A proteome-wide dependency map of protein interaction motifs

Sara M. Ambjørn[1,2,7], Bob Meeusen[1,2,7], Johanna Kliche[1,2], Juanjuan Wang[2,4], Dimitriya H. Garvanska[1,2], Thomas Kruse[1,2], Blanca Lopez Mendez[2], Matthias Mann[2], Niels Mailand[1,2], Emil P.T. Hertz[2,5,8], Norman E. Davey[3,6,8] , Jakob Nilsson[1,2,8]

[1] Center for Epigenetic Cell Memory, Danish Cancer Institute, Copenhagen, Denmark.

[2] Novo Nordisk Foundation Center for Protein Research, ICMM, University of Copenhagen, Copenhagen, Denmark.

[3] Division of Cancer Biology, The Institute of Cancer Research, 237 Fulham Road, SW3 6JB, London, UK

[4] Present address: Department of Clinical Biochemistry, Copenhagen University Hospital Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark

[5] Present address: Skape Bio ApS, Copenhagen, Denmark

[6] Department of Cell and Molecular Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, US

[7] Equal contribution

[8] Correspondence: Jaknil@cancer.dk, norman.davey@stjude.org, emil.hertz@cpr.ku.dk

Abstract

Short linear motifs (SLiMs) are the most ubiquitous protein interaction modules within unstructured regions of the human proteome, yet their contribution to cellular homeostasis remains poorly understood. To systematically assess SLiM function, we applied base editing to mutate all reported and a set of computationally predicted SLiMs defined by SLiM-like evolutionary patterns. By screening 7,293 SLiM-containing regions with 80,473 mutations in HAP1 cells, we define a SLiM dependency map identifying 450 known and 264 predicted SLiMs required for normal cell proliferation. Mutational consequences were highly reproducible in RPE1 cells, with differences attributed to cell line-specific gene essentiality. We show that most essential predicted SLiMs belong to novel classes and identify binding partners for several of these, providing mechanistic insight into a disease-associated ANKRD17 mutation. Our study provides a proteome-wide resource on SLiM essentiality and uncovers numerous uncharacterized yet essential SLiMs.

Introduction

The molecular details of how specific protein complexes contribute to cellular functions is a fundamental question in cell biology and forms the basis for interpreting disease causing mutations. Short linear motifs (SLiMs; also referred to as molecular recognition features (MoRFs), molecular recognition elements (MREs), or eukaryotic linear motifs (ELMs)) are an important and undercharacterized class of protein interaction modules, likely constituting the largest uncharted part of the human interactome[1-3]. SLiMs reside in intrinsically disordered regions (IDRs), protein segments that do not adopt a stable three-dimensional structure, though they often form stable secondary structure in their bound state. They are typically encoded in short stretches of ~10 amino acids where 2-4 residues serve as key binding determinants, directly interacting with binding partners, which are most commonly globular protein domains (Fig. 1a-b), but can also include DNA, RNA, and lipid membranes[4-6]. Due to the compact footprint of these interfaces, SLiMs generally bind with low micromolar affinity and mediate dynamic, conditional, yet specific protein interactions[7]. The simple binding determinants of SLiMs also facilitate the acquisition of motifs by *ex nihilo* evolution through random mutation, resulting in multiple proteins convergently evolving a similar SLiM that binds the same motif-binding partner[8]. Likewise, SLiM function can readily be perturbed by single amino acid mutations and this can contribute to human diseases[2,9,10]. The unique properties of SLiM-mediated interactions are widely utilized by the cell, and SLiMs play a wide range of roles to support key cellular functions[2], including driving the assembly of dynamic complexes, recruiting modifying enzymes, controlling protein stability, and encoding subcellular localisation[5].

Decades of detailed research has uncovered thousands of human SLiMs, with the potential for several thousand more to be discovered[1]. The recent development and application of a range of novel high-throughput methods to identify peptides binding to a globular domain has increased the rate of SLiM discovery[3]. Although detailed mechanistic and functional studies have been conducted for a small subset of the known SLiMs, most have not been functionally characterized. Functional analysis remains a time-consuming bottleneck, and a significant portion of SLiMs are validated solely by *in vitro* experiments, biophysical measurements or structural studies. Functional characterization experiments often solely rely on exogenous expression of proteins, which does not address function at the endogenous protein level. Since the affinities of SLiMs are finely tuned for proper biological function, it is crucial that SLiMs are disrupted in the endogenous proteins to maintain physiological concentrations and

70 spatiotemporal expression[7,11]. Thus, most SLiMs are incompletely characterized, leaving a gap
71 in our understanding of SLiM contributions to cell fitness. Approaches to functionally
72 characterize SLiMs at scale would help in the interpretation of the large amount of interaction
73 and disease mutation data, and pinpoint key interactions required for cellular functions. Here,
74 we utilize a base editor screening approach to perform large-scale, programmable mutagenesis
75 of all reported and thousands of computationally predicted SLiMs in their endogenous cellular
76 context to investigate their impact on cellular proliferation.

77 **Results**

78 **Design of a base editor screen targeting SLiMs in the human proteome**

79 As our target set, we generated a list of 7,293 SLiM-containing regions in 4,430 proteins, where
80 each region contains one or more overlapping motifs (Fig. 1c, Table S1). This dataset contains
81 all reported SLiMs (the *reported motif set*) and an extensive set of predicted SLiMs (the
82 *predicted motif set*). The *reported motif set* consists of 3,562 motif-containing regions
83 comprising a comprehensive set of 2,750 curated motifs from the experimental literature
84 (compiled in the Eukaryotic Linear Motif resource (ELM)[5] and from manual curation) and 812
85 motif instances extracted from PDB structures. The *predicted motif set* contains 3,731 protein
86 regions with SLiM-like evolutionary patterns predicted by screening for highly conserved
87 groups of amino acids in rapidly evolving disordered regions (see methods section).
88
89 We reasoned that base editors would be ideal to precisely disrupt SLiM-mediated interactions
90 by introducing point mutations in key residues required for binding, allowing the functional
91 interrogation of SLiMs at scale and at endogenous protein levels. Base editors are composed
92 of a catalytically impaired Cas protein (typically Cas9 nickase) fused to either a cytosine or
93 adenine deaminase. The gRNA RNA (gRNA) contains a 20 nucleotide RNA sequence which
94 gRNAs the Cas enzyme to the complimentary location in the genome and displaces the
95 opposite DNA strand, which is then deaminated by the deaminase. In this manner, followed by
96 replication and repair, cytosine base editors (CBEs) convert C-G base pairs to T-A while
97 adenine base editors (ABEs) convert A-T base pairs to G-C[12] (Fig. 1d). Cas binding also
98 requires the presence of a protospacer adjacent motif (PAM), which for SpCas9 is an NGG
99 sequence located immediately downstream of the gRNA binding sequence. Editing takes place
100 within a specific nucleotide "editing window" at a defined distance from the PAM (Fig. 1e).
101 This allows for gRNA RNA (gRNA) directed mutations in the genome[12], with the amino acid
102 mutational outcome dictated by the underlying codon sequence. In this study, we selected the

103 ABE ABE8e[13] and the CBE BE3.9max[14], both of which target a 5-nucleotide editing window
104 spanning nucleotides 4-8 in the target DNA matching the gRNA (Fig. 1e). These were fused to
105 either SpCas9 nickase (D10A) or the engineered variant SpG nickase[15,16], which recognizes a
106 relaxed NGN PAM sequence, allowing for an expanded range of targetable residues in SLiMs.
107 We will henceforth refer to these base editors as: ABE_{NGG}, $CBE_{NGG,}$ ABE_{NGN}, and CBE_{NGN}.
108
109 A set of gRNAs was designed to target each SLiM with the four base editors. For the prediction
110 of mutations, we assumed full editing within the editing window and retained only gRNAs
111 predicted to result in non-synonymous mutations. The resulting gRNA library consisted of
112 90,213 gRNAs targeting 80,473 unique mutations to 56,083 residues in the 7,293 target SLiMs
113 (Fig. 1f, Table S2). The library also contained 62 previously described positive control gRNAs
114 targeting splice sites of essential genes and 255 negative controls including gRNAs targeting
115 intergenic regions of the genome and non-targeting gRNAs that do not map to the genome[16].
116 Only gRNAs targeting NGG PAMs were cloned into lentiviral vectors encoding SpCas9-base
117 editor fusions (ABE_{NGG} and $CBE_{NGG,}$) while all gRNAs targeting NGG or NGN PAMs were
118 cloned into lentiviral vectors encoding SpG-base editor fusions (ABE_{NGN} and CBE_{NGN}) to
119 produce four lentiviral plasmid libraries.
120
121 **Design and benchmarking of the SLiM dependency screen.**
122 HAP1, a near-haploid human cell line derived from a chronic myelogenous leukemia (CML)
123 cell line, had shown superior performance in previous base editor screens[16,17] and was chosen
124 as the cell model to interrogate SLiM dependency. HAP1 cells were transduced in biological
125 triplicates with each of the lentiviral libraries at low multiplicity of infection and high gRNA
126 coverage (>675-fold) (Fig. 1g). After transduction and selection, cells were grown
127 competitively for 21 days with passaging every three days. Genomic DNA was extracted at
128 day 0, 18, and 21 post selection, and gRNA abundances were quantified by next generation
129 sequencing (NGS) (Fig. 1g). The fold changes in gRNA abundances comparing early and late
130 time points were quantified along with the significance of that change using the limma gene
131 expression analysis package to identify gRNAs which were enriched, stable, or depleted over
132 the time course of the screen[18] (Table S3).
133
134 Positive and negative control gRNAs were examined to assess the discriminatory power of
135 each base editor to identify functionally relevant edits. We observed substantial depletion of
136 positive control gRNAs compared to negative control gRNAs (Fig. S1a) with good

137 discriminatory power for three out of four base editors (Fig. 1h, ABE_{NGN} AUC=0.88; ABE_{NGG}
138 AUC=0.92; CBE_{NGN} AUC=0.63; CBE_{NGG} AUC=0.94). The CBE_{NGN} base editor displayed a
139 notably weaker depletion of positive controls, consistent with observations by others[16], which
140 was also reflected in lower discriminatory power. Base editors targeting the NGG PAM
141 outperformed those targeting the NGN PAM (Fig. 1h). As the discriminatory power at day 18
142 was highly similar to that at day 21, we focused on day 18 in the following analyses (Fig. S1b).
143 A p-value cut-off of 0.01 showed a false positive rate of ~1% across all four screens, with a
144 true positive rate of up to 79% (Fig. 1i), and was set as the threshold for considering gRNAs to
145 significantly influence cell proliferation.
146
147 We next examined the effect of the SLiM-targeting gRNAs. The majority of these showed
148 minimal to no change in abundance during the screen; however, a small proportion exhibited
149 significant depletion or enrichment (Fig. 1i; ABE_{NGN} 3.5%, ABE_{NGG} 4.0%, CBE_{NGN} 1.5%, and
150 CBE_{NGG} 5.1%). We observed a strong correlation between replicates as well as between the
151 gRNAs that changed significantly in both NGN and NGG screens for the same editor (Fig.
152 S1c-h). As anticipated, given the different targeting and editing outcomes of the CBE and ABE,
153 we observed distinct gRNA changes for each editor, advocating the use of both for deeper
154 functional investigation (Fig. 1j).
155
156 We next confirmed that the base editors introduced the predicted mutations by focusing on a
157 SLiM from the *predicted motif set* in Cleavage and polyadenylation specificity factor subunit
158 1 (CPSF1) that was required for cell fitness. We deep sequenced the genomic locus after
159 transduction with 15 single gRNAs in either ABE_{NGN} or CBE_{NGN} base editors (Fig. S2). The
160 ABE_{NGN} gRNAs introduced mutations with high efficiency at day 0 (up to 61% allelic
161 frequency) with good overlap with the predicted edits and showed limited bystander edits (Fig.
162 S2, Table S4 and Supplemental data). The CBE_{NGN} gRNAs also introduced the predicted
163 mutations, though at lower efficiency than the ABE_{NGN}, displaying up to 10% indel generation
164 and evident bystander editing, which is in line with findings by others[14] (Fig. S2, Table S4 and
165 Supplemental data). We further observed that specific mutations introduced by gRNAs that
166 scored in the proliferation screen were depleted from the cell population from day 0 to day 18,
167 supporting that these mutations lead to loss of function that impacts cellular fitness (Fig. S2,
168 Table S4).

169 In summary, the screens had strong discriminatory power for gRNAs targeting regions of the
170 genome that are required for normal cell proliferation and showed significant depletion or
171 enrichment of a subset of the SLiM-targeting gRNAs.
172
173 **Defining a SLiM dependency map**
174 Next, we defined a SLiM dependency map by analyzing the changes in gRNA abundances in
175 a motif-centric manner. We reasoned that SLiMs targeted by multiple significant gRNAs are
176 more likely to be required for cellular fitness. In accordance with this, we observed that 62.3%
177 of the significantly changing gRNAs were found in only 11.5% of the SLiMs, suggesting that
178 a subset of SLiMs are affecting proliferation. We further filtered the predicted motifs using
179 AlphaFold structural information to more stringently enrich for motifs residing in intrinsically
180 unstructured regions[19]. Next, we categorized the remaining 6,489 SLiMs into three classes of
181 motifs: depleted motifs, which result in a proliferation defect upon mutation; enriched motifs,
182 which provide a proliferation advantage upon mutation; and stable motifs, which show no
183 observable effect upon mutation (examples in Fig. 2a-c). We applied three metrics to quantify
184 essentiality and prioritize motifs by confidence across the four screens: the number of unique
185 significant gRNAs per motif; the mean fold change of the gRNAs for each motif; and the
186 distribution of these fold changes against the distribution of all gRNAs in the screen (Table
187 S5). This data is made available through an interactive web resource for the community to
188 explore SLiM function (https://slim.icr.ac.uk/projects/base_editing) (Fig. S3).
189
190 Overall, across the four screens, we identified 714 SLiMs in 626 proteins targeted by at least
191 two unique gRNAs with significant depletion or enrichment (Fig. 2d). We consider these motifs
192 the hits of our screen and will refer to them as 'essential SLiMs', that is SLiMs which affect
193 normal cell proliferation when mutated. In contrast, the majority of motifs (4,362; 67.2%) were
194 targeted solely by gRNAs that remained stable through the screening (Fig. 2d). We consider
195 these to be non-essential for proliferation, although we cannot exclude the lack of phenotype
196 to be due to low editing efficiency (see discussion). Of the essential SLiMs, 565 were depleted
197 and 149 enriched upon mutation and 450 were from the *reported motif set* while 264 were from
198 the *predicted motif set* (Fig. 2e-f).
199
200 We next categorized the *reported motif set* into high-level functionalities (Fig. 2g) and
201 observed a small increase in essentiality for the binding (7.4%) and degradation (8.0%) classes
202 of motifs compared to the modification, localization, and docking motif classes (5.8%, 6.0%,

203 and 5.8%, respectively). We noted that many highly conserved SLiMs are not essential for cell
204 proliferation and, contrary to expectation, that essential SLiMs are not more highly conserved
205 than non-essential SLiMs (Mann-Whitney p-value 0.53, Fig. 2h, Table S5). Finally, as
206 expected, when comparing SLiM essentiality to gene essentiality defined by HAP1 gene
207 knockout screens[20], essential motifs are more likely to be encoded in essential genes (Mann-
208 Whitney p-value 1.69×10^{-27}, Fig. 2i). However, we did observe essential motifs in proteins
209 encoded by non-essential genes including the instance in Phosphatidate phosphatase LPIN1
210 (encoded by *LPIN1*) $WLWGELP_{273-279}$ and Tyrosine-protein phosphatase non-receptor type 1
211 (PTP1B encoded by *PTPN1*) $EVRSR_{369-373}$ investigated later.

212 Several of the SLiMs from the *reported motif set* with the highest number of significant gRNAs
213 in the screen are in proteins with clear links to core biological functions (Fig. 2j). These motifs
214 include the polyadenylate-binding protein-1 (PABP) binding motif in eIF4G required for the
215 expression of polyadenylated mRNAs[21] and the APC/C degradation motif (D box) of Cyclin
216 B1 (CCNB1) required for mitotic exit[22]. Mutations in motifs with anti-apoptotic functions in
217 many cases resulted in reduced cell proliferation, for example, mutations to motifs in Apoptosis
218 regulator Bcl-2 family member BCL2L1[23]. Conversely, mutation of the BH3 motif in
219 Apoptosis regulator BAX (BAX) and tetramerization motif in the Tumor protein p73 (TP73),
220 both of which have pro-apoptotic functions[24-26], resulted in a proliferative advantage. We
221 performed a Reactome pathway analysis of the proteins harboring the essential SLiMs of the
222 *reported motif set* compared to all proteins in the set. This showed an enrichment in essential
223 motifs in pathways related to core biological processes such as transcription, RNA processing,
224 cell cycle and apoptosis (Fig. S4a, Table S6). For the *predicted motif set*, functions were more
225 diverse with no enrichment in any Reactome pathway. Mapping essential SLiMs to reported
226 protein complexes retrieved from the Complex Portal[27] suggested that numerous key
227 complexes require multiple cooperative motif-mediated interactions for their function (Fig.
228 S4b-d, Table S7). Three complexes stood out due to the number of essential motifs with five
229 each: the Mitotic Checkpoint Complex (motifs in BUBR1 and CDC20) (Fig. S4b); Histone-
230 lysine N-methyltransferase complex SET1A variant (motifs in SETD1A, RBBP5 and ASH2L)
231 (Fig. S4c) and pre-mRNA cleavage and polyadenylation specificity factor complex (motifs in
232 CPSF1, FIP1L1 and WDR33) (Fig. S4d). The latter two are of particular interest as they include
233 several uncharacterized, yet essential motifs.

234 In summary, we used the base editing screens to define a dependency map of the SLiM-
235 mediated interactome, identifying reported and predicted SLiMs affecting cell proliferation.

236 **gRNA behavior and editing outcomes are maintained across cell lines**

237 To further confirm beyond the CPSF1 motif (Fig. S2) that introduction of the predicted edits
238 by the base editor cause the observed effects on cellular fitness, we focused on the 714 essential
239 motifs defined in the SLiM dependency map for validation. A major limitation of base editing
240 screening approaches is that base editing outcomes are not measured in the screen but requires
241 tedious post-screen validation experiments to confirm editing events (Fig. S2). During the
242 preparation of this work, several base editing activity 'sensors' were published[28-31] in which a
243 surrogate target site is linked to the gRNA construct in the lentiviral vector, allowing for
244 indirect measurement of base editing outcomes by NGS of the sensor (Fig. 3a). We thus
245 designed a SLiM validation base editing library using a gRNA-sensor set-up (Fig. 3a-b). We
246 focused on ABE_{NGN} due to its high performance with the relaxed NGN PAM requirement (Fig.
247 1h) and designed 8505 gRNA-sensor constructs targeting the 714 essential motifs defined in
248 the SLiM dependency map, including both significant and non-significant gRNAs from the
249 previous ABE_{NGN} screen (Table S8). In addition to positive and negative controls, eight gRNAs
250 per SLiM-containing gene were designed to target splice sites to generate knockout-like
251 phenotypes, allowing us to compare the effects of residue-level to gene level perturbations.
252

253 To evaluate the reproducibility of the HAP1 screen across cell types, the sensor screen was
254 performed both in HAP1 cells and in a non-transformed RPE1 p53-/- cell line, which we have
255 previously used for base editing screening[32]. Cells were transduced with the lentiviral library
256 at low multiplicity of infection and high gRNA coverage (>1000x) and allowed to grow
257 competitively for 18 days. gRNA abundance and sensor editing in genomic DNA at day 0, 3
258 and 18 post selection were quantified by NGS (Fig. 3b), and changes in gRNA abundances
259 from day 0 to day 18 were analyzed as for the primary screen (Table S9-S10). Editing outcome
260 and frequency on the sensor in the starting population was evaluated at day 3 as this timepoint
261 had higher editing frequencies than day 0 (Fig. 3c, Fig. S5a-b). Positive and negative controls
262 behaved as expected, and no threshold was set on minimal editing frequency of a gRNA since
263 it did not substantially improve the discriminatory power of the screen (Fig. S5c-d). First, we
264 compared the changes in gRNA abundance to the primary screen data with the same editor
265 (ABE_{NGN}) to confirm reproducibility across screens (Fig. S5e). The gRNA behaviors were
266 highly correlated although changes were less pronounced in the sensor screen.

267

268 A comparison of HAP1 and RPE1 cells revealed highly similar sensor editing frequencies
269 (Pearson correlation $r = 0.93$, $p < 0.001$), enabling a direct comparison of gRNA behavior (Fig.
270 3c). Similarly, the gRNA abundance changes were correlated (positive controls: $r = 0.73$, $p <$
271 0.001, SLiM-targeting: $r = 0.41$, $p < 0.001$), despite a large proportion of the motif-targeting
272 gRNAs showing no change, suggesting gRNA behavior is generally reproducible across cell
273 lines (Fig. 3d, Table S9). However, we also observed some notable differences between HAP1
274 and RPE1 cells including the enrichment of gRNAs targeting specific motifs (Fig. 3d). In RPE1
275 cells, gRNAs targeting the AMBRA1 degron of Cyclin D1 and the 14-3-3 binding site in WW
276 domain-containing transcription regulator protein 1 (WWTR1) are more highly enriched than
277 in HAP1 cells. Conversely, gRNAs targeting the Sin3 motif of the Max-binding protein MNT
278 (MNT) and the nuclear export signal of Kelch-like ECH-associated protein 1 (KEAP1) are
279 depleted in HAP1 but enriched in RPE1 (Fig. 3d, Fig. S5f). These differences between cell
280 lines likely reflect variation in the essentiality of the corresponding genes. As the library
281 included gRNAs targeting splice sites of the genes harboring the SLiMs, we were able to
282 compare the effect of SLiM mutation to gene essentiality from 'knockouts' created by base
283 edited splice sites (mean $\log_2$ fold change of eight splice sites), which correlates well with
284 DepMap[33] knockout gene effects (Fig. S5g-h). Strikingly, for Cyclin D1 and WWTR1,
285 mutation of the SLiM produces the opposite effect of the gene knockout, likely reflecting the
286 growth suppressing role of these SLiMs, where motif loss would result in stabilisation of Cyclin
287 D1[34-36] and activation of WWTR1[37]. The motif-targeting gRNAs for MNT and KEAP1 follow
288 the same pattern observed for their respective splice site mutations (Fig. 3d-e, Fig. S5f). Indeed,
289 there was a trend for the log fold changes of significant motifs to align with the splice site-
290 targeting gRNAs, confirming that motif essentiality tends to follow overall gene essentiality
291 (Fig. S5i-j, Table S10).
292

293 Next, we looked more closely into reported SLiMs where the conserved key binding
294 determinants (defined positions) are well-characterized to determine if mutations in these
295 positions were more likely to cause a fitness defect than non-conserved flanking positions
296 (wildcard positions) (Fig. 3f). We considered only the motifs which were significantly
297 changing (Mann-Whitney p-value < 0.01, Table S10) and further separated the gRNAs into the
298 50% lowest and 50% highest editing. We observed that gRNAs targeting defined positions are
299 more likely to cause a fitness effect than gRNAs targeting wildcard positions and that this effect
300 is more prominent with high editing gRNAs than low editing gRNAs (Fig. 3f). This is

301 consistent with the fitness effect being caused by disruption of SLiM-mediated binding through
302 mutation of defined key binding residues.
303
304 Next, we wanted to further validate our top hits by probing editing in their genomic loci and
305 track mutations during competitive growth. Here, we chose the top-scoring ABE_{NGN} gRNA for
306 the top 25 predicted motifs and cloned a lentiviral gRNA-sensor library, which were transduced
307 into HAP1 cells, followed by competitive growth and NGS of the sensor as described above
308 (Fig. 3g). In addition, we performed pooled amplification and NGS of the 25 genomic target
309 loci using rhAmpSeq^{TM} to obtain detailed information about editing at the genomic loci (Fig.
310 3g, Table S11, see also methods section). Inspecting the sensor and rhAmpSeq editing data, we
311 observed highly similar outcomes across predicted, sensor-derived, and genomic locus edits
312 (Fig. 3h, Fig. S6a-b). Exceptions included Eukaryotic peptide chain release factor GTP-binding
313 subunit ERF3A (encoded by *GSPT1*) where K87E rather than the predicted K87G was the
314 more predominant sensor and genomic locus edit and Kinase suppressor of Ras 1 (*KSR1*) where
315 predicted and sensor editing matched (S453P) but the most predominant genomic locus edit
316 was S453Y (Fig. S6a). In most cases, we observed that the relative frequency of the protein-
317 level mutations changed from early to late timepoints (Fig. 3i, Fig. S6a). For example,
318 ANKRD17 I2427V and EIF3G W12R were depleted over time, indicating that cells with these
319 mutations were outcompeted from the population, consistent with the depletion of the
320 respective gRNAs in the screens.
321
322 Together, these results demonstrate that i) the screen results were reproducible across two cell
323 lines, ii) differences between the cell lines can largely be explained by differences in gene
324 essentialities, iii) predicted, sensor, and genomic locus editing were highly similar, and iv)
325 depleted gRNAs from the screen introduce mutations in the genomic target loci which are
326 outcompeted over time.
327
328 **Base editing of a novel SLiM in CPSF1 links proliferation defects to loss of CSTF3**
329 **binding.**
330 Next, to validate that the introduction of the predicted edits by the base editor caused the
331 observed effects on cellular fitness through disruption of SLiM-mediated interactions, we first
332 focused on one of the top scoring SLiM from the *predicted motif set*, $DDEEEMLYG_{741-749}$ in
333 CPSF1 (Fig. 4a, Table S5). This SLiM shares significant similarity with the Cleavage
334 stimulation factor subunit 3 (CSTF3)-binding motif in FIP1L1[38], that is also a strongly depleted

335 motif in our screen, particularly at core LYG and N-terminal acidic residues (Fig. 4a). We will
336 refer to it as the LYG motif from here on. First, to explore whether mutations in this motif
337 would lead to loss of binding partners, we transiently transfected HeLa cells with YFP-tagged
338 full length CPSF1 wild-type or a mutant with alanine substitutions in the two amino acids that
339 were mutated in the most depleted gRNAs targeting the motif, E743 and E744 (Table S3). The
340 constructs were affinity purified (AP) from HeLa cells and a quantitative interactome
341 comparison was performed using mass spectrometry (MS)-based proteomics (we will refer to
342 this approach as AP-MS). This revealed that the CPSF1 LYG motif is required for binding to
343 Cleavage Stimulation Factor complex subunits (CSTF1 and CSTF3), confirming that the LYG
344 motif in CPSF1 presents another instance of a CSTF3 binding motif (Fig. 4b, Fig. S7a, Table
345 S12). Consistent with this, an AlphaFold3 (AF3) prediction modelling of the full length version
346 of the CPSF1-WDR33-CPSF4-RNA-CSTF3 complex positions the acidic LYG motif of
347 CPSF1 in the basic SLiM-binding pocket on CSTF3 that was reported to bind the FIP1L1
348 motif[38] (Fig. 4c, Fig. S7b).
349 To directly link the observed effects on cellular fitness of the base editor-induced
350 mutations in CPSF1 to the loss of binding to CSTF3, we tested the impact of 10 predicted
351 single amino acid mutations achievable by base editing on CSTF3 binding. Isothermal titration
352 calorimetry (ITC) affinity measurements of CPSF1 peptides to recombinant $CSTF3_{21-549}$
353 revealed reduced binding of all mutants compared to wild-type (Fig. 4d, Fig. S8, Table S13).
354 These data were consistent with the co-affinity purification of CSTF3 with wild-type but not
355 mutant full length YFP-tagged CPSF1 from HeLa cells (Fig. 4e). The affinity measurements
356 revealed that the ABE mutations introduced at E743, E744, L747 and Y748 had the largest
357 impact on binding, consistent with the key buried residues in the reported structure of the
358 FIP1L1 LYG motif bound to CSTF3[38] (Fig. 4c, Fig. S7b).
359
360 Collectively, these data demonstrate that single amino acid mutations introduced into a SLiM
361 by base editing are sufficient to disrupt protein interactions resulting in a fitness defect that can
362 be measured by proteome-wide screening.
363
364 **Essential predicted motif instances largely represent novel motif families**
365 We next investigated the remaining 263 essential SLiMs from the *predicted motif set* to
366 determine if they belong to known or novel motif families (Fig. 5a for top scoring motifs and
367 Table S5). We first compared these SLiMs to the specificity determinants of previously
368 characterized domain-binding motifs (see method section), revealing 39 predicted instances

369 with significant similarity to existing SLiM families (Fig. 5b, Table S14). The majority of these
370 examples belong to well characterized motif classes. For example, the ERTPLL$_{306-311}$ motif in
371 the E3 ubiquitin-protein ligase RNF167 (RNF167) strongly resembles an Adaptin binding
372 Endosome-Lysosome-Basolateral sorting signal (Fig. 5c), aligning with the role of RNF167 in
373 regulating endosome trafficking[39]. Another example is the SWADQVE$_{11-17}$ motif in Eukaryotic
374 translation initiation factor 3 subunit G (EIF3G), which had significant similarity to the binding
375 determinants of a motif binding to MIF4GD. The motif class had previously only been studied
376 in fish, with MIF4G domain-binding motifs in *Danio rerio* Stem-loop binding protein (SLBP)
377 and mRNA-export factor DBP5 (DBP5)[40]. Therefore, the predicted EIF3G instance would
378 extend the taxonomic range of the motif to human proteins (Fig. 5d). To test this, we again
379 used an AP-MS approach comparing full length YFP-tagged human EIF3G wild-type to a
380 mutant with the amino acids of the predicted motif substituted with alanine residues. This
381 showed strong and specific binding to human MIF4G that was lost upon mutation of the SLiM,
382 consistent with the observations in *Danio rerio* (Fig. 5e). Next, we compared the binding
383 determinants of the 264 predicted essential motifs to each other and observed only 16 instances
384 with similarity to other predicted motifs. Those that shared binding determinants consisted of
385 motifs in paralogous proteins, proline-rich motifs, and a cluster of transcription factors that
386 shared hydrophobic motifs resembling transactivation domain[41] (Table S14). Strikingly, the
387 specificity determinant analysis indicated that 225 predicted essential SLiMs showed no
388 significant similarity to any known motif class, suggesting that the majority of the predicted
389 SLiMs do not belong to existing motif families (Fig. 5b, Table S14).
390
391 To further explore the predicted essential motifs, we expanded our AP-MS approach,
392 uncovering putative binding partners for seven of the additional 11 baits tested (Fig. 5f, Fig.
393 S9, Table S12). In WDR33, the proliferation screen had uncovered an essential IPG motif
394 (IPGMG$_{441-445}$), and further inspection of this region revealed an extra conserved copy of this
395 consensus motif (IPGLDWG$_{471-477}$) as well as an additional LPGM$_{420-423}$ motif (Fig. S10a).
396 Mutation of all three resulted in the loss of binding to the Cleavage Stimulation Factor subunits
397 CSTF1 and CSTF3 in AP-MS (Fig. S9). Direct binding of the IPG motifs in WDR33 to CSTF3
398 was validated by pull-down of CSTF3$_{21-549}$ with a WDR33 fragment covering both IPG motifs,
399 WDR33$_{431-485}$, and mutation of both motifs disrupted binding (Fig. S10b). This unstructured
400 region of WDR33 would be in close proximity to CSTF3 based on the structure of the CPSF1-
401 WDR33-CPSF4-RNA-CSTF3 complex (6URO) which contains the domain directly N-
402 terminal of the IPG motifs[42], and our results suggest that the IPG motifs help stabilise the

403 interaction. This possibility is supported by AF3 predictions using full length proteins, where
404 the essential WDR33$_{441-444}$ motif was confidently modelled in a binding pocket in CSTF3 (Fig.
405 S10c). Another interesting observation was the loss of binding between the tyrosine
406 phosphatase PTP1B and Adenomatous Polyposis Coli (APC) protein upon mutation of the
407 EVRSR$_{369-373}$ motif in PTP1B (Fig. S9, Table S12). This novel interaction suggests that the
408 actions of PTP1B and APC in regulating β−catenin during Wnt signaling might be
409 coordinated[43]. This is an example of an essential SLiM in a non-essential protein, where
410 mutation leads to a fitness defect, possibly through the creation of a dominant negative version
411 of the protein causing misregulation of β−catenin.
412

413 **A disease linked mutation in ANKRD17 disrupts SLiM binding to the CCR4-NOT**
414 **complex**

415 We next focused on whether we could obtain novel mechanistic insight into disease-causing
416 mutations potentially affecting SLiM function. Of the 714 essential SLiMs, 104 had reported
417 pathogenic disease mutations (extracted from PepTools[44]) within the motif or in the immediate
418 flanking regions (3 amino acids on either side) with 57 of these being point mutations similar
419 to the outcome of base editing (Table S15). We focused our attention on a PIGTER$_{2426-2431}$
420 motif present in the mRNA binding[45] protein Ankyrin repeat and KH domain-containing 17
421 (ANKRD17), representing a novel uncharacterised motif family (Fig. 5g). The rhAmpSeq
422 screen confirmed that mutation of the SLiM (I2427V) in the genomic locus resulted in
423 depletion from the cell population over time, consistent with a fitness defect (Fig. 3i). An
424 additional instance of the motif in the ANKRD17 paralogue Ankyrin repeat and KH domain-
425 containing 1 (ANKHD1), mapped in the specificity determinant analysis (Table S14), was also
426 found to be essential in our screen. In ANKRD17, a disease mutation (R2434->G) of unknown
427 consequence was present C-terminal to the motif, PIGTERSA**R**$_{2426-2434}$. This mutation is
428 classified as likely pathogenic in ClinVar, linked to Chopra-Amiel-Gordon syndrome (CAGS),
429 a neurodevelopmental syndrome caused by heterozygous loss of ANKRD17 function[46]. Using
430 our AP-MS approach, we found that upon mutation of the entire motif in ANKRD17 binding
431 to CNOT1, CNOT2, and CNOT3 was lost, suggesting that the PIGTER motif mediates binding
432 to the NOT module of the CCR4-NOT deadenylase complex involved in mRNA regulation,
433 consistent with published proximity mapping data[47,48] (Fig. 5h). We validated that the predicted
434 amino acid mutations introduced by the base editors in ANKRD17 disrupted the binding to
435 CNOT1 by affinity purification from HeLa cells, and importantly, we confirmed that the

436 R2434G CAGS disease mutation disrupted binding to CNOT1 (Fig. 5i). We speculate that the
437 defect in ANKRD17 function in patients with this mutation is due to deregulation of specific
438 mRNAs bound by ANKRD17 that are no longer channeled to the CCR4-NOT complex[46].
439
440 **Discussion**
441 SLiMs constitute one of the most abundant regulatory elements in unstructured regions of the
442 human proteome and can readily arise *ex nihilo* or be disrupted through disease mutations.
443 Until now, our understanding of SLiM contribution to cell function has been limited due to
444 lack of efficient endogenous functional screening methods. In this study, we utilize base editor
445 screening to create a comprehensive map of cellular SLiM dependency in HAP1 cells and show
446 that mutations introduced by base editors are sufficient to disrupt SLiM mediated interactions,
447 allowing us to couple disruption of interactions to cellular fitness effects. Importantly, the
448 interpretation of our data is not confounded by structural considerations as we target SLiMs in
449 unstructured regions. Our screens reveal that numerous motifs are required for optimal cell
450 proliferation both in HAP1 and RPE1 cells. By defining a set of essential SLiMs, we provide
451 a priority list for future studies, as well as gRNAs to specifically target these. We provide a
452 user-friendly web-based resource (https://slim.icr.ac.uk/projects/base_editing) that allows for
453 interactive exploration of our comprehensive dataset, thereby providing a valuable tool for the
454 research community. We furthermore show that binding partners of essential SLiMs can be
455 identified by a streamlined AP-MS approach comparing affinity-tagged wild-type protein to a
456 variant where the SLiM is mutated. This approach can now be applied to many of the predicted
457 SLiMs we identify as essential, followed up by mutational binding analyses and modelling to
458 characterize mode of binding.
459
460 While we cannot exclude that some motifs do not score due to low editing efficiency of the
461 gRNAs targeting them, our sensor screen suggests that the majority of gRNAs, at least for
462 ABE_{NGN}, are editing efficiently (Fig. 3c). We therefore favor that the ~90% non-scoring motifs
463 are not required for proliferation in HAP1 cells under normal cell culture conditions.
464 Importantly, our data show that essential SLiMs are more likely to occur within essential
465 proteins (Fig. 2j), arguing that such fitness screens are mainly able to measure mutational
466 effects in essential proteins. Thus, we consider the non-scoring motifs a large unexplored SLiM
467 space with context specific functions. Our gRNA libraries now provide a powerful tool to
468 explore these context specific functions of SLiMs by applying the libraries on additional cell
469 types or during cell perturbations. Our comparison of RPE1 and HAP1 cells already confirms

470 the utility of our libraries to identify cell line specific differences in SLiM dependencies. A
471 further avenue to explore is to employ the gRNA libraries under conditions where cells are
472 treated with therapeutic agents, potentially identifying SLiM interactions that can constitute
473 novel vulnerabilities that can be exploited for drug development[49].
474
475 In conclusion, our work highlights that base editing technologies can be applied to discover
476 novel protein binding regions and prioritize interaction interfaces based on their importance to
477 cell fitness. Our approach provides a blueprint for future efforts to further dissect SLiM
478 function.
479
480 **Acknowledgments**
481 Work at the Center for Epigenetic Cell Memory, Danish Cancer Institute is supported by The
482 Danish National Research Foundation (DNRF195). Work at the Novo Nordisk Foundation
483 Center for Protein Research is supported by NNF14CC0001. Work in the Nilsson lab is
484 supported by grants from the Novo Nordisk Foundation (NNF0082227 and NNF0065098),
485 DFF-FNU (4283-00188B) and the Danish Cancer Society (R269-A15586-B71). N.E.D. is
486 funded by a Cancer Research UK Senior Cancer Research Fellowship (C68484/A28159).
487 E.P.T.H. is funded by a Lundbeck Foundation Postdoc Grant (R380-2021-1221). We thank the
488 protein production and characterization facility at Novo Nordisk Foundation Center for Protein
489 Research for all their help with the project. We thank the Genomics Platform at Novo Nordisk
490 Foundation Center for Protein Research and Center for Stem Cell Medicine (reNEW) for
491 technical support and use of instruments. NGS data processing and analysis were performed
492 using the DeiC National Life Science Supercomputer at Technical University of Denmark
493 (www.computerome.dk).
494

Methods

SLiM library design

SLiM-containing regions were collected from three sources (time of download: April 2022): 1) manually *curated* motif information and 2) motif-containing *structure* processing to retrieve experimentally validated motifs (as a source for the *reported* dataset); and 3) *evolutionary footprinting* to define putative motifs (source for the *predicted* dataset). *Curated* data was extracted from the motif literature, or retrieved programmatically from the ELM datasets. *Structural* data on motif instances was programmatically extracted from PDB by screening all PDB structures that contain a biological assembly with 2+ chains; defining disordered peptides in these structures using FLIPPER[50]; and extracting SLiMs using the discriminator of bound footprint <1500 $Å^2$, mono-partite interface as defined by a single distinct secondary structural properties in the bound conformation, and defined by FLIPPER as disordered across the entire chain. *Evolutionary footprinting* data was defined using the SLiMPrints algorithm to define conserved regions with SLiM-like evolutionary patterns in the IDRs of the human proteome[51]. Significant predicted motifs were defined with two cut-offs to enrich for functionally important motifs, for instances in DepMap essential proteins with mean gene effect < -0.5 or overlapping disease relevant SNVs, a SLiMPrints significance cutoff of 0.005 was used, otherwise a cut-off of 0.0005 was applied. Accessibility filter was then applied to the SLiMPrints prediction to enrich for accessible intracellular motif instances by filtering predictions with an IUPred score less than 0.33 and removing peptides overlapping Transmembrane, Extracellular or Lumenal annotation in UniProt[52,53].

gRNA library design

gRNAs were designed to target the protein level of the motifs in the SLiM library based on the PAM (NGG or NGN), editing window (nucleotides 4-8 in the target DNA of the gRNA for all editors) and editing preference (ABE A→G, CBE C→T) of each base editor applied in the screen (ABE_{NGN}, CBE_{NGN}, ABE_{NGG}, and CBE_{NGG}). The UniProt defined protein region was mapped to the corresponding transcript in Ensembl. The genomic sequence information of the gene, the oligonucleotide sequence of the gene in the genome, the CDS (CoDing Sequence) of the transcript, and the intronless oligonucleotide sequence product of the gene that is translated into a protein, were retrieved from Ensembl. The position of the target nucleotides in the CDS were determined by mapping the UniProt start and stop positions to the CDS, and the fidelity of the mapping was confirmed by mapping the peptide sequence to the translated CDS of the region. In the cases where no translated Ensembl transcript maps to UniProt protein sequence,

529 the longest Ensembl transcript was retrieved, the target peptide region was mapped to the
530 longest translated Ensembl transcript and the target positions are adjusted to account for the
531 remapping. Next, the Ensembl CDS and genomic sequence were aligned and the target
532 genomic sequence nucleotides for editing were determined through the translated Ensembl
533 CDS. This step allows for editing of nucleotides flanking intron-exon boundaries. Potential
534 PAM sites upstream and downstream of the target genomic sequence nucleotides were
535 identified on both strands. PAM sites that positioned the edit window of the base editor over
536 the target genomic sequence nucleotides were investigated further. The editing window was
537 searched for nucleotides corresponding to the preference of the base editor to find potential
538 nucleotide edits. If one or more editable nucleotides were available for a window all edits were
539 made simultaneously *in silico* and the resulting protein-coding changes were compared to the
540 wild-type protein sequence. gRNAs predicted to result in non-synonymous mutations were
541 retained as part of the SLiM dependency map gRNA library. A set of previously described
542 control gRNAs[16] including 62 positive control targeting splice sites of essential genes and 255
543 negative controls targeting intergenic regions of the genome and non-targeting gRNAs were
544 also included in the library design.
545
546 **Library cloning**
547 Each gRNA in the library was extended to create a pool of oligonucleotide sequences (83 nt)
548 following a previously published design[16]. The oligonucleotide consisted of primer sites for
549 differential amplification, flanking appropriate overhang sequences with Esp3I restriction
550 enzyme recognition sites and the gRNA: 5'-[Forward primer, 20
551 nt]CGTCTCACACCG[sgRNA, 20 nt]GTTTCGAGACG[Reverse Primer, 20 nt]. The
552 oligonucleotide pool was synthesised by GenScript (CustomArray). Individual subpools were
553 amplified in a reaction of 25 µl 2x NEBnext PCR master mix (New England Biolabs), 1 µl
554 oligonucleotide pool (10 ng/µl), 2.5 µl each of forward and reverse primer at a final
555 concentration of 0.5 uM, and 19 µl water, using a unique primer combination for each subpool
556 (Table S16).
557
558 The number of PCR reactions per subpool was determined by scaling to 24 reactions per 92,000
559 gRNAs. PCR cycling conditions were: (1) 98 °C for 30 s; (2) 53 °C for 30 s; (3) 72 °C for 30
560 s; (4) go to (1) x 19; (5) 72 °C for 60 s; (6) 10 °C hold. Amplicons were PCR-purified
561 (NucleoSpin Gel and PCR Clean-up, Macherey-Nagel). Library vector was pre-digested with
562 FastDigest Esp3I (Thermo Fisher Scientific) and gel-purified (NucleoSpin Gel and PCR

563 Clean-up, Macherey-Nagel). The following library vectors were used: CBE_{NGG}: pRDA_256
564 (BE3.9Max-SpCas9); CBE_{NGN}: pRDA_478 (BE3.9Max-SpG); ABE_{NGG}: pRDA_426 (ABE8e-
565 SpCas9); and ABE_{NGN}: pRDA_479 (ABE8e-SpG). pRDA_256 (Addgene #158581),
566 pRDA_478 (Addgene #179096), pRDA_426 (Addgene #179097), pRDA_479 (Addgene
567 #179099) were gifts from John Doench & David Root. Amplicons were cloned into library
568 vectors by Golden Gate reactions containing 25 µl 2x Rapid Ligase Buffer (Qiagen
569 Enzymatics), 0.25 µl Bovine Serum Albumin (20 mg/mL) (New England Biolabs), 1 µl
570 FastDigest Esp3I (1 U/µl), 0.25 µl T7 Ligase (3x10^6 U/mL) (Qiagen Enzymatics), 15 ng
571 purified amplicon, 100 ng digested vector, and water in a total volume of 50 uL. The number
572 of Golden Gate reactions performed per subpool was determined by scaling to 56 reactions per
573 92,000 gRNAs. Golden Gate reaction conditions were: (1) 37 °C for 5 min; (2) 20 °C for 5
574 min; (3) go to (1) x 24; (4) 10 °C hold. Ligated plasmid product was purified first by PCR-
575 purification and then by isopropanol precipitation in the following reaction: 300 µl PCR-
576 purified plasmid, 300 µl isopropanol, 3 µl GlycoBlue Coprecipitant (final conc. 75 ng/µl)
577 (Thermo Fisher Scientific), 50 mM NaCl. Samples were vortexed (10 s full speed) and
578 incubated 15 min at RT before centrifugation at 16,000xg for 15 min at RT. Precipitated DNA
579 was washed twice with ice-cold 80% Ethanol, air-dried and resuspended in TE buffer (1 µl TE
580 buffer per Golden Gate reaction) at 55 °C for 10 min. The number of electroporations per
581 subpool was determined by scaling to 16 electroporations per 92,000 gRNAs. The purified
582 plasmid library was electroporated into Endura Electrocompetent cells (Lucigen) and grown at
583 30 °C for 16 h on agar with 100 µl/mL carbenicillin. Colonies were scraped off and plasmid
584 DNA was prepared (NucleoBond Xtra Maxi EF, Macherey-Nagel). To confirm library
585 representation and distribution, inserts were amplified from plasmid library in a reaction of 25
586 µl 2x NEBnext PCR master mix, 1 µl plasmid library (10 ng/µl), 2.5 µl each of i5 and i7 primers
587 at a final concentration of 0.5 uM, and 19 µl water, using unique primer combinations (Table
588 S15). PCR cycling conditions were: (1) 98 °C for 30 s; (2) 98 °C for 10 s; (3) 55 °C for 30 s;
589 (4) 65 °C for 15 s; (5) go to (2) x 19; (6) 65 °C for 5 min; (7) 10 °C hold. Amplicons were first
590 gel-purified and then re-purified by PCR-purification. Samples were sequenced on a
591 NextSeq500 or a NextSeq2000 (Illumina) with a 2% spike-in of PhiX with 21 (NextSeq500)
592 or 22 (NextSeq2000) dark cycles. Fastq files were generated using bcl2fastq v.2.19.1 and reads
593 were trimmed to 20 bp using cutadapt 1.18, removing a variable number of basepairs at the
594 start and end depending on the size of the primer stagger. MAGeCK[54] 0.5.8 was used to assign
595 the trimmed reads to the gRNAs in the library and create the gRNA count matrix.
596

597 **Virus production**

598 HEK 293T/17 cells (ATCC, CRL-11268) were seeded in DMEM GlutaMax (Gibco) + 10%
599 FBS (Gibco) + 100 U/mL penicillin–streptomycin (Gibco). Medium was exchanged with Opti-
600 MEM (Gibco) + 5% FBS 24h later for co-transfection of the base editing sgRNA library with
601 lentiviral packaging plasmids pMD2.G and psPAX2 using Lipofectamine 3000 (Invitrogen). 6
602 hours after transfection, the medium was exchanged for DMEM GlutaMax + 10% FBS +
603 100 U/mL penicillin–streptomycin + 1% bovine serum albumin. 48 hours after transfection,
604 viral particles were collected, filtered through a 0.45 µm syringe filter, and stored at -80ºC.
605 pMD2.G and psPAX2 were gifts from Didier Trono (Addgene plasmid # 12259;
606 http://n2t.net/addgene:12259; RRID:Addgene_12259 and Addgene plasmid # 12260;
607 http://n2t.net/addgene:12260; RRID:Addgene_12260).

608

609 **Competitive Growth Screen**

610 HAP1 cells (#C859, Horizon Discovery, SNB: 35369) were cultured in IMDM GlutaMax
611 (Gibco) supplemented with 10% FBS and 100 U/mL penicillin–streptomycin and passaged
612 every three days by trypsinization in 0.05% trypsin-EDTA (Gibco). Each screen was performed
613 as a triplicate with three separate transductions at a coverage above 675-fold sgRNA
614 representation, which was maintained throughout the screen. HAP1 cells were transduced with
615 the indicated lentiviral libraries at a low multiplicity of infection (MOI; 0.3-0.4) by treating
616 cells with 8 µg/mL polybrene and lentiviral supernatants for 24 hours. Viral titers required to
617 achieve an MOI of ~0.3 were pre-determined in a titration experiment. Transduced cells were
618 selected by treatment with 1.5 µg/mL puromycin for 48 hours. After selection, cells were
619 cultured for 21 days. Genomic DNA was extracted from cell pellets harvested after completion
620 of selection, which we consider the start of the screen (d0), and after 18 (d18), and 21(d21)
621 days of culturing. ~600 bp plasmid library fragments containing the sgRNA sequences were
622 amplified from the genomic DNA by PCR using NEBNext Ultra II Q5 Master Mix with the
623 LCV2_forward and LCV2_reverse primers (Table S16). A second PCR reaction introduced i5
624 and i7 multiplexing barcodes (Table S16), whereafter ~220 bp PCR products were gel-purifed
625 and sequenced on Illumina NextSeq2000 with a 2% spike-in of PhiX with 22 dark cycles. Fastq
626 files were generated using bcl2fastq v.2.19.1 and reads were trimmed to 20 bp using cutadapt
627 1.18, removing a variable number of basepairs at the start and end depending on the size of the
628 primer stagger. MAGeCK 0.5.8 [54]was used to assign the trimmed reads to the gRNAs in the
629 library (motif-targeting and control gRNAs (essential splice sites, non-targeting, intergenic))
630 and create the count matrix.

631

632 **Base editing data processing**

633 NGS count data from the four screens (three timepoints and three replicates each) were
634 processed to define significantly changing gRNAs using a limma statistical analysis[18]. Low-
635 abundance gRNAs were removed (counts <30). The vast majority of gRNAs had a single exact
636 match in the genome, however, non-specific gRNAs with off-target genome binding sites
637 showed increased dropout correlated with the number of off-target binding sites. Consequently,
638 gRNAs with 5 or more off-target matches in the genome were removed from downstream
639 analysis, leading to the exclusion of 483 gRNAs. Raw sequencing counts were normalized for
640 input into the limma software by dividing the number of reads for a specific gRNA by the total
641 number of reads from the sample, multiplying by a million to the define the transcripts per
642 million (TPM) and calculating the $\log_2$ of the gRNA TPM to define the $\log_2$ transcripts per
643 million (log2TPM). The log2TPM values were compared using limma, resulting in fold change
644 (FC) and p-values for each gRNA. These p-values were compared for the positive and negative
645 controls, and a p-value cut-off of 0.01 showed a false positive rate of ~1% across all four
646 screens, with a true positive rate of up to 79%. Consequently, a p-value of less than 0.01 was
647 set as the threshold considered to impact cell proliferation.

648 Next, the gRNAs mapping to each motif-containing region were analyzed to define
649 groups of significantly changing gRNAs. The main metrics used was the number of unique
650 significant gRNAs per motif across all screens. Per screen, the direction of the effect on
651 proliferation was quantified as the mean fold change of the gRNAs to define depletion or
652 enrichment, and the fold change of the gRNAs for a region were compared against the fold
653 change for all gRNAs in the analysis using a Mann–Whitney U test to calculate an enrichment
654 or depletion p-value. AlphaFold2 was applied as post-screen filtering to flag motifs that are
655 buried or in regions incompatible with binding. The structural context of each region was
656 defined using AlphaFold2 models. Full-length protein AlphaFold2 models were retrieved from
657 the AlphaFold2 protein structure database (https://alphafold.ebi.ac.uk/). The structured regions
658 of the proteins were extracted using the xProtCAS tool[55] using a graph-based community
659 detection algorithm on the AlphaFold2 predicted aligned error (PAE) matrix. Based on the
660 xProtCAS, the determined structural state of each residue was classified as Domain, Loop or
661 Unstructured. The unstructured proportion contains IDRs and coiled coils; coiled coil regions
662 were defined as predicted helices greater than 25 amino acids in length; the remaining residues
663 were classified as IDR. Potential intramolecular IDR - Domain contacts were also determined
664 as the first or last region in an autonomous structural module of length less than 30 connected

665 by a loop of greater than 20 amino acids. The SLiMPrints portion of the initial library contained
666 760 predicted motif instances and 10,606 gRNAs that were filtered after AlphaFold2 predicted
667 they were located in coiled coil or domain regions. As a result, 6,489 of the 7,293 motif-
668 containing regions were retained for downstream processing and analysis.
669
670 **Region annotation**
671 Regions were annotated using the PepTools peptide annotation tool
672 (https://slim.icr.ac.uk/tools/peptools/) to define overlapping mutagenesis (source: UniProt),
673 disease relevant SNVs (source: gnomAD, dbSNP, COSMIC curated, TOPMed, NCI-TCGA
674 COSMIC, NCI-TCGA, ExAC, Ensembl, ESP, ClinVar, 1000 Genomes retrieved from the EBI
675 API), and other features of interest (source: UniProt)[44]. Protein complex members were
676 retrieved from the Complex Portal[27]. Pathway enrichment analysis was performed using the
677 PANTHER enrichment tool[56] to identify enriched Reactome pathway annotation using the
678 PANTHER API (https://www.pantherdb.org/services/) by comparing essential motif-
679 containing proteins to the protein set tested in the base editing analysis. The specificity
680 determinant comparison analysis was performed with the CompariPSSM tool[57]. The specificity
681 determinants of the regions targeted in the base editing analysis were extracted as conservation-
682 derived PSSMs. An orthologue alignment was created for each region by running the
683 GOPHER[58] algorithm on the region-containing protein against a search database of metazoan
684 proteomes. For each region, the columns corresponding to the edited region were extracted
685 from the alignment and converted into a frequency PSSM. Frequency PSSMs were also created
686 from aligned experimentally tested motif instances for motif classes from the ELM resource
687 (http://elm.eu.org/) and a more extensive in-house curated MoMaP motif dataset
688 (https://slim.icr.ac.uk/momap/). The base-edited PSSM set was then compared against the
689 ELM PSSM dataset, MoMaP dataset, and itself using the CompariPSSM tool. Significant
690 matches were defined as comparisons with a CompariPSSM similarity significance of $<1x10^{-}$
691 ^{5}, dissimilarity score of <1 and PSSM overlap >75%.
692
693 **Library design for the sensor and rhAmpSeq screens**
694 The *sensor screen* library includes all gRNAs targeting the 714 motifs with 2+ unique
695 significant gRNAs listed in Table S5, filtered to retain only gRNAs predicted to result in non-
696 synonymous mutations with ABE_{NGN} (ABE8e-SpG; total 8505 gRNAs, Table S8). To generate
697 the sensor region, each gRNA was mapped back to the genome, and a 41nt sequence
698 overlapping the targeted site was extracted, consisting of an 11nt upstream flank, the 20nt

699 gRNA target, and a 6nt downstream flank. Eight gRNAs predicted to mutate splice sites of
700 each targeted gene were also included in the library. In addition to the positive and negative
701 controls gRNAs from the original screen, the library included 500 positive control gRNAs
702 targeting splice sites of essential genes and 500 negative controls targeting intergenic regions
703 of the genome (250 gRNAs) and non-targeting gRNAs (250 gRNAs). The library for the
704 *rhAmpSeq screen* constitues the top scoring ABE_{NGN} gRNA for the top 25 novel motifs in a
705 gRNA-sensor set-up (Table S8). Each gRNA and sensor construct in the libraries were
706 designed inspired by the general architecture from published work[28] with an additional
707 inclusion of a 4 nt barcode at the 5' end of each insert[31] to create a pool of oligonucleotide
708 sequences. Each oligonucleotide consists of primer sites for differential amplification flanking
709 appropriate overhang sequences with Esp3I restriction enzyme recognition sites and the gRNA,
710 gRNA scaffold sequence, termination site, sensor site (gRNA with flanking sequences) and bar
711 code (206 nt):
712 [Fw_primer,20nt]CGTCTCACACCG[gRNA,20nt]GTTTGAGAGCTAGAAATAGCAAGT
713 TCAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTT
714 TTT[flank,11nt][gRNA,20nt][flank,6nt][barcode,4nt]GTTTCGAGACG[Rv_primer,20nt].
715

716 **Sensor and rhAmpSeq screen library cloning**

717 An oligopool comprising both *sensor* and *rhAmpSeq* screen libraries was ordered as a Sureprint
718 OLS Hifi Oligo library 171-210 nt, 15K (Agilent). The libraries were amplified in a total of 30
719 reactions for the *sensor* library and 4 reactions for the *rhAmpSeq* library, each reaction
720 consisting of 25 µl 2x NEBnext PCR master mix, 1 µl oligonucleotide pool (1 ng/µl), 0.25 µl
721 each of forward and reverse primer at a final concentration of 0.5 uM, and 23.5 µl water, using
722 a unique primer combination for each library (Table S16). PCR cycling conditions were: (1)
723 98 °C for 30 s; (2) 53 °C for 30 s; (3) 72 °C for 30 s; (4) go to (1) x 14; (5) 72 °C for 60 s; (6)
724 10 °C hold. Amplicons were PCR-purified. Library vector pLenti-U6-sensor-ABE8e-
725 nSpG(D10A) was cloned by KpnI+EcoRI restriction of pRDA_479 and insertion of synthetic
726 DNA carrying U6 promoter followed by BsmBI cassette, thus removing the gRNA scaffold
727 sequence to allow insert of the gRNA-scaffold-sensor cassette. The vector was pre-digested
728 with FastDigest Esp3I and gel-purified. Amplicons were cloned into pLenti-U6-sensor-
729 ABE8e-nSpG by Golden Gate cloning in a total of 15 reactions for the *sensor library* and 6
730 reactions for the *rhAmpSeq library*. Each reaction contained 25 µl 2x Rapid Ligase Buffer,
731 0.25 µl Bovine Serum Albumin (20 mg/mL), 1 µl FastDigest Esp3I (1 U/µl), 0.25 µl T7 Ligase
732 (3x10^6 U/mL), 15 ng purified amplicon, 100 ng digested vector, and water in a total volume

733 of 50 µL. Golden Gate reaction conditions were: (1) 37 °C for 5 min; (2) 20 °C for 5 min; (3)
734 go to (1) x 24; (4) 10 °C hold. Ligated plasmid product was purified first by PCR-purification
735 and then by isopropanol precipitation as described above. The purified plasmid library was
736 electroporated into Endura Electrocompetent cells (Lucigen) and grown at 30 °C for 16 h in
737 LB with 100 µl/mL carbenicillin. Cells were harvested by centrifugation and plasmid DNA
738 was prepared. To confirm library representation and distribution, inserts were amplified from
739 plasmid library in a reaction of 25 µl 2x NEBnext PCR master mix, 1 µl plasmid library (10
740 ng/µl), 0.25 µl forward and 0.25 µl reverse primer (a 1:1:1:1:1:1 mix of primers
741 Fw_sensor_P5_S0-Fw_sensor_P5_S5 and Rv_sensor_P7_S0-Rv_sensor_P7_S5 with various
742 stagger lengths to create sequence diversity in the samples; Table S16) at a final concentration
743 of 0.5 uM, and 23.5 µl water. PCR cycling conditions were: (1) 98 °C for 30 s; (2) 98 °C for
744 10 s; (3) 65 °C for 30 s; (4) 72 °C for 20 s; (5) go to (2) x 16; (6) 72 °C for 3 min; (7) 10 °C
745 hold. A second PCR reaction of 25 µl NEBnext PCR master mix, 10 ng PCR1 product, 2.5 µl
746 forward and reverse primer (Table S16) in a total of 50 µl introduced i5 and i7 multiplexing
747 barcodes using the following PCR cycling conditions: (1) 98 °C for 30 s; (2) 98 °C for 10 s;
748 (3) 65 °C for 75 s; (4) go to (2) x 7; (6) 72 °C for 5 min; (7) 10 °C hold. Amplicons were first
749 gel-purified and then re-purified by PCR-purification. Samples were sequenced on a
750 NextSeq2000 (Illumina) with a 25% spike-in of PhiX. Fastq files were generated using
751 bcl2fastq v.2.19.1 and reads were trimmed using cutadapt 1.18, removing a variable number of
752 basepairs at the start and end depending on the size of the primer stagger. MAGeCK[54] 0.5.8
753 was used to assign the trimmed reads to the gRNAs in the library and create the gRNA count
754 matrices.
755
756 **Virus production and competitive growth for sensor and rhAmpSeq screens**
757 Lentiviral libraries were produced as described above and the screens were conducted as
758 described above as a triplicate (three separate transductions) at low multiplicity of infection
759 (MOI ~0.3) and high coverage (>1000X) with the following alterations: A) 24 hours after
760 transduction, transduced HAP1 cells were selected by splitting into medium with 1 µg/mL
761 puromycin. B) The screen was terminated at day 18. Additionally, the *sensor screen* was
762 performed in hTERT RPE1 p53-/- cells (a kind gift from D. Durocher) with the following
763 alterations compared to HAP1 cells: A) Cells were cultured in DMEM GlutaMax (Gibco) +
764 10% FBS (Gibco) + 100 U/mL penicillin–streptomycin (Gibco) and 0.25% trypsin-EDTA
765 (Gibco) was used for trypsinization. B) Transduced cells were selected by splitting into 15
766 µg/mL puromycin and incubation for 24 hours, followed by an additional split into 15 µg/mL

767 puromycin and incubation for 24 hours. For sensor libraries, a ~255 bp plasmid library
768 fragments containing the sgRNA sequences including the sensor were amplified from the
769 genomic DNA by PCR using NEBNext Ultra II Q5 Master Mix with a 1:1:1:1:1:1 mix of
770 primers Fw_sensor_P5_S0-Fw_sensor_P5_S5 and Rv_sensor_P7_S0-Rv_sensor_P7_S5 with
771 various stagger lengths to create sequence diversity in the samples (Table S16). A second PCR
772 reaction introduced i5 and i7 multiplexing barcodes (Table S16), whereafter ~330 bp PCR
773 products were gel-purifed and sequenced (paired end, Read 1: 190 cycles ; read 2: 75 cycles)
774 on Illumina NextSeq2000 with a 25% spike-in of PhiX. Fastq files were generated using
775 bcl2fastq v.2.19.1 and reads were trimmed using cutadapt 1.18, removing a variable number of
776 basepairs at the start and end depending on the size of the primer stagger. MAGeCK[54] 0.5.8
777 was used to assign the trimmed reads to the gRNAs in the libraries and create the gRNA count
778 matrix.
779
780 **rhAmpSeq pooled amplicon sequencing**
781 Genomic DNA from the *rhAmpSeq* screen samples were subjected to pooled amplification of
782 the 25 genomic loci targeted by the gRNAs (Table S17) using the rhAmpSeq CRISPR analysis
783 system (IDT) following the instructions of the manufacturer with a custom 25 primer pair
784 rhAmpSeq CRISPR panel. Briefly, the amplification reaction consisted of 5 µl 4x rhAmpSeq
785 CRISPR Library mix 1 (IDT), 11 µl genomic DNA (4.55 ng/µl) and 2 µl each of forward and
786 reverse primer mixes at a concentration of 250 nM. PCR cycling conditions were: (1) 95 °C
787 for 10 min; (2) 95 °C for 15 s; (3) 61 °C for 8 min; (4) go to (2) x 13; (5) 99.5 °C for 15 min;
788 (6) 4 °C hold. The PCR products were diluted 1:20 in ultrapure water and immediately used as
789 template in a second PCR reaction that introduced i5 and i7 multiplexing barcodes. The
790 amplification reaction consisted of 5 µl 4x rhAmpSeq CRISPR Library mix 2 (IDT), 11 µl
791 diluted PCR1 product and 2 µl each of i5 and i7 indexing primers in a unique combination at a
792 concentration of 1 µM (Table S16). PCR cycling conditions were: (1) 95 °C for 3 min; (2) 95
793 °C for 15 s; (3) 60 °C for 30 s; (4) 72 °C for 30 s; (5) go to (2) x 23; (6) 72 °C for 1 min; (6) 4
794 °C hold. The indexed PCR reactions were pooled and purified by adding 100 µl Agencourt®
795 AMPure® XP Beads to 100 µl pooled PCR reaction, followed by incubation for 10 min at RT
796 and placing in a magnetic separation rack for 5 min. The beads were washed twice in 80%
797 ethanol and dried, and the DNA eluted in 22 µl IDTE pH 8.0 by vortexing. The tube was placed
798 in the magnetic separation rack, and the eluate was collected without beads and sequenced on
799 Illumina NextSeq2000 (paired end, Read 1: 161 cycles ; read 2: 161 cycles) with a 25% spike-
800 in of PhiX. Fastq files were generated using bcl2fastq v.2.19.1 and Read 1 and 2 were stiched.

801

802 **Sensor and rhAmpSeq screen data processing**

803 *Sensor screen:* For each timepoint and replicate, paired-end NGS data were generated with the
804 gRNA sequence located in Read 1 and the corresponding sensor sequence in Read 2. gRNA
805 regions were extracted from Read 1 using the construct flanking sequences and reads lacking
806 a designed gRNA were removed. gRNA counts were determined and processed as described
807 for the initial base editing screen. For each Read 1 containing a valid gRNA, the paired Read
808 2 was processed by extracting the sensor region using the designed flanking sequences. Each
809 sensor was verified against the expected barcode for its gRNA, and barcode–gRNA
810 mismatches were discarded. Sensor sequences were then trimmed to the designed region,
811 grouped by gRNA, and converted into gRNA-centric nucleotide-level edit count tables. These
812 nucleotide edits were mapped to protein-level edits, including splice-site and premature stop
813 annotations where relevant, and counts for identical protein-level editing outcomes were
814 merged. The final output included nucleotide editing frequencies, protein editing frequencies,
815 and detailed frequencies for specific protein, splice-site, and stop edits. Analysis focused on
816 editing data obtained at day 3. *rhAmpSeq screen, amplicon data:* The rhAmpSeq amplicon
817 NGS data were processed largely as described for the sensor sequence in the sensor screen with
818 the following exceptions: a single stitched read was used; reads were trimmed using unique
819 genomic flanking sequences to recover amplicon sequences aligned to the designed constructs;
820 and reads corresponding to unedited amplicons were removed, as they likely originated from
821 cells transduced with a different gRNA. *rhAmpSeq screen, sensor data:* NGS data was
822 processed using the pipeline described for the *sensor screen*. Final editing outcome frequencies
823 for the *rhAmpSeq screen's* sensor and amplicon data were aggregated at the protein residue
824 level to generate position-specific editing frequency matrices for visualization as sequence
825 logos.

826

827 **Deep sequencing of CPSF1 locus**

828 Individual sgRNAs targeting CPSF1 were ordered as DNA oligonucleotides (TAG
829 Copenhagen) with overhangs for Esp3I cloning: Forward oligo: 5'-CACCG[gRNA (20nt)]-3';
830 Reverse oligo: 5'-AAAC[reverse complement gRNA (20nt)]C-3' (Table S16). The
831 oligonucleotides were annealed and phosphorylated by T4 PNK (NEB) and cloned into either
832 base editor construct pRDA_478 or pRDA_479 using golden gate cloning with Esp3I and T7
833 ligase. Each sgRNA was cloned into the backbone where it was predicted to give an edit. Empty
834 pRDA_478 and pRDA_479 were used as controls. Ligated plasmids were transformed into

835 NEB Stable Competent E. coli and grown on LB agar containing ampicillin at 30°C for 24 h,
836 followed by plasmid preparation from single colonies. Virus particles for each sgRNA-
837 expressing base editor construct were produced as described above. HAP1 cells were
838 transduced as a single biological replicate with each sgRNA-expressing base editor construct,
839 selected, and cultured for 18 days after selection (d0-d18) as described above. Genomic DNA
840 was extracted from cell pellets harvested at d0 and d18 using the QIAamp DNA Blood Mini
841 Kit (Qiagen). A PCR reaction was set up to amplify the genomic CPSF1 locus, where editing
842 was predicted to occur, and to introduce Illumina TruSeq Adaptor flaps. This was done using
843 NEBNext Ultra II Q5 Master Mix with primers: CPSF1_Fw and CPSF1_Rv (Table S16). A
844 second PCR reaction introduced unique combinations of Illumina i5 and i7 multiplexing
845 barcodes using NEBNext Ultra II Q5 Master Mix and Illumina TruSeq primers listed in Table
846 S16. Gel-purified PCR products were sequenced on Illumina NextSeq2000 with a 25% spike-
847 in of PhiX and analysed by CRISPResso2[59] (Supplemental Data 1).
848
849 **Cloning**
850 pcDNA5/FRT/TO plasmids encoding venus-tagged (an enhanced YFP derivate referred to as
851 YFP for simplicity) baits (Table S16) in a wild-type (WT) or mutant (Mut) version were
852 generated by gene synthesis, mutagenesis, and subcloning into pcDNA5/FRT/TO/venus
853 (GeneArt, ThermoScientific). pcDNA5/FRT/TO/Venus-CPSF1 and NCAPH were cloned
854 from in house constructs and mutated by site-directed mutagenesis PCR. WDR33 431-485
855 sequences were amplified by PCR and cloned into pGEX4T-1. The E.coli expression construct
856 for CSTF3 21-549 was done by LIC cloning into an in house expression vector. All constructs
857 were verified by sequencing.
858
859 **Affinity purification coupled to mass spectrometry (AP-MS) sample preparation**
860 HeLa cells were transfected with plasmids encoding YFP-tagged full length protein baits (or
861 larger fragments in the case of ANKRD17$_{1950-2603}$ and WDR33$_{1-525}$) (Table S16) using
862 Lipofectamine 2000 24 or 48 hours prior to harvesting. The bait proteins were either wild-type
863 (WT) or motif mutant (Mut) (Fig. S9, Table S16). Cells were lysed on ice in lysis buffer (50
864 mM Tris pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.1% NP-40) supplemented with 1 mM DTT,
865 complete protease inhibitor cocktail (Roche), and 4U/mL RNAse, and sonicated for 10 min
866 (30''ON/30''OFF, max) using Bioruptor Plus (Diagenode). Lysates were cleared by
867 centrifugation at 20,000 g at 4°C for 30 minutes, and YFP-tagged baits were affinity purified
868 using GFP-trap beads (ChromoTek). Purified protein complexes were washed twice in lysis

869 buffer and once in PBS. Partial on-bead digestion was performed with buffer (2 M urea; 2 mM
870 DTT; 20 µg/ml Trypsin; 50 mM Tris, pH 7.5) incubated at 37°C for 30 min at 1,400 rpm.
871 Supernatants were transferred to new tubes, alkylated with 25 mM chloroacetamide (CAA) and
872 further digested overnight at RT and 1,000 rpm. Digestion was terminated with 1%
873 trifluoroacetic acid (TFA). Peptides were desalted and purified using styrenedivinylbenzene-
874 reversed phase sulfonate (SDB-RPS) StageTips prepared in 0.2% TFA. Peptides were washed
875 and eluted with an elution buffer (80% acetonitrile (ACN); 1% ammonia) prior to vacuum-
876 drying. Dried peptides were reconstituted in 2% ACN and 0.1% TFA. Technical quadruplicates
877 were performed.
878
879 **Liquid chromatography-mass spectrometry (LC-MS) analysis**
880 All samples were loaded onto Evotips Pure and measured with a data-independent acquisition
881 (DIA) method. 200 ng of peptides were loaded onto Evotips following the manufacturer's
882 instructions, and then analyzed with an Evosep One LC system (Evosep Biosystems) coupled
883 online to an Orbitrap mass spectrometer (Orbitrap Astral or Orbitrap Exploris 480, Thermo
884 Fisher Scientific)[60-62]. Eluted peptides were separated on a 8-cm-long PepSep column (150 µm
885 inner diameter packed with 1.5 µm of Reprosil-Pur C18 beads (Dr Maisch)) in a standard
886 gradient method with a stainless emitter (30 µm inner diameter). The mobile phases were 0.1%
887 formic acid in liquid chromatography (LC)–MS-grade water (buffer A) and 0.1% formic acid
888 in acetonitrile (buffer B). Data was acquired in DIA mode.
889
890 For the samples acquired in the Orbitrap Exploris 480, peptides were eluted online from the
891 EvoTips using an Evosep One LC system and analysed at 30 samples per day (44-min
892 gradient). The mass spectrometer was operated in positive mode using the DIA mode. Full scan
893 precursor spectra (350–1,000 m/z) were recorded in profile mode using a resolution of 120,000,
894 with a normalised automatic gain control target (AGC) of 300% and a maximum injection time
895 of 45 ms. Fragment spectra were then recorded in profile mode, fragmenting 49 consecutive
896 13.3 m/z windows covering the mass range 350–1,000 m/z and using a resolution of 15,000.
897 Isolated precursors were fragmented in the higher-energy collisional dissociation (HCD) cell
898 using 27% normalised collision energy, a normalised AGC target of 3,000 % and a maximum
899 injection time of 22 ms.
900
901 For the samples acquired in the Orbitrap Astral, peptides were eluted online from the EvoTips
902 using an Evosep One LC system and analysed at 60 samples per day (21-min gradient). The

903 Orbitrap Astral mass spectrometer was operated in positive mode using the DIA mode, at a
904 resolution of 240,000 with a full scan range of 380-980 m/z. The normalised AGC was set to
905 500%. Fragment ion scans were recorded at a maximum injection time of 3.5 ms, 200 windows
906 of 3Th scanning from 380-980 m/z. Ions were fragmented using HCD with 25% normalised
907 collision energy.
908
909 **Mass spectrometry raw data processing and bioinformatic analysis**
910 Raw files were analysed with directDIA workflow in Spectronaut v.19.0[63]. Default settings
911 were used. Data filtering was set to 'Qvalue'. 'Cross run normalisation' was enabled with the
912 strategy of 'local normalisation' based on rows with 'Qvalue complete'. FDR was set to 1% at
913 both the protein and peptide precursor levels. Raw data was searched against the human
914 proteome reference database (including 20,593 entries, Uniprot March 2023). Data was filtered
915 for 75% valid values (in at least 3 out 4 replicates) across WT or Mutant samples (protein
916 groups with >25% missing values were excluded from downstream statistical analysis). Protein
917 intensities were then log2 transformed for downstream statistical and bioinformatics
918 analysis. For each protein, missing values were sampled from a normal distribution with the
919 following parameters: mean = median of all quantified values across all proteins in each sample
920 minus 1.8 times the standard deviation, standard deviation = the standard deviation of all
921 quantified values across all proteins in each sample multiplied by 0.3. For ZNHIT-IP, we did
922 extra z-score normalization to reduce the technical variations. The significance of protein
923 enrichment of Mutants relative to WT was determined through an unpaired two-sided *t*-test
924 corrected by Benjamini–Hochberg at FDR of 0.1 and fold change of 1.5[64,65]. Data analysis and
925 visualisation were done in Jupyter Notebook in the environment of Python 3.9.
926
927 **AP-WB**
928 HeLa cells were transfected with plasmids encoding the specified YFP-tagged baits in a wild-
929 type version or with the indicated mutations using Lipofectamine 2000 24 hours prior to
930 harvesting. Cells were lysed on ice in lysis buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1 mM
931 EDTA, 0.1% NP-40) supplemented with 1 mM DTT and complete protease inhibitor cocktail
932 (Roche) and sonicated for 10 min (30''ON/30''OFF, max) using Bioruptor Plus (Diagenode).
933 Lysates were cleared by centrifugation at 20,000 g at 4°C for 30 minutes, and YFP-tagged baits
934 were affinity purified using GFP-trap beads (ChromoTek). Precipitated protein complexes
935 were washed thrice in lysis buffer and eluted in 2X LDS sample buffer before SDS-PAGE and
936 Western blot analysis with the indicated antibodies (CSTF3: HPA040168 (Sigma), CNOT1:

937 14276-1-AP (proteintech), YFP: GFP in house made antibody to full length GFP). Uncropped
938 WBs for panel 4e and 5i are shown in Fig. S11.
939
940 **Protein production**
941 CSTF3$_{21-549}$ wild-type and GST-WDR33$_{431-485}$ wild-type and mutant were expressed in
942 BL21(DE3) cells overnight at 18 degrees and cells were harvested by centrifugation. For
943 CSTF3 purification cells were resuspended in Buffer L (50 mM NaP pH 7.5; 300 mM NaCl;
944 10 mM imidazole; 10% glycerol; 0.5 mM TCEP; 1xComplete EDTA-free tablets (Roche)) and
945 lysed. Following clarification by centrifugation, lysate was loaded on a 5 ml HiTrap column
946 and His-CSTF3$_{21-549}$ eluted with an imidazole gradient. The peak fractions were pooled, diluted
947 with 50 mM NaP pH 7.5 to 50 mM NaCl and loaded on a SP HP HiTrap column and proteins
948 eluted with a 50mM-1M NaCl salt gradient. Peak fractions were pooled.
949 For purification of GST-WDR33$_{431-485}$, cell pellets were resuspended in 25 mL lysis buffer (50
950 mM Tris-HCl pH 7.4, 300 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol) and sonicated
951 for 10 minutes on ice. Lysates were cleared in a fixed angle rotor for 40 min 145000 rpm and
952 supernatants were incubated with 350 µL pre-equilibrated Glutathione Sepharose Fast Flow
953 (GE healthcare) for 1h. Beads were washed 3 times with 30 mL lysis buffer and used directly
954 for binding experiments.
955
956 **GST-WDR33 pull down assay.**
957 10 µL GST-WDR3$_{431-485}$ wild-type or mutant beads in 200 µL buffer B (150 mM NaCl, 50
958 mM Tris pH 7.4 1 mM DTT) were incubated with 20 µg CTSF3 for 1h at 4C with rotation.
959 Beads were washed 2 times with buffer B and eluted in 25 µL 2x LDS sample buffer (Thermo
960 Fisher Scientific) with 10 mM DTT. For control conditions unbound glutathione beads were
961 used. Data is obtained from three replicates.
962
963 **ITC**
964 Peptides were purchased from Peptide 2.0 Inc (Chantilly. VA, USA). The purity obtained in
965 the synthesis was 95–98% as determined by high performance liquid chromatography (HPLC)
966 and subsequent analysis by mass spectrometry. Prior to ITC experiments both CSTF3$^{21-549}$ and
967 the peptides were extensively dialyzed against 50 mM sodium phosphate pH 7.5, 150 mM
968 NaCl, 0.5 mM TCEP. CSTF3$^{21-549}$ and peptide concentrations were determined using a
969 spectrometer by measuring the absorbance at 280 nm and applying values for the extinction
970 coefficients computed from the corresponding sequences by the ProtParam program

971 (http://web.expasy.org/protparam/). Peptides at approximately 500 µM concentration were
972 loaded into the syringe and titrated into the calorimetric cell containing CSTF3$^{21-549}$ at ~ 40 µM
973 (monomer concentration). The reference cell was filled with distilled water. In all assays, the
974 titration sequence consisted of a single 0.4 µl injection followed by 19 injections, 2 µl each,
975 with 150 s spacing between injections to ensure that the thermal power returns to the baseline
976 before the next injection. The stirring speed was 750 rpm. Control experiments with peptides
977 injected in the sample cell filled with buffer were carried out under the same experimental
978 conditions. These control experiments showed negligible heats of dilution. Experiments were
979 caried out as a single replicate. The heats per injection normalised per mole of the injectant
980 *versus* the molar ratio [Peptide]/[CSTF3$^{21-549}$] were fitted to a single-site model. Data were
981 analysed with MicroCal PEAQ-ITC (version 1.1.0.1262) analysis software (Malvern
982 Instruments Ltd.).
983
984 **AlphaFold predictions**
985 AlphaFold models were generated by submitting full length protein sequences of CPSF1,
986 WDR33, CPSF4, CSTF3 and PAS-RNA (inspired by the human Pre-mRNA 3'-End
987 Processing Machinery structure PDB:6URO) to AlphaFold3[66]
988 (https://alphafoldserver.com) and was ran using standard parameters without templates.
989 Overall predicted template modelling (PTM) and interface predicted template modelling
990 (iPTM) scores were 0.67 and 0.74, respectively. The 5 predictions were further inspected for
991 confidence in ChimeraX[67] by aligning the predictions to the PDB:6URO structure[42] using the
992 *matchmaker* function. Confidence of the CPSF1$_{742-749}$ - CSTF3 interaction was assessed by
993 aligning the models to the FIP1L1$_{1-35}$-CSTF3 crystal structure (PDB: 7ZY4)[38], confirming the
994 same CPSF1-CSTF3 interaction mode for all 5 predictions, which for the motifs was highly
995 similar to that of FIP1L1. For the WDR33$_{441-445}$ motif, confidence of the predictions was
996 confirmed given that all 5 predictions modelled the IPGMG$_{441-445}$ in the same binding pocket
997 of CSTF3 with average atom pLDDT of IPGMG$_{441-445}$ ranging from 64.44-65.57.
998
999 **Data availability**
1000 The proteomic raw data is available at PRIDE ID PXD055750
1001 Username: reviewer_pxd055750@ebi.ac.uk
1002 Password: lyWJc1nnFmbq
1003 All screen sequencing data is available as gRNA count files in supplementary tables.

Code availability

Code will be made available upon request.

Conflict of interest

The authors have no conflict of interest

1010 **Figure legends**

1011

1012

1013 **Figure 1. Design and benchmarking of a SLiM-targeting base editor screen**

1014 **a**, SLiM-mediated interaction exemplified by the p53 MDM2 degron binding to the MDM2
1015 SWIB domain. **b**, Sequence logo (top), peptide sequence (middle) and consensus (bottom) of
1016 the MDM2 degron in p53, showing the evolutionary conservation of key affinity and specificity
1017 determintants (defined positions) and non-conserved flanking residues (wildcard positions
1018 denoted with 'x'). **c**, Composition of the SLiM-targeting base editor library. **d**, Editing
1019 properties of the Cytosine Base Editor (CBE) and Adenine Base Editor (ABE) classes used in
1020 the screen. **e**, Example of a gRNA target DNA sequence with optimal base editing window and
1021 protospacer adjacent motif (PAM) highlighted. **f**, Statistics of predicted mutations introduced
1022 in target SLiMs by gRNAs cloned into ABE and/or CBE. **g**, Schematic of the base editing
1023 screening experimental pipeline. **h**, Receiver Operating Characteristic (ROC) curve based on
1024 the gRNA fold change for the positive and negative control gRNAs across the four base editors.
1025 AUC, area under the curve. **i**, Rank plots of the gRNA $\log_2$ fold change showing the
1026 significantly enriched, depleted, and unchanged gRNAs. TPR: True positive rate. FPR: False
1027 positive rate. **j**, Comparison of the gRNA $\log_2$ fold change in the ABE_{NGG} and CBE_{NGG} screens.

1028

1029 **Figure 2: Identification of SLiMs required for optimal fitness**

1030 **a-c**, Examples of SLiM-containing regions with depleted (red) (**a**), enriched (blue) (**b**) or
1031 unchanged (grey) (**c**) gRNAs. Left panels: Y-axes indicate $\log_2$-fold changes of all gRNAs
1032 targeting the indicated motif summarised for the 4 screens ('All BEs') and individual screens.
1033 Circle size represents limma p-value. Right panels: Y-axes indicate the $\log_2$-fold change of a
1034 given gRNA predicted to target a given amino acid residue within the represented motif
1035 (indicated on the x-axis), with connected dots indicating the same gRNA edit across the
1036 biological triplicate of the ABE_{NGN} screen. In case gRNAs are predicted to target multiple
1037 residues, the log2 fold changes of the guide is shown for each residue. **d**, **e**, Number of SLiMs
1038 targeted by the indicated significantly enriched or depleted gRNAs. **f,** Mann-Whitney p-value
1039 versus $\log_2$ fold change using the most significant p-values for each motif from the four
1040 different screens. Red: *reported motif set*. Grey: *predicted motif set*. Size represents the
1041 percentage of significant gRNAs for each motif. **g**, Subdivision of motifs in panel **f** by ELM
1042 motif classes. **h**, Conservation of motifs categorised as essential versus non-essential. *RLCp*:
1043 SLiMPrint relative local conservation probability[35]. **i**, Log_2-fold change of motifs categorised

1044 by HAP1 gene essentiality. **j**, Top enriched (blue) and depleted (red) reported SLiMs in the
1045 four screens. *Score:* log10 Mann-Whitney p-value, based on fold change, score is negative for
1046 depleted motifs, positive for enriched motifs. Circle color denotes general function of the motif-
1047 containing protein.
1048
1049 **Figure 3: Validation of gRNA performance and editing outcomes**
1050 **a**, Schematic of the base editing sensor setup. **b**, Schematic of the experimental pipeline of the
1051 sensor screen targeting the 714 essential motifs and 8 splice sites per motif-encoding gene in
1052 HAP1 and RPE1 p53-/- cells. **c**, Comparison of sensor editing frequencies in HAP1 and RPE1
1053 cells. **d**, Comparison of SLiM-targeting gRNA $\log_2$ fold changes for the HAP1 and RPE1 cells.
1054 **e**, Comparison of mean $\log_2$ fold changes for gRNA sets targeting splices sites of SLiM-
1055 encoding genes for the HAP1 and RPE1 cells. **f,** Log_2 fold changes of gRNAs classified based
1056 on the ELM resource[5] consensus as *defined* (specific, binding determining residues) or
1057 *wildcard* (flexible positions denoted as 'x') for significantly changing motifs ($\log_{10}$ Mann–
1058 Whitney $p < 0.01$). gRNAs were further subdivided in the 50% highest and 50% lowest sensor
1059 editing efficiencies. **g**, Schematic of the experimental pipeline of the rhAmpSeq screen
1060 targeting the 25 predicted essential motifs with the top scoring gRNA for each, reading out
1061 sensor editing and genomic editing by NGS after pooled rhAmpSeq amplification of the 25
1062 SLiM-containing loci. **h,** Comparison of the predicted or observed protein level edits between
1063 sources of edit information: *predicted*, based on any predicted edit for the ABE optimal editing
1064 window; *sensor*, based on the most common single edit in the sensor; and *genomic*, based on
1065 the most common single edit in the rhAmpSeq amplification of the genomic locus. **i,**
1066 Comparison of the *predicted* edit*, sensor* edit and *genomic* edits at day 0, 3, and 18 for four
1067 targets in the rhAmpSeq screen. Logos show the protein level edit frequency for the *sensor* and
1068 *genomic* data.
1069
1070 **Figure 4: Validation of an essential predicted CSTF3 interaction motif in the cleavage**
1071 **and polyadenylation specificity factor CPSF1**
1072 **a**, Evolutionary conservation of the CPSF1 LYG (DDEEEMLYG$_{741\text{-}749}$) motif and the
1073 CSTF3-binding FIP1L1 motif. **b**, Volcano plot showing mass spectrometry analysis of YFP-
1074 tagged full length CPSF1 WT vs E743A/E744A mutant (Mut) affinity purified from HeLa cells
1075 (n=4). The dashed lines indicate fold-change (1.5) and p-value (0.1) significance cut-offs. **c**,
1076 AlphaFold3 (AF3) prediction modelling the CPSF1-WDR33-CPSF30-RNA-CSTF3

complex (PTM score = 0.67; iPTM score = 0.74). Top panel: CPSF1 is presented as cartoon, the other proteins are represented as surface. Bottom left panel: detail of CPSF1$_{742-749}$ (teal, cartoon representation) binding to CSTF3 (light purple, surface representation). Bottom right panel: detail of the structural alignement of CPSF1-WDR33-CPSF30-RNA-CSTF3 AF3 model with the FIP1L1-CSTF3 structure (PDB: 7ZY4), showing FIP1L1$_{20-28}$ (grey, cartoon representation) binding to the CPSF1-binding pocket of CSTF3 (light purple, surface representation). **d**, ITC binding data of CPSF1$_{737-751}$ wild-type (WT) and mutant peptides to CSTF3$_{21-549}$. **e**, Western blot analysis of full length WT and mutant YFP-CPSF1 affinity purifications from HeLa cells probed for CSTF3 and YFP, representative of two independent experiments.

Figure 5: Characterisation of essential predicted motifs

a, Top essential predicted SLiMs in the four screens. *Score:* log10 Mann-Whitney p-value, based on fold change. Total number of significant gRNAs are indicated. **b**, Comparison of the essential motifs to the specificity determinants of the ELM resource[5] using CompariPSSM[57]. *Similarity p-value*, quantifies the similarity of the binding determinants. *Dissimilarity score*, quantifies key binding determinants with strong dissimilarity. Bright red dots correspond to motif pairs with with similarity scores above the default CompariPSSM cutoff. **c**, Conservation logo of the ELM Adaptin-binding Endosome-Lysosome-Basolateral sorting signals class (TRG_DiLeu_LyEn_5) and the conservation logo of RNF167 motif ERTPLL$_{306-311}$ **d**, Conservation logo of EIF3G motif SWADQVE$_{11-17}$ and *Danio rerio* DBP5 and SLBP5 MIF4G binding motifs. **e**, Volcano plot showing affinity purification coupled to mass spectrometry (AP-MS) analysis of YFP-tagged full length EIF3G WT vs motif mutant (Mut; AAAAAAA$_{11-17}$) purified from HeLa cells (n=4). The dashed lines indicate fold-change (1.5) and p-value (0.1) significance cut-offs. **f,** Summary of AP-MS analysis of YFP-tagged full length bait proteins (except for ANKRD17$_{1950-2603}$ and WDR33$_{1-525}$ where fragments were used) purified from HeLa cells. Top significantly lost binder or, if relevant, top lost complex is shown, comparing WT to motif mutants. n/b, no binder. **g**, Conservation logo of ANKRD17 PIGTER$_{2426-2431}$ motif with surrounding sequence and homologous region of ANKHD1. R2434 mutated in Chopra-Amiel-Gordon syndrome is circled. **h**, Volcano plot of AP-MS for motif-mutant (Mut; AAAAAAA$_{2425-2431}$) versus WT YFP-ANKRD17$_{1950-2603}$, purified from HeLa cells (n=4). The dashed lines indicate fold-change (1.5) and p-value (0.1) significance cut-offs. **i**, Representative Western blot and quantification of affinity purified YFP-ANKRD17$_{1950-2603}$

WT and mutant from HeLa cells, probed for CNOT1 and YFP (n=3, error bars represent mean +/- standard deviation).

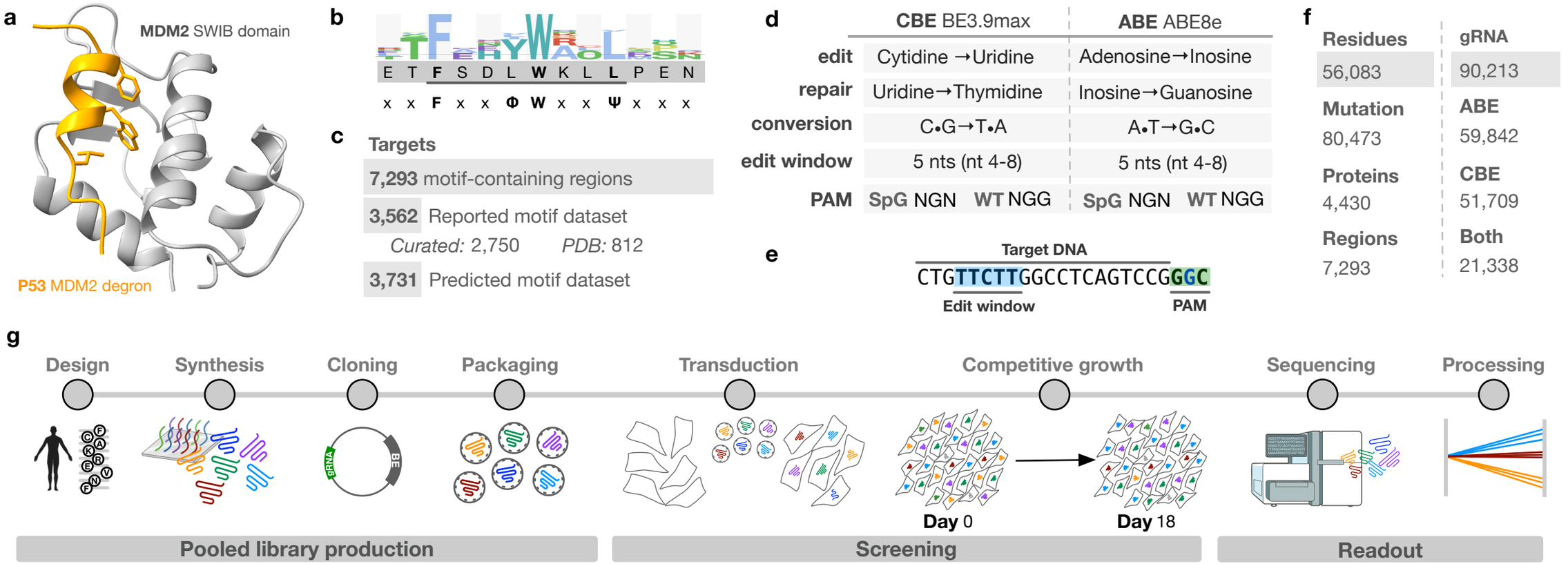
Bibliography

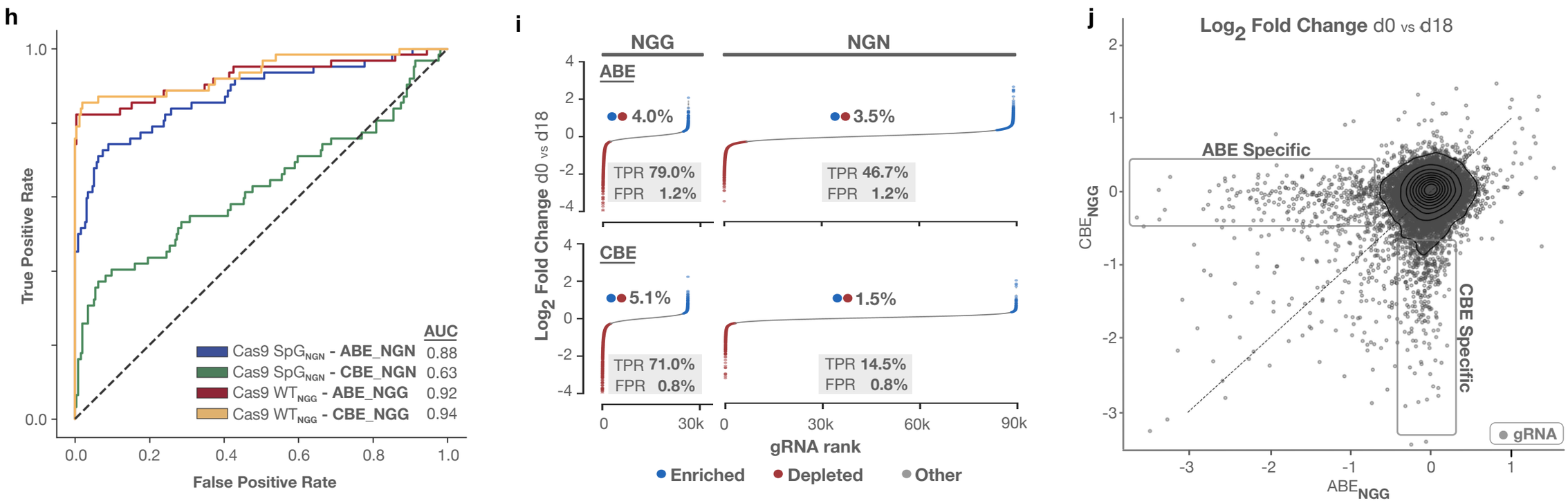
1. Tompa, P., Davey, N.E., Gibson, T.J. & Babu, M.M. A million peptide motifs for the molecular biologist. *Mol Cell* **55**, 161-9 (2014).
2. Van Roey, K. et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev* **114**, 6733-78 (2014).
3. Davey, N.E., Simonetti, L. & Ivarsson, Y. The next wave of interactomics: Mapping the SLiM-based interactions of the intrinsically disordered proteome. *Curr Opin Struct Biol* **80**, 102593 (2023).
4. Davey, N.E. et al. Attributes of short linear motifs. *Mol Biosyst* **8**, 268-81 (2012).
5. Kumar, M. et al. ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids Res* **52**, D442-d455 (2024).
6. Holehouse, A.S. & Kragelund, B.B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol* **25**, 187-211 (2024).
7. Teilum, K., Olsen, J.G. & Kragelund, B.B. On the specificity of protein-protein interactions in the context of disorder. *Biochem J* **478**, 2035-2050 (2021).
8. Davey, N.E., Cyert, M.S. & Moses, A.M. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun Signal* **13**, 43 (2015).
9. Kliche, J. et al. Proteome-scale characterisation of motif-based interactome rewiring by disease mutations. *Mol Syst Biol* **20**, 1025-1048 (2024).
10. Mészáros, B., Kumar, M., Gibson, T.J., Uyar, B. & Dosztányi, Z. Degrons in cancer. *Sci Signal* **10**(2017).
11. Roy, J., Li, H., Hogan, P.G. & Cyert, M.S. A conserved docking site modulates substrate affinity for calcineurin, signaling output, and in vivo function. *Mol Cell* **25**, 889-901 (2007).
12. Lue, N.Z. & Liau, B.B. Base editor screens for in situ mutational scanning at scale. *Mol Cell* **83**, 2167-2187 (2023).
13. Richter, M.F. et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat Biotechnol* **38**, 883-891 (2020).
14. Koblan, L.W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat Biotechnol* **36**, 843-846 (2018).
15. Walton, R.T., Christie, K.A., Whittaker, M.N. & Kleinstiver, B.P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290-296 (2020).
16. Sangree, A.K. et al. Benchmarking of SpCas9 variants enables deeper base editor screens of BRCA1 and BCL2. *Nat Commun* **13**, 1318 (2022).
17. Hanna, R.E. et al. Massively parallel assessment of human variants with base editor screens. *Cell* **184**, 1064-1080.e20 (2021).
18. Ritchie, M.E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

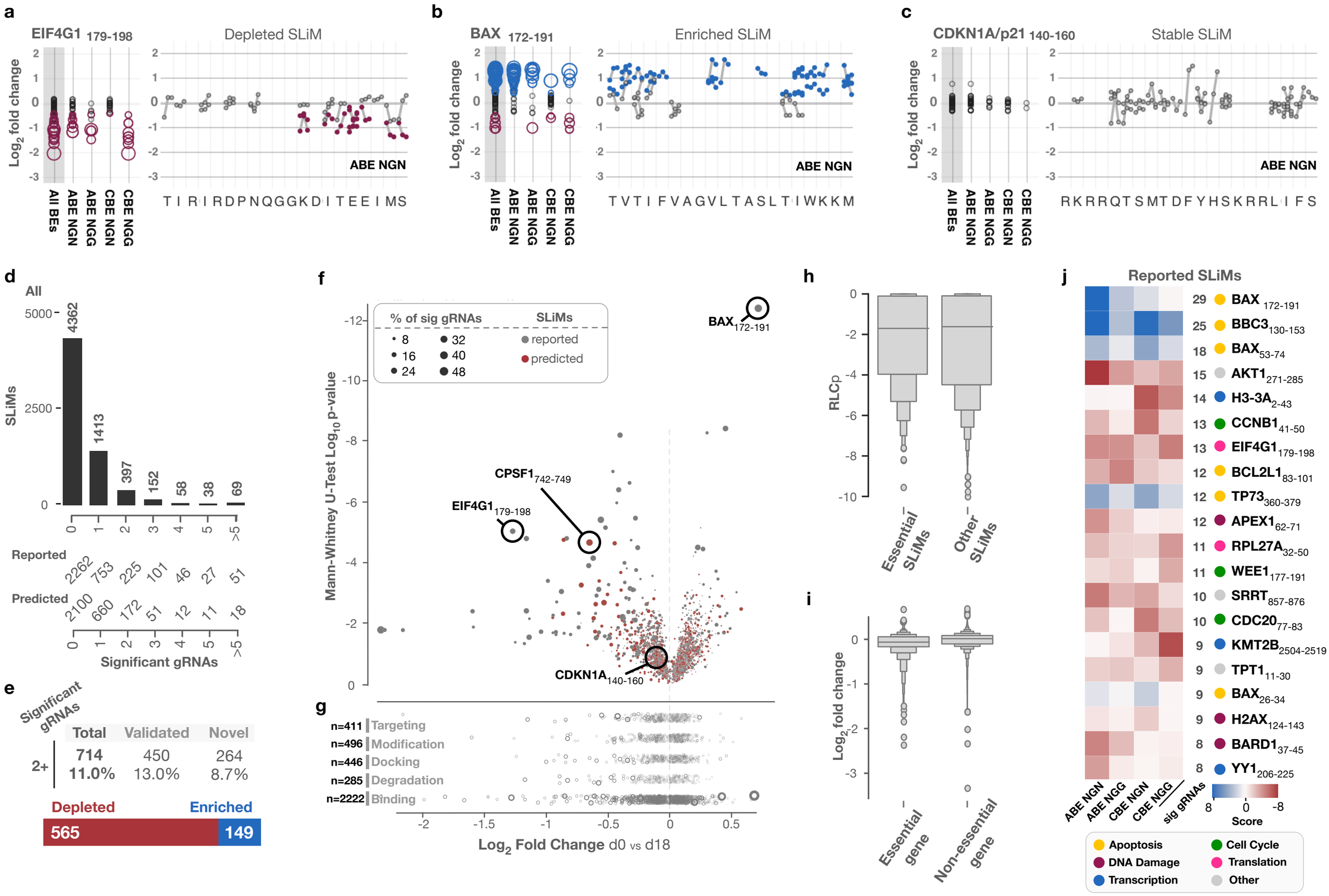
1155 19. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the
1156 human proteome. *Nature* **596**, 590-596 (2021).
1157 20. Blomen, V.A. et al. Gene essentiality and synthetic lethality in haploid human
1158 cells. *Science* **350**, 1092-6 (2015).
1159 21. Safaee, N. et al. Interdomain allostery promotes assembly of the poly(A) mRNA
1160 complex with PABP and eIF4G. *Mol Cell* **48**, 375-86 (2012).
1161 22. Clute, P. & Pines, J. Temporal and spatial control of cyclin B1 destruction in
1162 metaphase. *Nat Cell Biol* **1**, 82-7 (1999).
1163 23. Lee, E.F. et al. The functional differences between pro-survival and pro-
1164 apoptotic B cell lymphoma 2 (Bcl-2) proteins depend on structural differences in
1165 their Bcl-2 homology 3 (BH3) domains. *J Biol Chem* **289**, 36001-17 (2014).
1166 24. Coutandin, D. et al. Conformational stability and activity of p73 require a second
1167 helix in the tetramerization domain. *Cell Death Differ* **16**, 1582-9 (2009).
1168 25. Czabotar, P.E. et al. Mutation to Bax beyond the BH3 domain disrupts
1169 interactions with pro-survival proteins and promotes apoptosis. *J Biol Chem*
1170 **286**, 7123-31 (2011).
1171 26. Suzuki, M., Youle, R.J. & Tjandra, N. Structure of Bax: coregulation of dimer
1172 formation and intracellular localization. *Cell* **103**, 645-54 (2000).
1173 27. Meldal, B.H.M. et al. Complex Portal 2018: extended content and enhanced
1174 visualization tools for macromolecular complexes. *Nucleic Acids Res* **47**, D550-
1175 d558 (2019).
1176 28. Sánchez-Rivera, F.J. et al. Base editing sensor libraries for high-throughput
1177 engineering and functional analysis of cancer-associated single nucleotide
1178 variants. *Nature Biotechnology* **40**, 862-873 (2022).
1179 29. Kim, Y. et al. High-throughput functional evaluation of human cancer-associated
1180 mutations using base editors. *Nat Biotechnol* **40**, 874-884 (2022).
1181 30. Song, M. et al. Sequence-specific prediction of the efficiencies of adenine and
1182 cytosine base editors. *Nat Biotechnol* **38**, 1037-1043 (2020).
1183 31. Ryu, J. et al. Joint genotypic and phenotypic outcome modeling improves base
1184 editing variant effect quantification. *Nat Genet* **56**, 925-937 (2024).
1185 32. Sellés-Baiget, S. et al. Catalytic and noncatalytic functions of DNA polymerase κ
1186 in translesion DNA synthesis. *Nat Struct Mol Biol* **32**, 300-314 (2025).
1187 33. Arafeh, R., Shibue, T., Dempster, J.M., Hahn, W.C. & Vazquez, F. The present and
1188 future of the Cancer Dependency Map. *Nat Rev Cancer* **25**, 59-73 (2025).
1189 34. Chaikovsky, A.C. et al. The AMBRA1 E3 ligase adaptor regulates the stability of
1190 cyclin D. *Nature* **592**, 794-798 (2021).
1191 35. Maiani, E. et al. AMBRA1 regulates cyclin D to guard S-phase entry and genomic
1192 integrity. *Nature* **592**, 799-803 (2021).
1193 36. Simoneschi, D. et al. CRL4(AMBRA1) is a master regulator of D-type cyclins.
1194 *Nature* **592**, 789-793 (2021).
1195 37. Lei, Q.Y. et al. TAZ promotes cell proliferation and epithelial-mesenchymal
1196 transition and is inhibited by the hippo pathway. *Mol Cell Biol* **28**, 2426-36
1197 (2008).
1198 38. Muckenfuss, L.M., Migenda Herranz, A.C., Boneberg, F.M., Clerici, M. & Jinek, M.
1199 Fip1 is a multivalent interaction scaffold for processing factors in human mRNA
1200 3' end biogenesis. *Elife* **11**(2022).

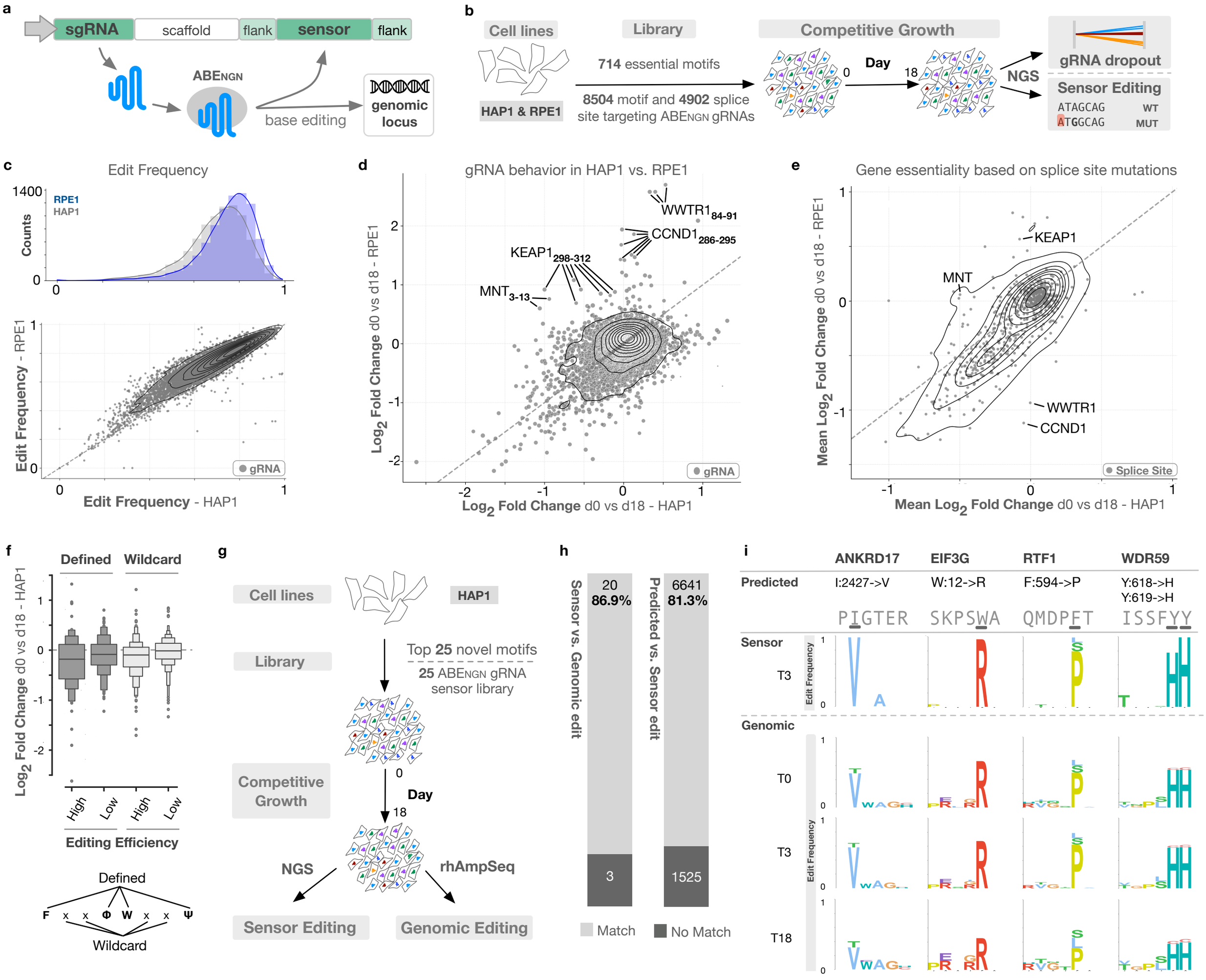
39. Yamazaki, Y. et al. Goliath family E3 ligases regulate the recycling endosome pathway via VAMP3 ubiquitylation. *Embo j* **32**, 524-37 (2013).
40. von Moeller, H. et al. Structural and biochemical studies of SLIP1-SLBP identify DBP5 and eIF3g as SLIP1-binding proteins. *Nucleic Acids Res* **41**, 7960-71 (2013).
41. Udupa, A., Kotha, S.R. & Staller, M.V. Commonly asked questions about transcriptional activation domains. *Curr Opin Struct Biol* **84**, 102732 (2024).
42. Zhang, Y., Sun, Y., Shi, Y., Walz, T. & Tong, L. Structural Insights into the Human Pre-mRNA 3'-End Processing Machinery. *Mol Cell* **77**, 800-809.e6 (2020).
43. Hernández, M.V., Wehrendt, D.P. & Arregui, C.O. The protein tyrosine phosphatase PTP1B is required for efficient delivery of N-cadherin to the cell surface. *Mol Biol Cell* **21**, 1387-97 (2010).
44. Benz, C. et al. Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Mol Syst Biol* **18**, e10584 (2022).
45. Castello, A. et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149**, 1393-1406 (2012).
46. Chopra, M. et al. Heterozygous ANKRD17 loss-of-function variants cause a syndrome with intellectual disability, speech delay, and dysmorphism. *Am J Hum Genet* **108**, 1138-1150 (2021).
47. Boland, A. et al. Structure and assembly of the NOT module of the human CCR4-NOT complex. *Nat Struct Mol Biol* **20**, 1289-97 (2013).
48. Youn, J.Y. et al. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol Cell* **69**, 517-532.e11 (2018).
49. Corbi-Verge, C. & Kim, P.M. Motif mediated protein-protein interactions as drug targets. *Cell Commun Signal* **14**, 8 (2016).
50. Monzon, A.M., Bonato, P., Necci, M., Tosatto, S.C.E. & Piovesan, D. FLIPPER: Predicting and Characterizing Linear Interacting Peptides in the Protein Data Bank. *J Mol Biol* **433**, 166900 (2021).
51. Davey, N.E. et al. SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res* **40**, 10628-41 (2012).
52. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827-39 (2005).
53. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699 (2018).
54. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554 (2014).
55. Kotb, H.M. & Davey, N.E. xProtCAS: A Toolkit for Extracting Conserved Accessible Surfaces from Protein Structures. *Biomolecules* **13**(2023).
56. Mi, H. et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703-721 (2019).
57. Tsitsa, I., Krystkowiak, I. & Davey, N.E. CompariPSSM: a PSSM-PSSM comparison tool for motif-binding determinant analysis. *Bioinformatics* **40**(2024).

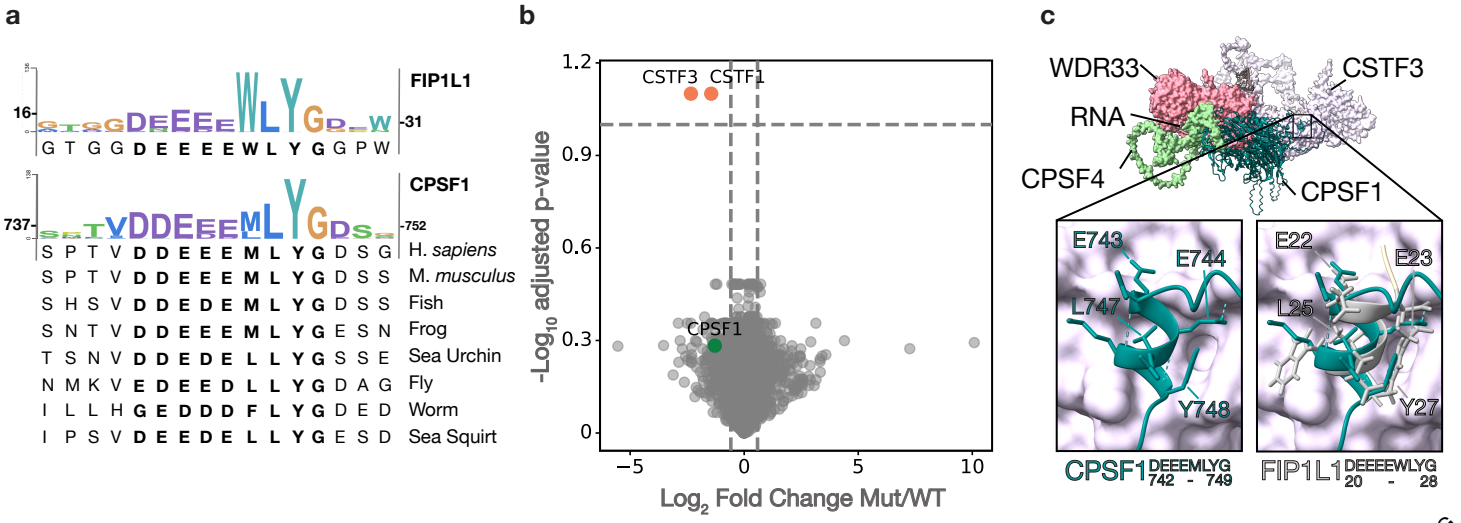
58. Davey, N.E., Edwards, R.J. & Shields, D.C. The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res* **35**, W455-9 (2007).
59. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* **37**, 224-226 (2019).
60. Bache, N. et al. A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. *Mol Cell Proteomics* **17**, 2284-2296 (2018).
61. Heil, L.R. et al. Evaluating the Performance of the Astral Mass Analyzer for Quantitative Proteomics Using Data-Independent Acquisition. *J Proteome Res* **22**, 3290-3300 (2023).
62. Guzman, U.H. et al. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat Biotechnol* **42**, 1855-1866 (2024).
63. Bruderer, R. et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* **14**, 1400-10 (2015).
64. Lin, M.H. et al. Benchmarking differential expression, imputation and quantification methods for proteomics data. *Brief Bioinform* **23**(2022).
65. Dowell, J.A., Wright, L.J., Armstrong, E.A. & Denu, J.M. Benchmarking Quantitative Performance in Label-Free Proteomics. *ACS Omega* **6**, 2494-2504 (2021).
66. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
67. Goddard, T.D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**, 14-25 (2018).


a
MDM2 SWIB domain
P53 MDM2 degron
b
E T F S D L W K L L P E N
x x F x x Φ W x x Ψ x x x
c
Targets
7,293 motif-containing regions
3,562 Reported motif dataset
Curated: 2,750 PDB: 812
3,731 Predicted motif dataset
d
CBE BE3.9max
ABE ABE8e
edit Cytidine →Uridine Adenosine→Inosine
repair Uridine→Thymidine Inosine→Guanosine
conversion C•G→T•A A•T→G•C
edit window 5 nts (nt 4-8) 5 nts (nt 4-8)
PAM SpG NGN WT NGG SpG NGN WT NGG
e
Target DNA
CTGTTCTTGGCCTCAGTCCGGGC
Edit window
PAM
f
Residues 56,083
Mutation 80,473
Proteins 4,430
Regions 7,293
gRNA 90,213
ABE 59,842
CBE 51,709
Both 21,338
g
Design
Synthesis
Cloning
Packaging
Transduction
Competitive growth
Sequencing
Processing
gRNA
BE
Day 0
Day 18
Pooled library production
Screening
Readout


h
True Positive Rate
False Positive Rate
AUC
Cas9 SpG$_{NGN}$ - ABE_NGN 0.88
Cas9 SpG$_{NGN}$ - CBE_NGN 0.63
Cas9 WT$_{NGG}$ - ABE_NGG 0.92
Cas9 WT$_{NGG}$ - CBE_NGG 0.94
i
NGG
NGN
ABE
CBE
4.0%
3.5%
5.1%
1.5%
TPR 79.0% FPR 1.2%
TPR 46.7% FPR 1.2%
TPR 71.0% FPR 0.8%
TPR 14.5% FPR 0.8%
Log$_2$ Fold Change d0 vs d18
gRNA rank
Enriched
Depleted
Other
j
Log$_2$ Fold Change d0 vs d18
ABE Specific
CBE Specific
CBE$_{NGG}$
ABE$_{NGG}$
gRNA


a
EIF4G1 179-198
Depleted SLiM
Log2 fold change
All BEs
ABE NGN
ABE NGG
CBE NGN
CBE NGG
T I R I R D P N Q G G K D I T E E I M S
ABE NGN
b
BAX 172-191
Enriched SLiM
T V T I F V A G V L T A S L T I W K K M
c
CDKN1A/p21 140-160
Stable SLiM
R K R R Q T S M T D F Y H S K R R L I F S
d
All
SLiMs
4362
1413
397
152
58
38
69
Reported
2262 753 225 101 46 27 51
Predicted
2100 660 172 51 12 11 18
Significant gRNAs
e
Significant gRNAs
Total Validated Novel
2+ 714 450 264
11.0% 13.0% 8.7%
Depleted Enriched
565 149
f
% of sig gRNAs
SLiMs
reported
predicted
Mann-Whitney U-Test Log10 p-value
BAX172-191
CPSF1742-749
EIF4G1179-198
CDKN1A140-160
g
n=411 Targeting
n=496 Modification
n=446 Docking
n=285 Degradation
n=2222 Binding
Log2 Fold Change d0 vs d18
h
RLCp
Essential SLiMs
Other SLiMs
i
Log2 fold change
Essential gene
Non-essential gene
j
Reported SLiMs
29 BAX172-191
25 BBC3130-153
18 BAX53-74
15 AKT1271-285
14 H3-3A2-43
13 CCNB141-50
13 EIF4G1179-198
12 BCL2L183-101
12 TP73360-379
12 APEX162-71
11 RPL27A32-50
11 WEE1177-191
10 SRRT857-876
10 CDC2077-83
9 KMT2B2504-2519
9 TPT111-30
9 BAX26-34
9 H2AX124-143
8 BARD137-45
8 YY1206-225
ABE NGN
ABE NGG
CBE NGN
CBE NGG
sig gRNAs
Score
Apoptosis
Cell Cycle
DNA Damage
Translation
Transcription
Other


a
sgRNA
scaffold
flank
sensor
flank
ABENGN
base editing
genomic locus
b
Cell lines
HAP1 & RPE1
Library
714 essential motifs
8504 motif and 4902 splice site targeting ABENGN gRNAs
Competitive Growth
0
Day
18
NGS
gRNA dropout
Sensor Editing
ATAGCAG WT
ATGGCAG MUT
c
Edit Frequency
RPE1
HAP1
Counts
Edit Frequency - RPE1
Edit Frequency - HAP1
gRNA
d
gRNA behavior in HAP1 vs. RPE1
WWTR1 84-91
CCND1 286-295
KEAP1 298-312
MNT 3-13
Log2 Fold Change d0 vs d18 - RPE1
Log2 Fold Change d0 vs d18 - HAP1
gRNA
e
Gene essentiality based on splice site mutations
KEAP1
MNT
WWTR1
CCND1
Mean Log2 Fold Change d0 vs d18 - RPE1
Mean Log2 Fold Change d0 vs d18 - HAP1
Splice Site
f
Defined
Wildcard
Log2 Fold Change d0 vs d18 - HAP1
High
Low
High
Low
Editing Efficiency
Defined
F x x Φ W x x Ψ
Wildcard
g
Cell lines
HAP1
Library
Top 25 novel motifs
25 ABENGN gRNA sensor library
Competitive Growth
0
Day
18
NGS
rhAmpSeq
Sensor Editing
Genomic Editing
h
Sensor vs. Genomic edit
20
86.9%
3
Predicted vs. Sensor edit
6641
81.3%
1525
Match
No Match
i
ANKRD17
EIF3G
RTF1
WDR59
Predicted
I:2427->V
W:12->R
F:594->P
Y:618->H
Y:619->H
PIGTER
SKPSWA
QMDPFT
ISSFYY
Sensor
T3
Genomic
T0
T3
T18
Edit Frequency



d

CSTF3$_{21\text{-}549}$	binding (ITC)
CPSF1$_{737\text{-}751}$	**K_D (μM)**
WT	23.1 ± 2.4
D741G	62.6 ± 16.4
D742G	65.2 ± 18.5
E743G	*n/b*
E743K	*n/b*
E744G	*n/b*
E744K	*n/b*
E745G	62.8 ± 20.0
M746V	96.5 ± 48.1
L747P	*n/b*
Y748H	*n/b*

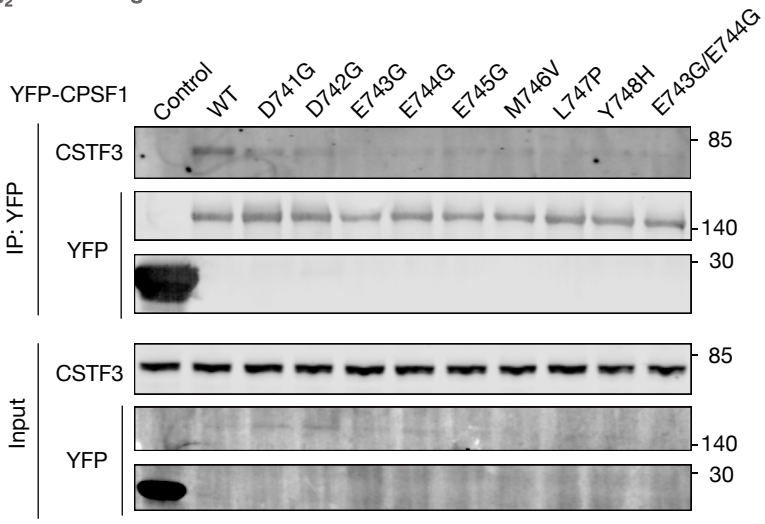
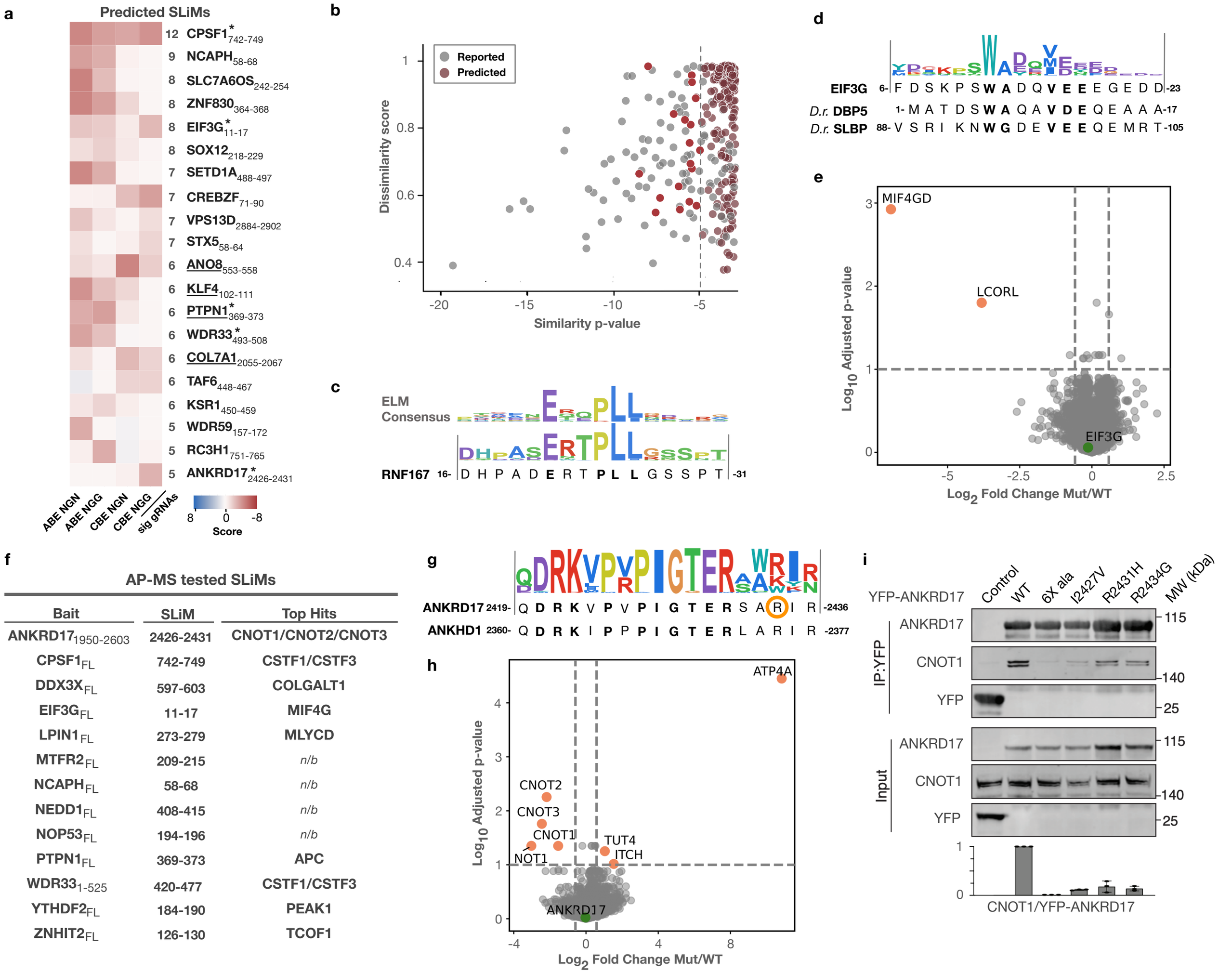


Figure 4


a
Predicted SLiMs
12 CPSF1* 742-749
9 NCAPH 58-68
8 SLC7A6OS 242-254
8 ZNF830 364-368
8 EIF3G* 11-17
8 SOX12 218-229
7 SETD1A 488-497
7 CREBZF 71-90
7 VPS13D 2884-2902
7 STX5 58-64
6 ANO8 553-558
6 KLF4 102-111
6 PTPN1* 369-373
6 WDR33* 493-508
6 COL7A1 2055-2067
6 TAF6 448-467
6 KSR1 450-459
5 WDR59 157-172
5 RC3H1 751-765
5 ANKRD17* 2426-2431
ABE NGN
ABE NGG
CBE NGN
CBE NGG
sig gRNAs
8 0 -8
Score
b
Reported
Predicted
Dissimilarity score
Similarity p-value
c
ELM Consensus
RNF167 16- D H P A D E R T P L L G S S P T -31
d
EIF3G 6- F D S K P S W A D Q V E E E G E D D -23
D.r. DBP5 1- M A T D S W A Q A V D E Q E A A A -17
D.r. SLBP 88- V S R I K N W G D E V E E Q E M R T -105
e
MIF4GD
LCORL
EIF3G
Log10 Adjusted p-value
Log2 Fold Change Mut/WT
f
AP-MS tested SLiMs
Bait | SLiM | Top Hits
ANKRD17 1950-2603 | 2426-2431 | CNOT1/CNOT2/CNOT3
CPSF1 FL | 742-749 | CSTF1/CSTF3
DDX3X FL | 597-603 | COLGALT1
EIF3G FL | 11-17 | MIF4G
LPIN1 FL | 273-279 | MLYCD
MTFR2 FL | 209-215 | n/b
NCAPH FL | 58-68 | n/b
NEDD1 FL | 408-415 | n/b
NOP53 FL | 194-196 | n/b
PTPN1 FL | 369-373 | APC
WDR33 1-525 | 420-477 | CSTF1/CSTF3
YTHDF2 FL | 184-190 | PEAK1
ZNHIT2 FL | 126-130 | TCOF1
g
ANKRD17 2419- Q D R K V P V P I G T E R S A R I R -2436
ANKHD1 2360- Q D R K I P P P I G T E R L A R I R -2377
h
ATP4A
CNOT2
CNOT3
CNOT1
NOT1
TUT4
ITCH
ANKRD17
Log10 Adjusted p-value
Log2 Fold Change Mut/WT
i
YFP-ANKRD17
Control
WT
6X ala
I2427V
R2431H
R2434G
MW (kDa)
IP:YFP
ANKRD17
CNOT1
YFP
Input
ANKRD17
CNOT1
YFP
115
140
25
CNOT1/YFP-ANKRD17