D. A., Jackson, S. P. & Tjian, R. Synergistic activation by
1082 the glutamine-rich domains of human transcription factor Sp1. *Cell* **59**, 827–836
1083 (1989).
- 1084 17. Mermod, N., O'Neill, E. A., Kelly, T. J. & Tjian, R. The proline-rich transcriptional
1085 activator of CTF/NF-I is distinct from the replication and DNA binding domain. *Cell*
1086 **58**, 741–753 (1989).
- 1087 18. Paulsen, R. E., Weaver, C. A., Fahrner, T. J. & Milbrandt, J. Domains regulating
1088 transcriptional activity of the inducible orphan receptor NGFI-B. *J. Biol. Chem.* **267**,
1089 16491–16496 (1992).
- 1090 19. Meijer, D., Graus, A. & Grosveld, G. Mapping the transactivation domain of the Oct-
1091 6 POU transcription factor. *Nucleic Acids Res.* **20**, 2241–2247 (1992).
- 1092 20. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Development* **144**,
1093 2719–2729 (2017).
- 1094 21. Alerasool, N., Segal, D., Lee, H. & Taipale, M. An efficient KRAB domain for
1095 CRISPRi applications in human cells. *Nat. Methods* **17**, 1093–1096 (2020).
- 1096 22. Keung, A. J., Bashor, C. J., Kiriakov, S., Collins, J. J. & Khalil, A. S. Using targeted
1097 chromatin regulators to engineer combinatorial and spatial transcriptional regulation.
1098 *Cell* **158**, 110–120 (2014).
- 1099 23. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-
1100 dependent regulatory functions. *Nature* **528**, 147–151 (2015).

- 1101 24. Arnold, C. D. *et al.* A high-throughput method to identify trans-activation domains
1102 within transcription factor sequences. *EMBO J.* **37**, (2018).
- 1103 25. Haberle, V. *et al.* Transcriptional cofactors display specificity for distinct types of
1104 core promoters. *Nature* **570**, 122–126 (2019).
- 1105 26. Alerasool, N., Leng, H., Lin, Z.-Y., Gingras, A.-C. & Taipale, M. Identification and
1106 functional characterization of transcriptional activators in human cells. *Mol. Cell* **82**,
1107 677-695.e7 (2022).
- 1108 27. Ptashne, M. How eukaryotic transcriptional activators work. *Nature* **335**, 683–689
1109 (1988).
- 1110 28. Gerber, H. P. *et al.* Transcriptional activation modulated by homopolymeric
1111 glutamine and proline stretches. *Science* **263**, 808–811 (1994).
- 1112 29. Tuttle, L. M. *et al.* Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic
1113 Fuzzy Protein-Protein Complex. *Cell Rep.* **22**, 3251–3264 (2018).
- 1114 30. Ross, S., Best, J. L., Zon, L. I. & Gill, G. SUMO-1 modification represses Sp3
1115 transcriptional activation and modulates its subnuclear localization. *Mol. Cell* **10**,
1116 831–842 (2002).
- 1117 31. Rocca, D. L., Wilkinson, K. A. & Henley, J. M. SUMOylation of FOXP1 regulates
1118 transcriptional repression via CtBP1 to drive dendritic morphogenesis. *Sci. Rep.* **7**,
1119 877 (2017).
- 1120 32. Verger, A., Perdomo, J. & Crossley, M. Modification with SUMO. A role in
1121 transcriptional regulation. *EMBO Rep.* **4**, 137–142 (2003).
- 1122 33. Torres-Machorro, A. L. Homodimeric and Heterodimeric Interactions among
1123 Vertebrate Basic Helix-Loop-Helix Transcription Factors. *Int. J. Mol. Sci.* **22**, (2021).
- 1124 34. Andrés, M. E. *et al.* CoREST: a functional corepressor required for regulation of
1125 neural-specific gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9873–9878
1126 (1999).

- 1127 35. Brayer, K. J. & Segal, D. J. Keep your fingers off my DNA: protein-protein
1128 interactions mediated by C2H2 zinc finger domains. *Cell Biochem. Biophys.* **50**,
1129 111–131 (2008).
- 1130 36. Koipally, J. & Georgopoulos, K. A molecular dissection of the repression circuitry of
1131 Ikaros. *J. Biol. Chem.* **277**, 27697–27705 (2002).
- 1132 37. McCarty, A. S., Kleiger, G., Eisenberg, D. & Smale, S. T. Selective dimerization of a
1133 C2H2 zinc finger subfamily. *Mol. Cell* **11**, 459–470 (2003).
- 1134 38. Boyle, P. & Després, C. Dual-function transcription factors and their entourage:
1135 unique and unifying themes governing two pathogenesis-related genes. *Plant*
1136 *Signal. Behav.* **5**, 629–634 (2010).
- 1137 39. Latchman, D. S. Eukaryotic transcription factors. *Biochem. J* **270**, 281–289 (1990).
- 1138 40. Dyson, H. J. & Wright, P. E. Role of Intrinsic Protein Disorder in the Function and
1139 Interactions of the Transcriptional Coactivators CREB-binding Protein (CBP) and
1140 p300. *J. Biol. Chem.* **291**, 6714–6722 (2016).
- 1141 41. Gillespie, M. A. *et al.* Absolute quantification of transcription factors in human
1142 erythropoiesis using selected reaction monitoring mass spectrometry. *STAR*
1143 *Protocols* **1**, 100216 (2020).
- 1144 42. Willy, P. J., Kobayashi, R. & Kadonaga, J. T. A basal transcription factor that
1145 activates or represses transcription. *Science* **290**, 982–985 (2000).
- 1146 43. Majello, B., De Luca, P. & Lania, L. Sp3 is a bifunctional transcription regulator with
1147 modular independent activation and repression domains. *J. Biol. Chem.* **272**, 4021–
1148 4026 (1997).
- 1149 44. Ma, J. Crossing the line between activation and repression. *Trends Genet.* **21**, 54–
1150 59 (2005).
- 1151 45. Loh, Y.-H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in
1152 mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).

- 1153 46. Heurtier, V. *et al.* The molecular logic of Nanog-induced self-renewal in mouse
1154 embryonic stem cells. *Nat. Commun.* **10**, 1109 (2019).
- 1155 47. Mann, R. S., Lelli, K. M. & Joshi, R. Hox specificity unique roles for cofactors and
1156 collaborators. *Curr. Top. Dev. Biol.* **88**, 63–101 (2009).
- 1157 48. Bürglin, T. R. & Affolter, M. Homeodomain proteins: an update. *Chromosoma* **125**,
1158 497–521 (2016).
- 1159 49. White, M. A. *et al.* A Simple Grammar Defines Activating and Repressing cis-
1160 Regulatory Elements in Photoreceptors. *Cell Rep.* **17**, 1247–1254 (2016).
- 1161 50. Friedman, R. Z. *et al.* Information content differentiates enhancers from silencers in
1162 mouse photoreceptors. *Elife* **10**, (2021).
- 1163 51. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1–13
1164 (2018).
- 1165 52. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V.
1166 CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically
1167 Disordered Proteins. *Biophys. J.* **112**, 16–21 (2017).
- 1168 53. Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative splice
1169 isoforms. *Nucleic Acids Res.* **41**, D110-7 (2013).
- 1170 54. Zulkower, V. & Rosser, S. DNA Chisel, a versatile sequence optimizer.
1171 *Bioinformatics* **36**, 4508–4509 (2020).
- 1172 55. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, 2009).