A proteome-wide dependency map of protein interaction motifs

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

[1] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark

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

[3] Equal contribution

[4] Correspondance: Jakob.nilsson@cpr.ku.dk, norman.davey@icr.ac.uk, emil.hertz@cpr.ku.dk

Abstract

Short linear motifs (SLiMs) are the most ubiquitous protein interaction modules in the unstructured regions of the human proteome. Despite their central role in protein function, our understanding of the contribution of SLiMs to cellular homeostasis remains limited. To address this, we designed base editor libraries to precisely mutate all curated SLiMs and a set of computationally predicted instances defined by SLiM-like evolutionary patterns. By targeting 7,293 SLiM containing regions with 80,473 mutations, we define a SLiM dependency map identifying 450 known and 264 predicted SLiMs required for normal cell proliferation. Notably, the vast majority of essential predicted SLiMs belong to novel classes of SLiMs. We also uncover the binding partners of several predicted SLiMs and provide mechanistic insight into disease causing mutations. Our study provides a proteome-wide resource on SLiM essentiality and highlights the presence of numerous uncharacterised essential SLiMs in the human proteome.

Introduction

Understanding the molecular detail of how specific protein complexes form and contribute to cellular functions is a fundamental question in cell biology and forms the basis for interpreting disease causing mutations. Short linear motifs (SLiMs) are an important and undercharacterised class of protein interaction modules, likely constituting the largest uncharted part of the human interactome[1–3]. SLiMs reside in the unstructured regions of proteins as evolutionary conserved islands and are typically encoded in short stretches of ~10 amino acids. Within these sequences, 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 utilised 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 have 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 of these have not been functionally characterised. 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. When in-cell validation is performed, experiments often rely on exogenous expression of proteins, which does not address function at endogenous protein levels. Since the affinities of SLiMs are finely tuned for proper biological function, it is crucial that SLiMs are investigated at endogenous protein concentrations and spatiotemporal expression[7,11]. Thus, most SLiMs are incompletely characterised, leaving a

knowledge gap in our understanding of SLiM contributions to cell function. Clearly, approaches to functionally characterise SLiMs at scale *in situ* would help in the interpretation of the large amount of interaction and disease mutation data, and pinpoint key interactions required for cellular functions. Here, we utilise a recently developed base editor screening approach to perform large-scale, programmable mutagenesis of all curated and thousands of computationally predicted SLiMs in their endogenous cellular context to investigate their requirement for cellular proliferation.

Results

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

We generated a dataset composed of 7,293 SLiM-containing regions in 4,430 proteins, where each region can contain one or more overlapping motifs (Fig. 1B-C, Table S1). The dataset contains all reported motifs (the *reported motif set*) and an extensive set of predicted motifs (the *predicted motif set*). The *reported motif set* consists of 3,562 motif-containing regions comprising a comprehensive set of 2,750 curated motifs from the experimental literature (compiled in the Eukaryotic Linear Motif resource[5] and from manual curation) and 812 motif instances extracted from PDB structures. The *predicted motif set* contains 3,731 protein regions with SLiM-like evolutionary patterns predicted by screening for highly conserved groups of amino acids in rapidly evolving disordered regions (see methods section).

We rationalised that base editors would be ideal to precisely disrupt SLiM-mediated interactions by mutating key residues required for binding, allowing the functional interrogation of SLiMs at scale and at physiological levels of proteins. Base editors are composed of a catalytically impaired Cas protein (typically Cas9 nickase) fused to either a cytosine deaminase (CBE) or deoxyadenosine deaminase (ABE) allowing for gRNA directed mutations to be introduced precisely into the genome [12], with the mutational outcome dictated by the underlying codon sequence. CBEs convert C-G base pairs to T-A while ABEs convert A-T base pairs to G-C within a specific nucleotide "editing window" at a defined distance from the PAM sequence of the gRNA (Fig. 1D-E). In this study, we selected the ABE8e[13] and BE3.9max[14] base editors, both of which target a 5-nucleotide editing window, fused to either SpCas9 nickase (D10A) or the engineered variant SpG nickase[15,16], the latter of which recognises a relaxed NGN PAM sequence allowing for an expanded range of targetable residues in SLiM-containing regions. We will henceforth refer to these base editors as: ABE_{NGG}, $CBE_{NGG,}$ ABE_{NGN}, and CBE_{NGN}.

A set of gRNAs was designed to target each motif with the four base editors. For the prediction of non-synonymous mutations we assumed full editing within the "editing window" spanning nucleotides 4-8 in the target DNA of the gRNA and retained only gRNAs predicted to result in protein level mutations. The resulting gRNA library consisted of 90,213 gRNAs targeting 80,473 unique mutations to 56,083 residues in the 7,293 target motif regions (Fig. 1F, Table S2). The library also contained 62 previously described positive control gRNAs targeting splice sites of essential genes and 255 negative controls including gRNAs targeting intergenic regions of the genome and non-targeting gRNAs that do not map to the genome[16]. Only gRNAs targeting NGG PAMs were cloned into lentiviral vectors encoding SpCas9-base editor fusions (ABE_{NGG} and $CBE_{NGG,}$) while all gRNAs targeting NGN or NGG PAMs were cloned into lentiviral vectors encoding SpG-base editor fusions (ABE_{NGN} and CBE_{NGN}) to produce four lentiviral plasmid libraries.

Design and benchmarking of the SLiM dependency screen.

HAP1, a near-haploid human cell line derived from a chronic myelogenous leukaemia (CML) cell line, has shown superior performance in previous base editor screens[16,17], and was chosen as the cell model to interrogate SLiM dependency. HAP1 cells were transduced in biological triplicates with the lentiviral libraries at low multiplicity of infection and high coverage (>675-fold). The gRNA abundance in genomic DNA at day 0, and after 18 or 21 days of competitive growth was quantified by next generation sequencing (Fig. 1G). The fold change in gRNA abundance comparing early and late time points was quantified along with the significance of that change using the limma gene expression analysis package[18] (Table S3)

Positive and negative control gRNAs were examined to assess the discriminatory power of each base editor to identify functionally relevant edits. We observed substantial depletion of positive control gRNAs compared to negative control gRNAs (Fig. S1A). However, the CBE_{NGN} base editor displayed a notably weaker effect, consistent with observations by others[16], which was reflected in its lower discriminatory power (Fig. 1H, CBE_{NGN} AUC=0.66 vs. ABE_{NGN} AUC=0.89). Base editors targeting the NGG PAM outperformed those targeting the NGN PAM (ABE_{NGN} AUC=0.92, CBE_{NGG} AUC=0.95) . The discriminatory power of the day 18 and day 21 screens were highly similar so we focused on day 18 in the following analyses (Fig. S1B).

The majority of motif targeting gRNAs showed minimal to no effect in the screen (Fig. 1I); however, a small proportion exhibited significant and replicated dropout (ABE_{NGN} 3.5%, ABE_{NGG} 4.0%, CBE_{NGN} 1.5%, and CBE_{NGG} 5.1%). A strong correlation was observed between the gRNAs showing significant changes in both NGN and NGG screens for the same editor (Fig. S1C). As anticipated, given the different targeting and editing outcomes of the CBE and ABE, we observed distinct effects for each editor, advocating the use of both for deeper mutational coverage (Fig. 1J). In summary, the screen had strong discriminatory power for gRNAs targeting regions of the genome that are required for normal cell proliferation and showed dropout of a subset of the motif-targeting gRNAs.

Defining a SLiM dependency map

Next, we defined a SLiM dependency map by analysing the gRNA dropout data in a motif-centric manner (Table S3-4). We hypothesised that motifs with multiple significant gRNAs are more likely to be required for cellular fitness. In accordance with this, we observed that 62.3% of the significantly changing gRNAs were found in only 11.5% of the motif-containing regions suggesting the dropout of a subset of motifs. We further filtered the predicted motifs using AlphaFold[19] structural information to more stringently enrich for motifs residing in intrinsically unstructured regions. Next, we categorised the remaining 6,489 regions into three classes of motifs: depleted motifs, which result in a proliferation defect upon mutation; enriched motifs, which provide a proliferation advantage upon mutation; and stable motifs, which show no observable effect upon mutation (examples in Fig. 2A-C respectively). Overall, across the four screens, we identified 714 motif-containing regions in 626 proteins with at least 2 unique significant gRNAs, while the majority of motif-containing regions (4,362; 67.2%) were targeted by gRNAs that remained stable through the screening (Fig. 2D). For regions where mutations had an impact on cell viability (for brevity, we will refer to these regions as “essential” motifs); 565 led to a decrease and 149 to an increase in cell proliferation (Fig. 2E). Of the essential motifs, 450 were from the *reported motif set* while 264 were from the *predicted motif set* (Fig. 2E). We applied three metrics to quantify essentiality and prioritise motifs by confidence: the number of unique significant gRNAs per motif; the mean fold change of the gRNAs for each motif; and the distribution of these fold changes against the distribution of all gRNAs in the screen (Fig. 2F, Table S4). This data is made available through an interactive web resource for the community to explore SLiM function (https://slim.icr.ac.uk/projects/base_editing) (Figure S2).

We next categorised the *reported motif set* into high-level functionalities (Fig. 2G) and observed a small but significant increase in essentiality for the binding (7.4%) and degradation (8.0%) classes of motifs compared to the modification, localisation, and docking motif classes (5.8%, 6.0%, and 5.8%, respectively). For the essential motifs with known consensus describing the motif binding determinants, gRNAs targeting key defined positions (Fixed positions) required for binding were not found to be more deleterious than other positions in the region (Mann-Whitney p-value 0.37), yet there is a enrichment of the most deleterious mutations in defined positions (Fig. 2H). We noted that many highly conserved motifs are not essential for cell proliferation and, contrary to expectation, that highly conserved motifs are not more likely to be essential (Mann-Whitney p-value 0.53, Fig. 2I). Finally, as expected, when comparing motif-containing region essentiality to protein essentiality defined by HAP1 gene knockout screens[20], essential regions are more likely to occur within essential proteins (Mann-Whitney p-value 1.69×10^{-27}, Fig. 2J). However, numerous essential motifs are present in proteins that are not essential including the instance in Phosphatidate phosphatase LPIN1 (LPIN1) $WLWGELP_{273\text{-}279}$ and Tyrosine-protein phosphatase non-receptor type 1 (PTPN1) $EVRSR_{369\text{-}373}$ investigated later.

Several of the motifs from the *reported motif set* with the highest number of significant gRNAs in the screen are in proteins with strong links to cell viability (Fig. 2K). These motifs include the polyadenylate-binding protein-1 (PABP) binding motif in eIF4G required for the expression of polyadenylated mRNAs[21] and the D box APC/C degradation motif of Cyclin B1 (CCNB1) required for mitotic exit[22]. We also observed strong effects for several apoptosis related SLiMs. Mutations in motifs with anti-apoptotic functions in many cases resulted in reduced cell proliferation, for example, mutations to motifs in Apoptosis regulator Bcl-2 family members BCL2L1[23]. Conversely, the BH3 motif in Apoptosis regulator BAX (BAX) and tetramerisation motif in the Tumor protein p73 (TP73) both of which have pro-apoptotic functions resulted in a proliferative advantage[24–26]. A Reactome pathway analysis of the proteins containing essential motifs from the *reported motif set* reflected these observations and showed an enrichment of essential motifs in pathways related to transcription, RNA processing, cell cycle and apoptosis in line with the expectation that mutations to motifs in proteins with core cellular functions are more likely to have an impact on cell proliferation (Fig. S3A, Table S5). When the instances from the *predicted motif set* were analysed, the functional role of proteins harbouring the predicted motifs were more diverse with no enrichment in any Reactome pathway. Investigating enriched motifs on the protein complex

level, we observed numerous motifs linking core components of key complexes as well as recruiting non-constitutive components to the complex (Table S6). Three complexes stood out due to the number of essential motifs with five each: the Mitotic Checkpoint Complex (motifs in BUBR1 and CDC20) (Fig. S3B); Histone-lysine N-methyltransferase complex SET1A variant (motifs in SETD1A, RBBP5 and ASH2L) (Fig. S3C) and pre-mRNA cleavage and polyadenylation specificity factor complex (motifs in CPSF1, FIP1L1 and WDR33) (Fig. S3D). The latter two are of particular interest as they include several uncharacterised, yet essential motifs.

In summary, the screens defined a dependency map of the SLiM-mediated interactome, which correctly reidentified motifs known to positively or negatively affect cell proliferation in fundamental cellular pathways, and identified 264 uncharacterised motifs that are essential for cell viability.

Base editing of SLiMs links mutagenesis to loss of binding.

To verify that introduction of the predicted edits by the base editor is causal for the observed effects on cellular fitness, we focused on the top scoring predicted motif, $DDEEEMLYG_{741-749}$ in CPSF1 (Fig. 2F, Fig. S3D, see also Fig. 4A). This motif shares significant similarity with the CSTF3-binding motif in the FIP1L1 motif[27], that is also a strongly depleted motif in our screen, particularly at a core LYG and N-terminal acidic residues (Fig. 3A). We will refer to it as the LYG motif from here on. To explore this, we developed an affinity purification coupled with mass spectrometry (AP-MS) pipeline. YFP-tagged CPSF1 WT and E743A/E744A were affinity purified from HeLa cells and a quantitative interactome comparison was performed using mass spectrometry-based proteomics. This revealed that the CPSF1 LYG motif is required for binding to Cleavage Stimulation Factor complex subunits (CSTF1 and CSTF3), confirming the motif as a novel CSTF3-binding motif (Fig. 3B, Fig. S4A, Table S7). Consistent with this, an AF3 prediction modelling the full length version of the CPSF1-WDR33-CPSF30-RNA-CSTF3 complex positions the LYG motif of CPSF1 into to the SLiM binding site on CSTF3 that was reported to bind the FIP1L1 motif[27](Fig. 3C).

We performed deep sequencing of the genomic locus encoding the LYG motif in CPSF1 from cells subjected to adenine and cytidine base editing with individual gRNAs targeting that motif. The ABE_{NGN} gRNAs introduced mutations with high efficiency (up to 61% allelic frequency), matched the expected edits and showed limited bystander edits (Fig. 3D-E and

S5-6, Table S8 and Data S1). The CBE_{NGN} gRNAs also introduced the predicted mutations, though at lower efficiency than the ABE_{NGN}, displaying up to 10% indel generation and evident bystander editing, which is in line with findings of others[14] (Fig. 3D-E and S5-6, Table S8 and Data S1). The gRNAs that were dropping out in the proteome-wide proliferation screen introduced specific mutations in the LYG motif which depleted from the cell population over time, suggesting a loss of function that impacts cellular fitness. Similarly, gRNAs that did not drop out in the proteome-wide screen introduced mutations that remained stable over time. To directly link the observed effects on cellular fitness of the base editor induced mutations in CPSF1 to the loss of binding to CSTF3, we tested the impact of 10 predicted single amino acid mutations achievable by ABE_{NGN} editing. ITC affinity measurements of CPSF1 peptides to recombinant CSTF3 revealed reduced binding of all mutants compared to wild type (Fig. 3F and S7, Table S9). These data were consistent with co-immunopurification of CSTF3 with wildtype and mutant full length CPSF1 in cells (Fig. 3G). The affinity measurements revealed that the ABE mutations introduced at E743, E744, L747 and Y748 had the largest impact on binding, consistent with the key buried residues in the reported structure of the FIP1L1 motif bound to CSTF3[27].

Collectively, these data demonstrate that single amino acid mutations introduced into a SLiM by base editing are sufficient to disrupt function, and that the proteome-wide screening readout is robust and reflects the observation in single gRNA base editing validation experiments.

Essential predicted motif instances largely represent novel motif families

We next investigated the 264 essential predicted motifs to determine if they belong to known or novel motif families (Fig. 4A for top scoring motifs and Table S4). We first compared these regions to the binding determinants of previously characterised domain-binding motifs, revealing 39 predicted instances with significant similarity to existing motif families (Fig. 4B, Table S10). The majority of these examples belong to well characterised motif classes. For example, the $ERTPLL_{306\text{-}311}$ motif in the E3 ubiquitin-protein ligase RNF167 (RNF167) strongly resembles an Adaptin binding Endosome-Lysosome-Basolateral sorting signal (Fig. 4C), aligning with the role of RNF167 regulating endosome trafficking[28]. Another example is the $SWADQVE_{11\text{-}17}$ motif in Eukaryotic translation initiation factor 3 subunit G (EIF3G), which had significant similarity to the binding determinants of a motif binding to MIF4GD. The motif class had previously only been studied in fish, with tested MIF4G domain-binding

motifs in *Danio rerio* Stem-loop binding protein (SLBP) and mRNA-export factor DBP5 (DBP5), therefore, the predicted EIF3G instance would extend the taxonomic range of the motif to human proteins (Fig. 4D)[29]. Consistent with this, AP-MS analysis showed strong and specific binding to human MIF4G which was lost upon mutation of the SLiM (Fig. 4E). We then compared the binding determinants of the 264 essential predicted motifs to each other and observed only 16 instances with similarity to other predicted motifs. Those that shared binding determinants consisted of regions in paralogous proteins, proline-rich motifs, and a cluster of transcription factors that shared hydrophobic motifs, likely to be a transactivation domain (Table S10).

The binding determinant analysis indicated that 225 predicted essential motifs showed no significant similarity to any known motif class, suggesting that the majority of the predicted motifs do not belong to existing motif families (Fig. 4B, Table S10). To explore this, we expanded our AP-MS approach to an additional 11 baits which uncovered the putative binding partner for several of these (Fig. 4F, Fig. S8, Table S7). In WDR33, we uncovered an essential IPG motif and inspection of this region in WDR33 revealed two conserved copies of this consensus and in human WDR33 an additional $LPGM_{420\text{-}423}$ motif (Fig. S4B) that when mutated resulted in the loss of binding to the Cleavage Stimulation Factor subunits (CSTF1 and CSTF3) (Fig. S8). Direct binding to CSTF3 was cross-validated with a WDR33 fragment, $WDR33_{431\text{-}485}$, covering both IPG motifs (Fig. S4C). This unstructured region of WDR33 would be in close proximity to CSTF3 based on the structure of the CPSF1-WDR33-CPSF30-RNA-CSTF3 complex (URO6) which contains the domain directly N-terminal of the IPG motifs[30] and our results suggest that the IPG motifs help stabilise the interaction. Another interesting observation was the loss of binding between the tyrosine phosphatase PTPN1 and Adenomatous Polyposis Coli (APC) protein upon mutation of the $EVRSR_{369\text{-}373}$ motif in PTPN1 (Fig. S8, Table S7). This suggests that the actions of PTPN1 and APC in regulating β−catenin might be coordinated[31] and provides an example of an essential motif in a nonessential protein.

Disease causing mutations in ANKRD17 disrupt SLiM binding to CCR4-NOT complex

We next focused on whether we could obtain novel mechanistic insight into disease-causing mutations potentially affecting SLiM function. Of the 6,489 motif-containing regions, 105 had reported pathogenic disease mutations within the motif or in the immediate flanking regions (3 amino acids on either side) (Table S11). We focused our attention on a PIG motif

present in the RNA binding protein Ankyrin repeat and KH domain containing 17 (ANKRD17), representing a novel uncharacterised motif family (Fig. 4G). An additional instance of the motif in ANKRD17 paralogue 1 (ANKHD1), mapped in the specificity determinant analysis, was also found to be essential in our screen. In ANKRD17, a disease mutation of unknown function just C-terminal to the PIG motif, PIGTERSA**R**$_{2426\text{-}2434}$ (R to G mutation) was present. This mutation is linked to Chopra-Amiel-Gordon syndrome (CAGS), a neurodevelopmental syndrome caused by heterozygous loss of ANKRD17 function[32]. Using our AP-MS approach, we found that upon mutation of the PIG motif in ANKRD17 binding to CNOT1, CNOT2, and CNOT3 was lost, suggesting binding to the NOT module of the CCR4-NOT deadenylase complex involved in mRNA regulation[33] (Fig. 4H). We validated that the expected amino acid mutations introduced by the base editors in ANKRD17 disrupted the binding to CNOT1 (Fig. 4I). Importantly, we also confirmed that the R2434G disease mutation disrupted binding to CNOT1, and we speculate that the defect in ANKRD17 function in patients with this mutation is due to deregulation of specific mRNAs bound by ANKRD17 that are not channelled to the CCR4-NOT complex[32].

Discussion

SLiMs constitute one of the most abundant regulatory elements in unstructured regions of the human proteome and can readily arise *ex nihilo* or be disrupted through disease mutations. Until now, our understanding of SLiM contribution to cell function has been limited due to lack of efficient endogenous functional screening methods. In this study, we utilise base editor screening to create a comprehensive map of cellular SLiM dependency and show that mutations introduced by base editors are sufficient to disrupt SLiM mediated interactions, allowing us to couple disruption of interactions to cellular fitness effects. Our screens reveal that numerous expected, and unexpected, motifs are required for optimal cell viability, and by defining a set of essential SLiMs we provide a priority list for future studies, as well as gRNAs to specifically target these regions. We provide a user-friendly web-based resource (https://slim.icr.ac.uk/projects/base_editing) that allows for interactive exploration of our comprehensive dataset, thereby providing a valuable tool for the motif research community. While we are confident about the motifs contributing to fitness in our screen, we are aware that multiple motifs that did not score could have been undetected due to poor editing of the target locus. Importantly however, the interpretation of our data is not confounded by structural considerations as we target SLiMs in unstructured regions.

Our screens reveal cellular dependency on both known and predicted motifs, exemplified by the numerous motifs identified in the polyA machinery. Affinity measurements of CPSF1 single-point mutations that were depleted in our screens uncover how tightly SLiM affinities are tuned to function and show that small perturbations of protein affinities can be detected in base editor screens. Interestingly, we identify essential SLiMs in proteins not reported to be essential in HAP1 cells highlighting a concept where SLiMs in nonessential proteins can be required for cellular function and proliferation. This emphasises the need for higher resolution fitness maps in addition to gene knockout data to understand protein function and impact of disease causing mutations.

In conclusion, our work highlights that base editing technologies can be applied to discover novel binding regions and prioritise interaction interfaces based on their importance to cell viability. Our approach provides a blueprint for future efforts to dissect SLiM function in different cell types and during cell perturbation.

Acknowledgments

Work at the Novo Nordisk Foundation Center for Protein Research is supported by NNF14CC0001. Work in the Nilsson lab is supported by grants from the Novo Nordisk Foundation (NNF0082227 and NNF0065098) and the Danish Cancer Society (R269-A15586-B71). N.E.D. is funded by a Cancer Research UK Senior Cancer Research Fellowship (C68484/A28159). E.P.T.H. is funded by a Lundbeck Foundation Postdoc Grant (R380-2021-1221). We thank the protein production and characterisation facility at Novo Nordisk Foundation Center for Protein Research for all their help with the project. We thank the Genomics Platform at Novo Nordisk Foundation Center for Protein Research and Center for Stem Cell Medicine for technical support and use of instruments. NGS data processing and analysis were performed using the DeiC National Life Science Supercomputer at Technical University of Denmark (www.computerome.dk).

Methods

SLiM library design

SLiM-containing regions were collected from three sources: manually curated motif information and motif-containing structure processing to retrieve experimentally validated motifs; and evolutionary footprinting to define putative motifs. *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[34]; and extract 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[35]. 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 [36,37].

gRNA design

gRNAs were designed to target protein level of the motifs in the SLiM library based on the PAM (NGG or NGN), edit 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, and the CDS (CoDing Sequence) of the transcript, 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, 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, the longest Ensembl transcript was retrieved, the target peptide

region was mapped to the longest translated Ensembl transcript and the target positions are adjusted to account for the remapping. Next, the Ensembl CDS and genomic sequence were aligned and the target genomic sequence nucleotides for editing were determined through the translated Ensembl CDS. This step allows for editing of nucleotide flanking intron-exon boundaries. Potential PAM sites upstream and downstream of the target genomic sequence nucleotides on both strands were identified. PAM sites that positioned the edit window of the base editor over the target genomic sequence nucleotides were investigated further. The edit window was searched for nucleotides corresponding to the preference of the base editor to find potential nucleotide edits. If one or more editable nucleotides were available for a window all edits were made simultaneously *in silico* and the resulting protein-coding changes were compared to the wildtype protein sequence. gRNAs predicted to result in non-synonymous mutations were retained as part of the motif dependency map gRNA library.

Library cloning

Each gRNA in the library was extended to create a pool of oligonucleotide sequences (83 nt) following a previously published design (Sangree et al., 2022, Nat Comms). The oligonucleotide consisted of primer sites for differential amplification flanking appropriate overhang sequences with Esp3I restriction enzyme recognition sites and the gRNA: 5'-[Forward primer, 20 nt]CGTCTCACACCG[sgRNA, 20 nt]GTTTCGAGACG[Reverse Primer, 20 nt]. The oligonucleotide pool was synthesised by GenScript (CustomArray). Individual subpools were amplified in a reaction of 25 uL 2x NEBnext PCR master mix (New England Biolabs), 1 uL oligonucleotide pool (10 ng/uL), 2.5 uL each of forward and reverse primer at a final concentration of 0.5 uM, and 19 uL water, using a unique primer combination for each subpool (Supplementary Table S12).

The number of PCR reactions per subpool was determined by scaling to 24 reactions per 92,000 gRNAs. PCR cycling conditions were: (1) 98 °C for 30 s; (2) 53 °C for 30 s; (3) 72 °C for 30 s; (4) go to (1) x 19; (5) 72 °C for 60 s; (6) 10 °C hold. Amplicons were PCR-purified (NucleoSpin Gel and PCR Clean-up, Macherey-Nagel). Library vector was pre-digested with FastDigest Esp3I (Thermo Fisher Scientific) and gel-purified (NucleoSpin Gel and PCR Clean-up, Macherey-Nagel). The following library vectors were used: pRDA_256 (BE3.9Max-SpCas9), pRDA_478 (BE3.9Max-SpG), pRDA_426 (ABE8e-SpCas9) and pRDA_479 (ABE8e-SpG). pRDA_256 (Addgene #158581), pRDA_478 (Addgene #179096), pRDA_426 (Addgene #179097), pRDA_479 (Addgene #179099) were gifts from

John Doench & David Root. Amplicons were cloned into library vectors by Golden Gate reactions containing 25 uL 2x Rapid Ligase Buffer (Qiagen Enzymatics), 0.25 uL Bovine Serum Albumin (20 mg/mL) (New England Biolabs), 1 uL FastDigest Esp3I (1 U/uL) (Thermo Fisher Scientific), 0.25 uL T7 Ligase (3x10^6 U/mL) (Qiagen Enzymatics), 15 ng purified amplicon, 100 ng digested vector, and water in a total volume of 50 uL. The number of Golden Gate reactions performed per subpool was determined by scaling to 56 reactions per 92,000 gRNAs. Golden Gate reaction conditions were: (1) 37 °C for 5 min; (2) 20 °C for 5 min; (3) go to (1) x 24; (4) 10 °C hold. Ligated plasmid product was purified first by PCR-purification (NucleoSpin Gel and PCR Clean-up, Macherey-Nagel) and then by isopropanol precipitation in the following reaction: 300 uL PCR-purified plasmid, 300 uL isopropanol, 3 uL GlycoBlue Coprecipitant (final conc. 75 ng/uL) (Thermo Fisher Scientific), 50 mM NaCl. Samples were vortexed (10 s full speed) and incubated 15 min at RT before centrifugation at 16,000xg for 15 min at RT. Precipitated DNA was washed twice with ice-cold 80% Ethanol, air-dried and resuspended in TE buffer (1 uL TE buffer per Golden Gate reaction) at 55 °C for 10 min. The number of electroporations per subpool was determined by scaling to 16 electroporations per 92,000 gRNAs. The purified plasmid library was electroporated into Endura Electrocompetent cells (Lucigen) and grown at 30 °C for 16 h on agar with 100 ug/mL carbenicillin. Colonies were scraped and plasmid DNA was prepared (NucleoBond Xtra Maxi EF, Macherey-Nagel). To confirm library representation and distribution, inserts were amplified from plasmid library in a reaction of 25 uL 2x NEBnext PCR master mix (New England Biolabs), 1 uL plasmid library (10 ng/uL), 2.5 uL each of unique i5 and i7 primers at a final concentration of 0.5 uM, and 19 uL water, using unique primer combinations (Supplementary Table 12). PCR cycling conditions were: (1) 98 °C for 30 s; (2) 98 °C for 10 s; (3) 55 °C for 30 s; (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 gel-purified and then re-purified by PCR-purification (NucleoSpin Gel and PCR Clean-up, Macherey-Nagel). Samples were sequenced on a NextSeq2000 (Illumina) with a 2% spike-in of PhiX with 22 dark cycles. Fastq files were generated using bcl2fastq v.2.19.1 and reads were trimmed to 20 bp using cutadapt 1.18, removing a variable number of basepairs at the start and end depending on the size of the primer stagger. MAGeCK 0.5.8 was used to assign the trimmed reads to the guides in the library and create the count matrix.

Virus production

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

Competitive Growth Screen

HAP1 cells (#C859, Horizon Discovery, SNB: 35369) were cultured in IMDM GlutaMax (Gibco) supplemented with 10% FBS and 100 U/mL penicillin–streptomycin and passaged every three days by trypsinization in 0.05% trypsin-EDTA (Gibco). Each screen was performed as a triplicate with three separate transductions at a coverage above 675-fold sgRNA representation, which was maintained throughout the screen. HAP1 cells were transduced with the indicated lentiviral libraries at a low multiplicity of infection (MOI; 0.3-0.4) by treating cells with 8 µg/mL polybrene and lentiviral supernatants for 24 hours. Viral titers required to achieve an MOI of ~0.3 were pre-determined in a titration experiment. Transduced cells were selected by treatment with 1.5 µg/mL puromycin for 48 hours. After selection, cells were cultured for 21 days. Genomic DNA was extracted from cell pellets harvested after completion of selection, which we consider the start of the screen (t0), and after 18 (t18), and 21(t21) days of culturing. ~600 bp plasmid library fragments containing the sgRNA sequences were amplified from the genomic DNA by PCR using NEBNext Ultra II Q5 Master Mix (New England Biolabs) with the LCV2_forward and LCV2_reverse primers (Supplementary Table S12). A second PCR reaction introduced i5 and i7 multiplexing barcodes (Supplementary Table S12), whereafter ~220 bp PCR products were gel-purifed and sequenced on Illumina NextSeq2000 with a 2% spike-in of PhiX with 22 dark cycles. Fastq files were generated using bcl2fastq v.2.19.1 and reads were trimmed to 20 bp using cutadapt 1.18, removing a variable number of basepairs at the start and end depending on the size of the primer stagger. MAGeCK 0.5.8 was used to assign the trimmed reads to the guides in the library and create the count matrix.

Data processing

The data for 90,213 gRNA were processed to define significantly changing gRNAs using a limma statistical analysis [18]. Next generation sequencing data from the day 0, 18 and 21 samples was converted to gRNA sequencing counts by mapping to motif-targeting and control gRNAs (essential splice sites, non-targeting , intergenic) in the motif dependency map gRNA library design. Three replicates were collected per time point, resulting in three data points per gRNA per time point. Low-abundance gRNAs were removed (counts <30). The vast majority of gRNAs had a single exact match in the genome, however, non-specific gRNAs with off-target genome binding sites showed increased dropout correlated with the number of off-target binding sites. Consequently, gRNAs with 5 or more off-target matches in the genome were removed from downstream analysis, leading to the exclusion of 483 gRNAs. Raw sequencing counts were normalised per sample to log2 transcripts per million (log2TPM). The log2TPM values were compared using limma, resulting in fold change (FC) and p-values for each gRNA. These p-values were compared to the positive and negative controls, and a p-value cut-off of 0.01 showed a false positive rate of ~1% across all four screens, with a true positive rate of up to 79%. Consequently, a p-value of less than 0.01 was set as the threshold considered to have an effect on cell proliferation. Next, the gRNAs mapping to each motif-containing region were analysed to define groups of significantly changing gRNAs. The main metrics used was the number of unique significant gRNAs per motif across all screens. Per screen, the direction of the effect on proliferation was quantified as the mean fold change of the gRNAs to define depletion or enrichment, and the fold change of the gRNAs for a region were compared against the fold change for all gRNAs in the analysis using a Mann–Whitney U test to calculate an enrichment or depletion p-value. AlphaFold2 was applied as post-screen filtering to flag motifs that are buried or in regions incompatible with binding. The structural context of each region was defined using AlphaFold2 models. Full-length protein AlphaFold2 models were retrieved from the AlphaFold2 protein structure database (https://alphafold.ebi.ac.uk/). The structured regions of the proteins were extracted using the xProtCAS tool [38] using a graph-based community detection algorithm on the AlphaFold2 predicted aligned error (PAE) matrix. Based on the xProtCAS, the determined structural state of each residue was classified as Domain, Loop or Unstructured. The unstructured proportion contains IDRs and coiled coils; coiled coil regions were defined as predicted helices greater than 25 amino acids in length; the remaining residues were classified as IDR. Potential intramolecular IDR - Domain contacts were also determined as the first or last region in an autonomous structural module of length less than

30 connected by a loop of greater than 20 amino acids. The SLiMPrints portion of the initial library contained 760 predicted motif instances and 10,606 gRNAs that were filtered after AlphaFold2 predicted they were located in coiled coil or domain regions. As a result, 6,489 of the 7,293 motif-containing regions were retained for downstream processing and analysis.

Region annotation

Regions were annotated using the PepTools peptide annotation tool (https://slim.icr.ac.uk/tools/peptools/) to define overlapping mutagenesis, disease relevant SNVs, and other features of interest[39]. Protein complex members were retrieved from the Complex Portal[40]. Pathway enrichment analysis was performed using the PANTHER enrichment tool[41] to identify enriched Reactome pathway annotation using the PANTHER API (https://www.pantherdb.org/services/) by comparing essential motif-containing proteins to the protein set tested in the base-editing analysis. The specificity determinant comparison analysis was performed with the CompariPSSM tool (https://github.com/ifigenia-t/CompariPSSM/). The specificity determinants of the regions targeted in the base-editing analysis were extracted as conservation-derived PSSMs. An orthologue alignment was created for each region by running the GOPHER[42] algorithm on the region-containing protein against a search database of metazoan proteomes. For each region, the columns corresponding to the edited region were extracted from the alignment and converted into a frequency PSSM. Frequency PSSMs were also created from aligned experimentally tested motif instances for motif classes from the ELM resource (http://elm.eu.org/) and a more extensive in-house curated MoMaP motif dataset (https://slim.icr.ac.uk/momap/). The base-edited PSSM set was then compared against the ELM PSSM dataset, MoMaP dataset, and itself using the CompariPSSM tool. Significant matches were defined as comparisons with a CompariPSSM similarity significance of $<1x10^{-5}$, dissimilarity score of <1 and PSSM overlap >75%.

Deep sequencing of CPSF1 locus

Individual sgRNAs targeting CPSF1 (Supplementary Table S12) were ordered as DNA oligonucleotides (TAG Copenhagen) with overhangs for Esp3I cloning: Forward oligo: 5'-CACCG[gRNA (20nt)]-3'; Reverse oligo: 5'-AAAC[reverse complement gRNA (20nt)]C-3'. The oligonucleotides were annealed and phosphorylated by T4 PNK (NEB) and cloned into either base editor construct pRDA_478 or pRDA_479 using golden gate cloning with Esp3I and T7 ligase. Each sgRNA was cloned into the backbone where it was predicted

to give an edit. Empty pRDA_478 and pRDA_479 were used as controls. Ligated plasmids were transformed into NEB Stable Competent E. coli and grown on LB agar containing ampicillin at 30°C for 24 h, followed by plasmid preparation from single colonies. Virus particles for each sgRNA-expressing base editor construct were produced as described above. HAP1 cells were transduced with each sgRNA-expressing base editor construct, selected, and cultured for 18 days after selection (t0-t18) as described above. Genomic DNA was extracted from cell pellets harvested at t0 and t18 using the QIAamp DNA Blood Mini Kit (Qiagen). A PCR reaction was set up to amplify the genomic CPSF1 locus, where editing was predicted to occur, and to introduce Illumina TruSeq Adaptor flaps. This was done using NEBNext Ultra II Q5 Master Mix with primers: Forward: *ACACTCTTTCCCTACACGACGCTCTTCCGATCT*TTCTTCGGGCCTCCTCCTT; Reverse: *GACTGGAGTTCAGACGTGTGCTCTTCCGATCT*TGGGGGCGCAGGTAGG (Illumina adaptor flaps in italic). A second PCR reaction introduced unique combinations of Illumina i5 and i7 multiplexing barcodes using NEBNext Ultra II Q5 Master Mix and Illumina TruSeq primers listed in Supplementary Table S12. Gel-purified PCR products were sequenced on Illumina NextSeq2000 with a 25% spike-in of PhiX and analysed by CRISPResso2 [43] .

Cloning

pcDNA5/FRT/TO plasmids encoding venus-tagged baits (Table S7) in a wildtype (WT) or mutant (MUT) version were generated by gene synthesis, mutagenesis, and subcloning into pcDNA5/FRT/TO/venus (GeneArt, ThermoScientific). pcDNA5/FRT/TO/Venus-CPSF1 were mutated by site-directed mutagenesis PCR.. WDR33 431-485 sequences were amplified by PCR and cloned into pGEX4T-1. The E.coli expression construct for CSTF3 21-549 was done by LIC cloning into an in house expression vector. All constructs were verified by sequencing.

IP-MS

HeLa cells were transfected with plasmids encoding mVenus-tagged baits (Table S7) using Lipofectamine 2000 24 or 48 hours prior to harvesting. Cells were lysed on ice in lysis buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.1% NP-40) supplemented with 1 mM DTT, complete protease inhibitor cocktail (Roche), and 4U/mL RNAse, and sonicated for 10 min (30’’ON/30’’OFF, max) using Bioruptor Plus (Diagenode). Lysates were cleared by centrifugation at 20,000 g at 4°C for 30 minutes, and venus-tagged baits were

immunoprecipitated using GFP-trap beads (ChromoTek). Precipitated protein complexes were washed twice in lysis buffer and once in PBS.

Affinity purification and mass spectrometry (AP-MS) sample preparation

mVenus-tag affinity purification was performed as described above. Partial on-bead digestion was performed with buffer (2 M urea; 2 mM DTT; 20 µg/ml Trypsin; 50 mM Tris, pH 7.5) incubated at 37°C for 30 min at 1,400 rpm. Supernatants were transferred to new tubes, alkylated with 25 mM chloroacetamide (CAA) and further digested overnight at RT and 1,000 rpm. Digestion was terminated with 1% trifluoroacetic acid (TFA). Peptides were desalted and purified using styrenedivinylbenzene-reversed phase sulfonate (SDB-RPS) StageTips prepared in 0.2% TFA. Peptides were washed and eluted with an elution buffer (80% acetonitrile (ACN); 1% ammonia) prior to vacuum-drying. Dried peptides were reconstituted in 2% ACN and 0.1% TFA.

Liquid chromatography-mass spectrometry (LC-MS) analysis

All samples were loaded onto Evotips Pure and measured with a data-independent acquisition (DIA) method. 200 ng of peptides were partially eluted from Evotips with <35% acetonitrile and analysed with an Evosep One LC system (Evosep Biosystems) coupled online to an Orbitrap mass spectrometer (Orbitrap Astral or Orbitrap Exploris 480, Thermo Fisher Scientific)[44–46]. Eluted peptides were separated on a 8-cm-long PepSep column (150 µm inner diameter packed with 1.5 µm of Reprosil-Pur C18 beads (Dr Maisch)) in a standard gradient method with a stainless emitter (30 µm inner diameter). The mobile phases were 0.1% formic acid in liquid chromatography (LC)–MS-grade water (buffer A) and 0.1% formic acid in acetonitrile (buffer B). Data was acquired in DIA mode.

For the samples acquired in the Orbitrap Exploris 480, peptides were eluted online from the EvoTips using an Evosep One LC system and analysed at 30 samples per day (44-min gradient). The mass spectrometer was operated in positive mode using the DIA mode. Full scan precursor spectra (350–1,000 m/z) were recorded in profile mode using a resolution of 120,000, with a normalised automatic gain control target (AGC) of 300% and a maximum injection time of 45 ms. Fragment spectra were then recorded in profile mode, fragmenting 49 consecutive 13.3 m/z windows covering the mass range 350–1,000 m/z and using a resolution of 15,000. Isolated precursors were fragmented in the higher-energy collisional dissociation

(HCD) cell using 27% normalised collision energy, a normalised AGC target of 3,000 % and a maximum injection time of 22 ms.

For the samples acquired in the Orbitrap Astral, peptides were eluted online from the EvoTips using an Evosep One LC system and analysed at 60 samples per day (21-min gradient). The Orbitrap Astral mass spectrometer was operated in positive mode using the DIA mode, at a resolution of 240,000 with a full scan range of 380-980 m/z. The normalised AGC was set to 500%. Fragment ion scans were recorded at a maximum injection time of 3.5 ms, 200 windows of 3Th scanning from 380-980 m/z. Ions were fragmented using HCD with 25% normalised collision energy.

Raw data processing and bioinformatic analysis

Raw files were analysed with directDIA workflow in Spectronaut v.19.0[47]. Default settings were used. Data filtering was set to 'Qvalue'. 'Cross run normalisation' was enabled with the strategy of 'local normalisation' based on rows with 'Qvalue complete'. FDR was set to 1% at both the protein and peptide precursor levels. Raw data was searched against the human proteome reference database (including 20,593 entries, Uniprot March 2023). Data was filtered for 60% valid values across WT or Mutant samples (protein groups with >40% missing values were excluded from downstream statistical analysis). Protein intensities were then log2 transformed for downstream statistical and bioinformatics analysis. Imputation of missing data was performed by random numbers drawn from a normal distribution with a width of 0.3 and a downshift of 1.8 applied to each sample. For ZNHIT-IP, we did extra symmetrical z-score normalisation. Significance was determined through an unpaired two-sided *t*-test corrected by Benjamini–Hochberg at FDR of 0.1 and fold change of 1.5[48,49]. Data analysis and visualisation were done in Jupyter Notebook in the environment of Python 3.9.

IP-WB

HeLa cells were transfected with plasmids encoding the specified venus-tagged baits in a wildtype version or with the indicated mutations using Lipofectamine 2000 24 hours prior to harvesting. Cells were lysed on ice in lysis buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.1% NP-40) supplemented with 1 mM DTT and complete protease inhibitor cocktail (Roche) and sonicated for 10 min (30''ON/30''OFF, max) using Bioruptor Plus (Diagenode). Lysates were cleared by centrifugation at 20,000 g at 4°C for 30 minutes, and venus-tagged

baits were immunoprecipitated using GFP-trap beads (ChromoTek). Precipitated protein complexes were washed thrice in lysis buffer and eluted in 2X LDS sample buffer before SDS-PAGE and Western blot analysis with the indicated antibodies (CSTF3: HPA040168 (Sigma), CNOT1 14276-1-AP proteintech, GFP in house made antibody to full length GFP).

Protein production

$CSTF3_{21-549}$ wild-type and $GST\text{-}WDR33_{431-485}$ wild-type and mutant were expressed in BL21(DE3) cells overnight at 18 degrees and cells harvested by centrifugation. For CSTF3 purification cells were resuspended in Buffer L (50 mM NaP pH=7,5; 300 mM NaCl; 10 mM imidazole; 10% glycerol; 0,5 mM TCEP; 1xComplete EDTA-free tablets (Roche)) and lysed. Following clarification by centrifugation lysate was loaded on a 5 ml HiTrap column and His-CSTF3 21-549 eluted with an imidazole gradient. The peak fractions were pooled diluted with 50 mM NaP pH=7,5 to 50 mM NaCl and loaded on a SP HP HiTrap column and proteins eluted with a 50mM-1M NaCl salt gradient. Peak fractions were pooled.

For purification of GST-WDR33 431-485 cell pellets were resuspended in 25 mL lysis buffer (50 mM Tris-HCl pH 7.4, 300 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol) and sonicated for 10 minutes on ice. Lysates were cleared in a fixed angle rotor for 40 min 145000 rpm and supernatants were incubated with 350 µL pre-equilibrated Glutathione Sepharose Fast Flow (GE healthcare) for 1h. Beads were washed 3 times with 30 mL lysis buffer and used directly for binding experiments.

GST-WDR33 pull down assay.

10 µL GST-WRD3 (wt or mutant) beads in 200 µL buffer B (150 mM NaCl, 50 mM Tris pH7,4 1mM DTT) were incubated with 20 µg CTSF3 for 1h at 4C and rotation. Beads were washed 2 times with buffer B and eluted in 25 µL 2x LSB(Thermo Fisher Scientific), 10mM DTTFor control conditions unbound glutathione beads were used.

ITC

Peptides were purchased from Peptide 2.0 Inc (Chantilly. VA, USA). The purity obtained in the synthesis was 95–98% as determined by high performance liquid chromatography (HPLC) and subsequent analysis by mass spectrometry. Prior to ITC experiments both $CSTF3^{21-549}$ and the peptides were extensively dialyzed against 50 mM sodium phosphate pH 7.5, 150 mM NaCl, 0.5 mM TCEP. $CSTF3^{21-549}$ and peptide concentrations were determined using a spectrometer by measuring the absorbance at 280 nm and applying values for the

extinction coefficients computed from the corresponding sequences by the ProtParam program (http://web.expasy.org/protparam/). Peptides at approximately 500 μM concentration were loaded into the syringe and titrated into the calorimetric cell containing CSTF3$^{21-549}$ at ~ 40 μM (monomer concentration). The reference cell was filled with distilled water. In all assays, the titration sequence consisted of a single 0.4 μl injection followed by 19 injections, 2 μl each, with 150 s spacing between injections to ensure that the thermal power returns to the baseline before the next injection. The stirring speed was 750 rpm. Control experiments with peptides injected in the sample cell filled with buffer were carried out under the same experimental conditions. These control experiments showed negligible heats of dilution. The heats per injection normalised per mole of the injectant *versus* the molar ratio [Peptide]/[CSTF3$^{21-549}$] were fitted to a single-site model. Data were analysed with MicroCal PEAQ-ITC (version 1.1.0.1262) analysis software (Malvern Instruments Ltd.).

Data availability

The proteomic raw data is available at PRIDE ID PXD055750

Username: reviewer_pxd055750@ebi.ac.uk

Password: lyWJc1nnFmbq

Conflict of interest

The authors have no conflict of interest

Figure legends

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

a, SLiM-mediated interaction exemplified by the p53 MDM2 degron binding to the MDM2 SWIB domain. **b**, Evolutionary conservation of key affinity and specificity determinants exemplified by the S100A binding SLiM in CBP/P300. **c**, Composition of the SLiM targeting base editor library. **d**, Editing properties of the Cytosine Base Editor (CBE) and Adenosine Base Editor (ABE) classes used in the screen. **e**, Example of gRNA target DNA sequence with optimal base editing window and protospacer adjacent motif (PAM) highlighted. **f**, Statistics of predicted mutations introduced in target motifs by gRNAs cloned into ABE and/or CBE. **g**, Schematic of the major steps of the experimental pipeline from library design to data processing. **h**, ROC curve of the gRNA fold change for the control gRNAs across the 4 base editors. AUC: area under the curve. **i**, Rank plots of the gRNA fold change. TPR: True positive rate. FPR: False positive rate. **j**, Comparison of the effect of gRNA fold change in the ABE_{NGG} and CBE_{NGG} screens.

Figure 2: Identification of SLiMs required for optimal fitness

a-c, Example of region with depleted (red) (**a**), enriched (green) (**b**) or unchanged (grey) (**c**) gRNAs. Left panels: Y-axes indicate Log2-fold change (LFC) of all gRNAs targeting the indicated motif summarised for the 4 screens ('All BEs') and individual screens. Dot size represents limma p-value. Right panels: Y-axes indicate LFC of a given gRNA predicted to target a given amino acid residue within the represented motif, with connected dots indicating the same guide edit across the biological triplicate of the ABE_{NGN} screen. **d**, **e**, Number of regions targeted by significantly enriched or depleted guides. **f,** M-W p-value versus fold change using the most significant p-values for each motif from the 4 different screens. Red: *reported motif set*. Grey: *predicted motif set*. Size represents significant gRNA count. **g**, Subdivision of motifs in panel **f** per ELM type. **h**, Fold change of gRNAs targeting fixed consensus residues required for binding or other residues in significantly scoring motif regions. **i**, Conservation of motifs categorised as essential (significantly depleted) versus non-essential (others). *RLCp*: SLiMPrint relative local conservation probability[35]. **j**, Log2-fold change of motifs categorised by HAP1 gene essentiality. **k**, Top enriched (blue) and depleted (red) reported motifs in the four screens. *Score:* log10 M-W p-value, based on fold change, score is negative for depleted motifs, positive for enriched motifs.

Figure 3: Validation of a essential predicted CSTF3 interaction motif in the cleavage and polyadenylation specificity factor CSPF1

a, Evolutionary conservation of the CSPF1 $DDEEEMLYG_{741-749}$ motif and the CSTF3-binding FIP1L1 motif. **b**, Volcano plot of AP-MS for motif-mutant versus wild-type CSPF1. The dashed lines indicate fold-change (1.5) and p-value (0.1) significance cut-offs. **c**, AF3 prediction modelling the CPSF1-WDR33-CPSF30-RNA-CSTF3 complex. The CSPF1 DEEEMLYG motif is indicated in bright green, motif-binding pocket on CSTF3 is indicated in red. **d**, LFC (t18 vs t0) of edited residues (>10% allelic frequency) in CSPF1 $DDEEEMLYG_{741-749}$ introduced by CBE_{NGN} (green) and ABE_{NGN} (blue) as validated by deep sequencing. **e**, Correlation of LCF of validated edits (>10% allelic frequency) versus LFC of respective guides in CBE_{NGN} and ABE_{NGN} screens. **f**, ITC binding data of CPSF1 (737-751) wild-type and mutant peptides to CSTF3. **g**, Representative Western blot of full length wild-type and mutant mVenus-CSPF1 GFP-traps probed for CSTF3 and GFP (mVenus) (n=3).

Figure 4: Characterisation of essential predicted motifs

a, Top depleted predicted motifs in the four screens. *Score:* log10 M-W p-value, based on fold change, score is negative for depleted motifs, positive for enriched motifs. **b**, Comparison of the essential motifs to the specificity determinants of the ELM resource[5] **c**, 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 motif. **e**, Volcano plot of AP-MS for motif-mutant versus wild-type EIF3G. The dashed lines indicate fold-change (1.5) and p-value (0.1) significance cut-offs. **f,** Summary of AP-MS experiments showing top significantly lost binder or, if relevant, top lost complex. **g**, Conservation logo ANKRD17 $PIGTER_{2426-2431}$ motif with surrounding sequence and homologous region of ANKHD1. **h**, Volcano plot of AP-MS for motif-mutant versus wild-type ANKRD17. The dashed lines indicate fold-change (1.5) and p-value (0.1) significance cut-offs. **i**, Representative western blot and quantification of YFP-ANKRD17 (YFP-ANK17) wild-type and mutant binding to CNOT1 **(**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**, 161–169 (2014).
2. Van Roey, K. *et al.* Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.* **114**, 6733–6778 (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**, 268–281 (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).

19. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

20. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).

21. Safaee, N. *et al.* Interdomain allostery promotes assembly of the poly(A) mRNA complex with PABP and eIF4G. *Mol. Cell* **48**, 375–386 (2012).

22. Clute, P. & Pines, J. Temporal and spatial control of cyclin B1 destruction in metaphase. *Nat. Cell Biol.* **1**, 82–87 (1999).

23. Lee, E. F. *et al.* The functional differences between pro-survival and pro-apoptotic B cell lymphoma 2 (Bcl-2) proteins depend on structural differences in their Bcl-2 homology 3 (BH3) domains. *J. Biol. Chem.* **289**, 36001–36017 (2014).

24. Coutandin, D. *et al.* Conformational stability and activity of p73 require a second helix in the tetramerization domain. *Cell Death Differ.* **16**, 1582–1589 (2009).

25. Czabotar, P. E. *et al.* Mutation to Bax beyond the BH3 domain disrupts interactions with pro-survival proteins and promotes apoptosis. *J. Biol. Chem.* **286**, 7123–7131 (2011).

26. Suzuki, M., Youle, R. J. & Tjandra, N. Structure of Bax: coregulation of dimer formation and intracellular localization. *Cell* **103**, 645–654 (2000).

27. Muckenfuss, L. M., Migenda Herranz, A. C., Boneberg, F. M., Clerici, M. & Jinek, M. Fip1 is a

multivalent interaction scaffold for processing factors in human mRNA 3' end biogenesis. *eLife* **11**, (2022).

28. Yamazaki, Y. *et al.* Goliath family E3 ligases regulate the recycling endosome pathway via VAMP3 ubiquitylation. *EMBO J.* **32**, 524–537 (2013).
29. von Moeller, H. *et al.* Structural and biochemical studies of SLIP1-SLBP identify DBP5 and eIF3g as SLIP1-binding proteins. *Nucleic Acids Res.* **41**, 7960–7971 (2013).
30. 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).
31. 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**, 1387–1397 (2010).
32. 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).
33. Boland, A. *et al.* Structure and assembly of the NOT module of the human CCR4-NOT complex. *Nat. Struct. Mol. Biol.* **20**, 1289–1297 (2013).
34. 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).
35. Davey, N. E. *et al.* SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res.* **40**, 10628–10641 (2012).
36. 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**, 827–839 (2005).
37. UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
38. Kotb, H. M. & Davey, N. E. xProtCAS: A Toolkit for Extracting Conserved Accessible Surfaces from Protein Structures. *Biomolecules* **13**, (2023).

39. Benz, C. *et al.* Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Mol. Syst. Biol.* **18**, e10584 (2022).

40. Meldal, B. H. M. *et al.* Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* **47**, D550–D558 (2019).

41. 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).

42. 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).

43. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).

44. 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).

45. 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).

46. Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-023-02099-7.

47. 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**, 1400–1410 (2015).

48. Lin, M.-H. *et al.* Benchmarking differential expression, imputation and quantification methods for proteomics data. *Brief. Bioinformatics* **23**, (2022).

49. 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).

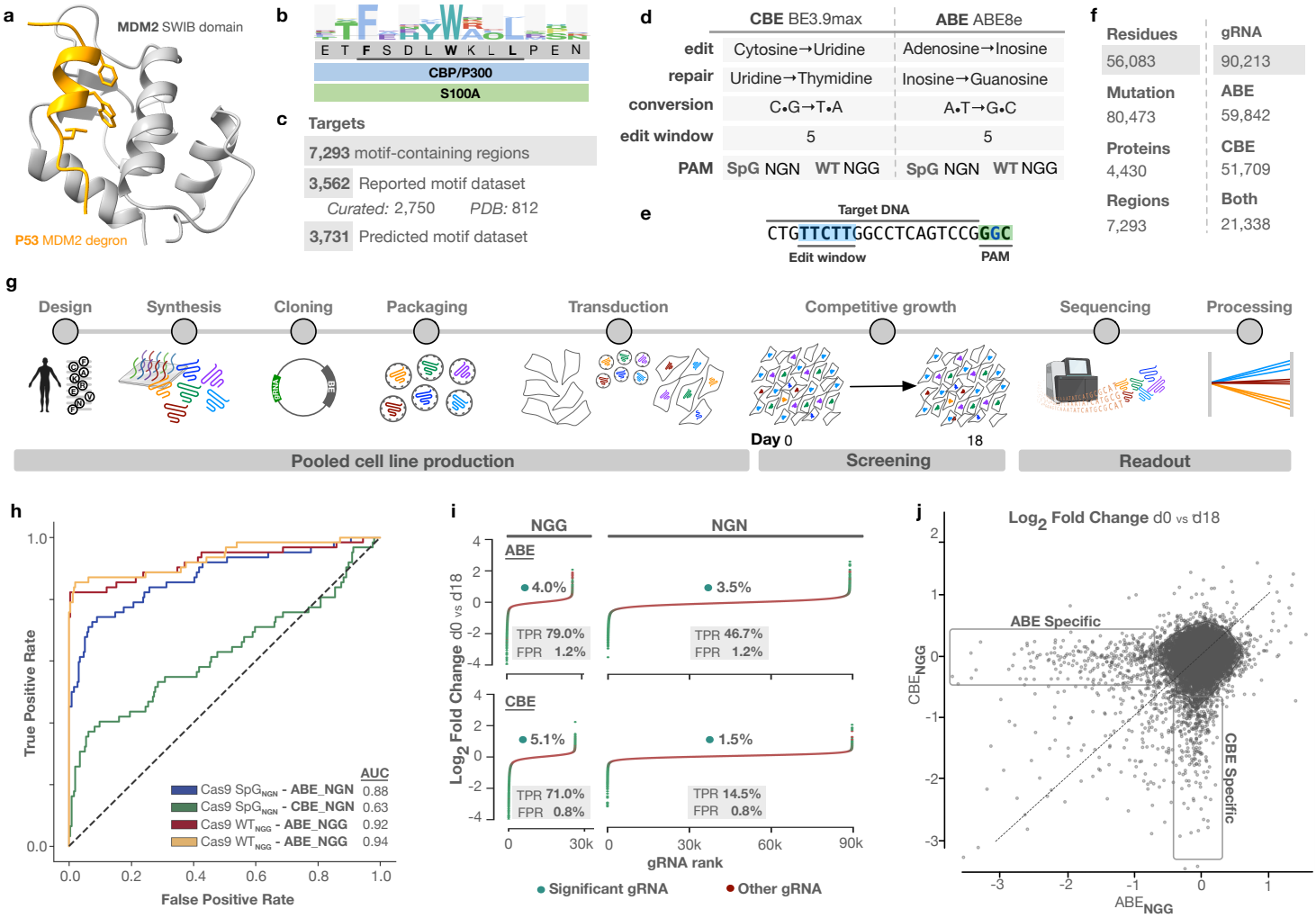

a
MDM2 SWIB domain
P53 MDM2 degron
b
E T F S D L W K L L P E N
CBP/P300
S100A
c
Targets
7,293 motif-containing regions
3,562 Reported motif dataset
Curated: 2,750 PDB: 812
3,731 Predicted motif dataset
d
CBE BE3.9max
ABE ABE8e
edit Cytosine→Uridine Adenosine→Inosine
repair Uridine→Thymidine Inosine→Guanosine
conversion C•G→T•A A•T→G•C
edit window 5 5
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
Day 0
18
Pooled cell line 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 - ABE_NGG 0.94
i
NGG
NGN
ABE
CBE
Log2 Fold Change d0 vs d18
4.0%
TPR 79.0%
FPR 1.2%
3.5%
TPR 46.7%
FPR 1.2%
5.1%
TPR 71.0%
FPR 0.8%
1.5%
TPR 14.5%
FPR 0.8%
gRNA rank
Significant gRNA
Other gRNA
j
Log2 Fold Change d0 vs d18
ABE Specific
CBE Specific
CBE NGG
ABE NGG

Figure 1

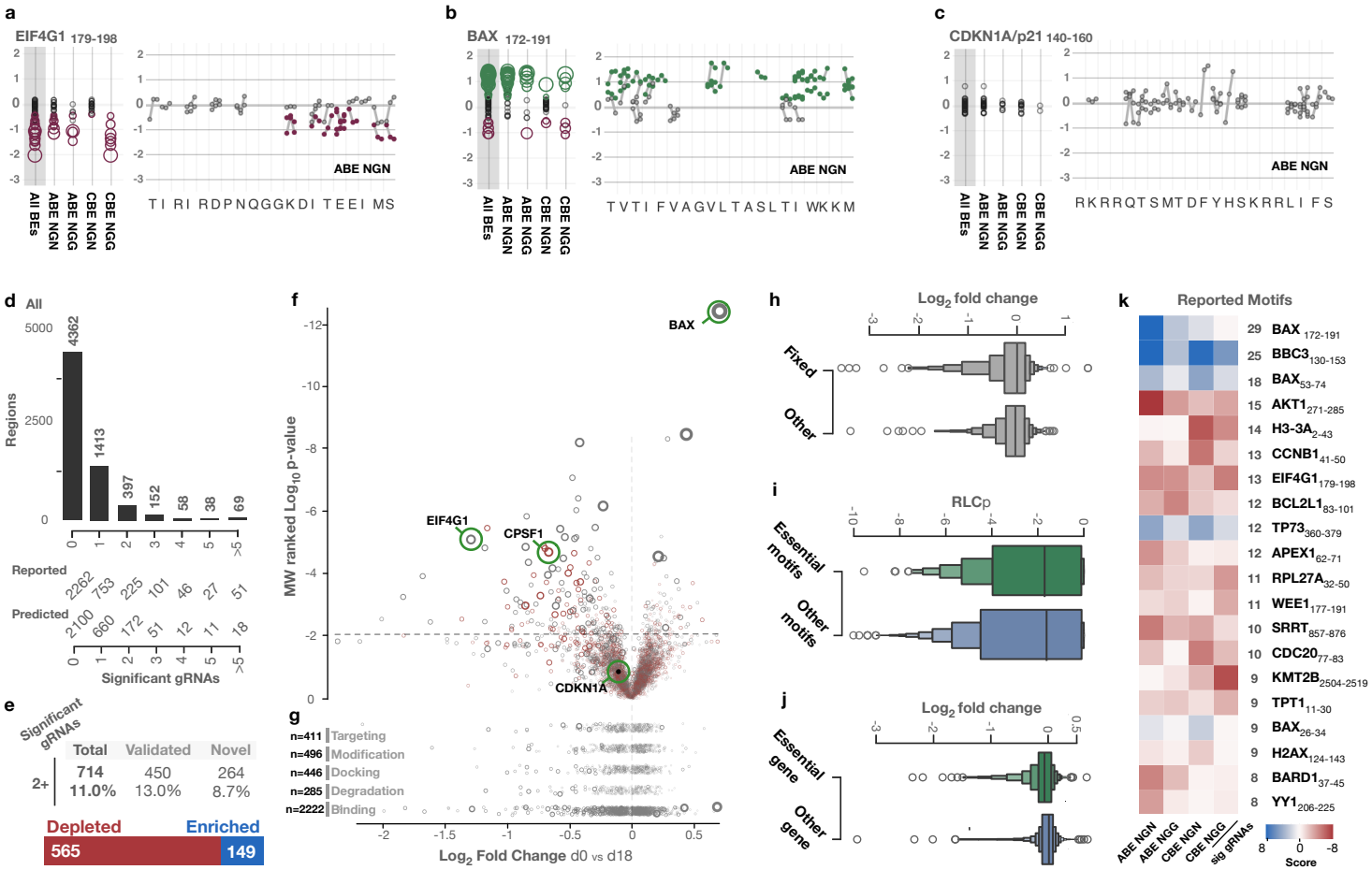

a
EIF4G1 179-198
All BEs
ABE NGN
ABE NGG
CBE NGN
CBE NGG
T I RI RDPNQGGKDI TEEI MS
ABE NGN
b
BAX 172-191
TVTI FVAGVLTASLTI WKKM
c
CDKN1A/p21 140-160
RKRRQTSMTDFYHSKRRLI FS
d All
Regions
Reported
Predicted
Significant gRNAs
e
Significant gRNAs
Total Validated Novel
2+ 714 450 264
11.0% 13.0% 8.7%
Depleted Enriched
565 149
f
MW ranked Log10 p-value
BAX
EIF4G1
CPSF1
CDKN1A
g
n=411 Targeting
n=496 Modification
n=446 Docking
n=285 Degradation
n=2222 Binding
Log2 Fold Change d0 vs d18
h
Log2 fold change
Fixed
Other
i
RLCp
Essential motifs
Other motifs
j
Log2 fold change
Essential gene
Other gene
k
Reported Motifs
29 BAX 172-191
25 BBC3 130-153
18 BAX 53-74
15 AKT1 271-285
14 H3-3A 2-43
13 CCNB1 41-50
13 EIF4G1 179-198
12 BCL2L1 83-101
12 TP73 360-379
12 APEX1 62-71
11 RPL27A 32-50
11 WEE1 177-191
10 SRRT 857-876
10 CDC20 77-83
9 KMT2B 2504-2519
9 TPT1 11-30
9 BAX 26-34
9 H2AX 124-143
8 BARD1 37-45
8 YY1 206-225
ABE NGN
ABE NGG
CBE NGN
CBE NGG
sig gRNAs
Score

Figure 2

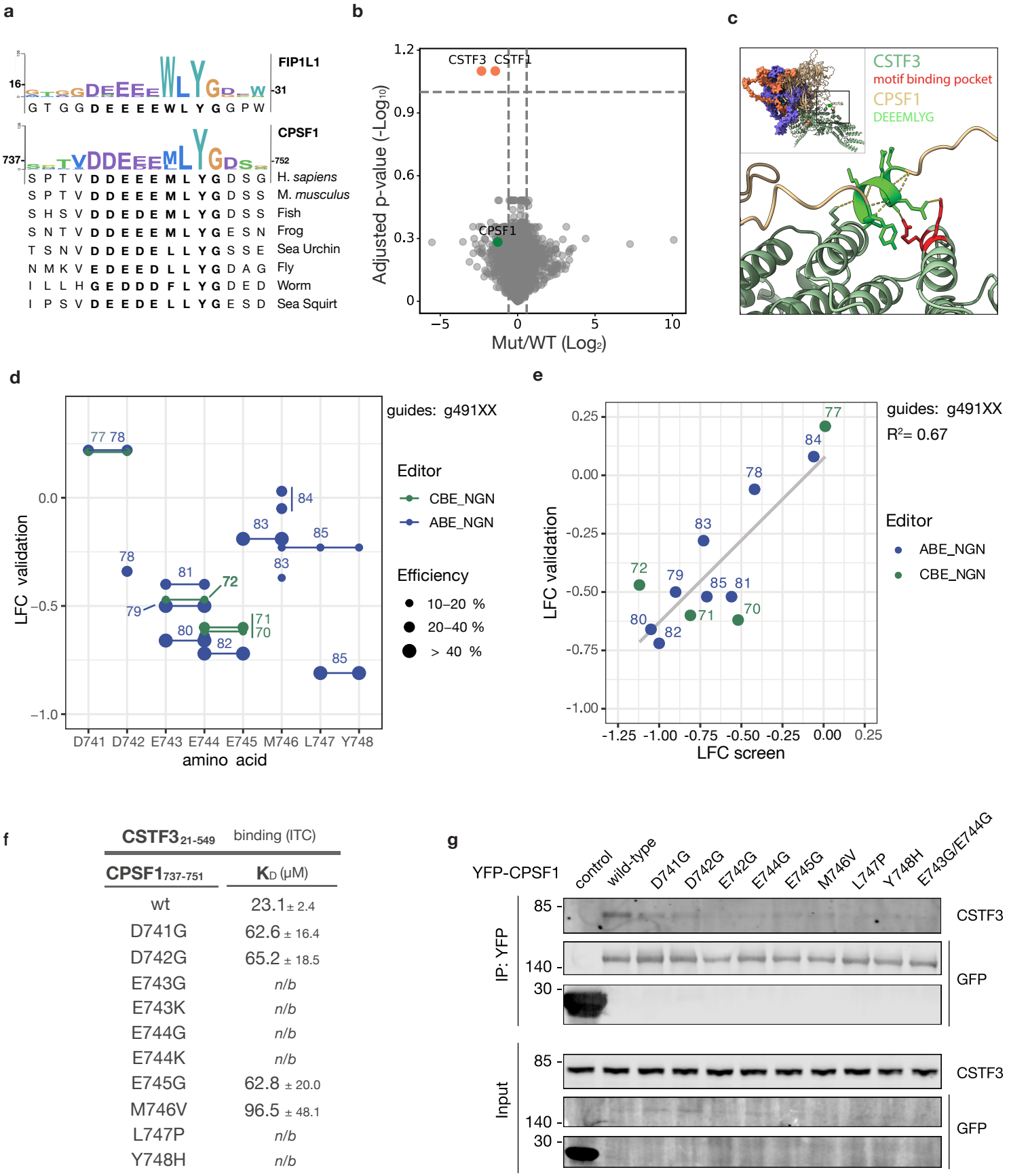
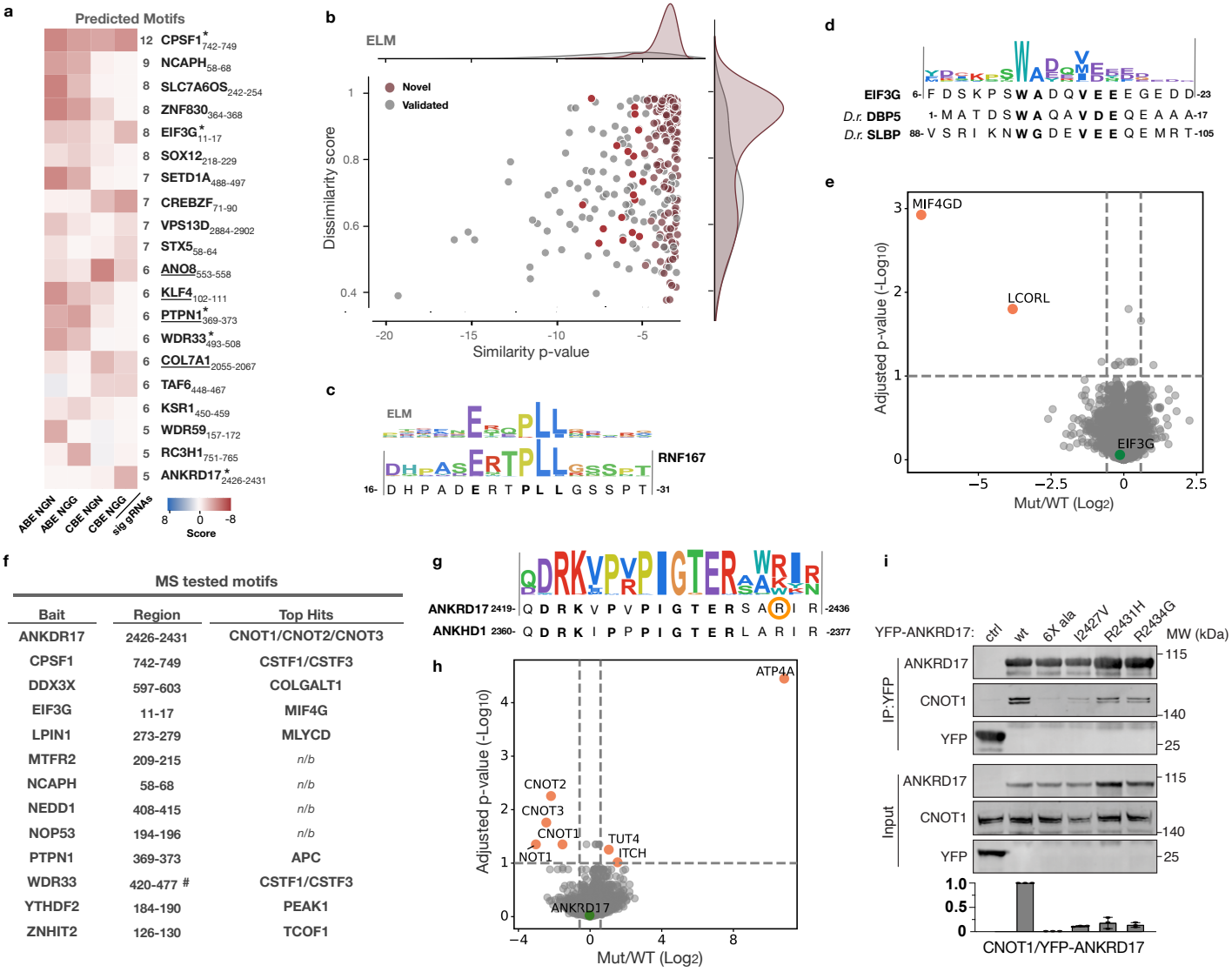

a
FIP1L1
16
31
G T G G D E E E E W L Y G G P W
CPSF1
737
752
S P T V D D E E E M L Y G D S G H. sapiens
S P T V D D E E E M L Y G D S S M. musculus
S H S V D D E D E M L Y G D S S Fish
S N T V D D E E E M L Y G E S N Frog
T S N V D D E D E L L Y G S S E Sea Urchin
N M K V E D E E D L L Y G D A G Fly
I L L H G E D D D F L Y G D E D Worm
I P S V D E E D E L L Y G E S D Sea Squirt
b
CSTF3 CSTF1
CPSF1
Adjusted p-value (-Log10)
Mut/WT (Log2)
c
CSTF3
motif binding pocket
CPSF1
DEEEMLYG
d
guides: g491XX
Editor
CBE_NGN
ABE_NGN
Efficiency
10–20 %
20–40 %
> 40 %
LFC validation
amino acid
D741 D742 E743 E744 E745 M746 L747 Y748
e
guides: g491XX
R²= 0.67
Editor
ABE_NGN
CBE_NGN
LFC validation
LFC screen
f
CSTF3 21-549 binding (ITC)
CPSF1 737-751 | KD (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
g
YFP-CPSF1
control
wild-type
D741G
D742G
E742G
E744G
E745G
M746V
L747P
Y748H
E743G/E744G
IP: YFP
Input
CSTF3
GFP
85
140
30

Figure 3



Bait	Region	Top Hits
ANKDR17	2426-2431	CNOT1/CNOT2/CNOT3
CPSF1	742-749	CSTF1/CSTF3
DDX3X	597-603	COLGALT1
EIF3G	11-17	MIF4G
LPIN1	273-279	MLYCD
MTFR2	209-215	*n/b*
NCAPH	58-68	*n/b*
NEDD1	408-415	*n/b*
NOP53	194-196	*n/b*
PTPN1	369-373	APC
WDR33	420-477 #	CSTF1/CSTF3
YTHDF2	184-190	PEAK1
ZNHIT2	126-130	TCOF1

Figure 4