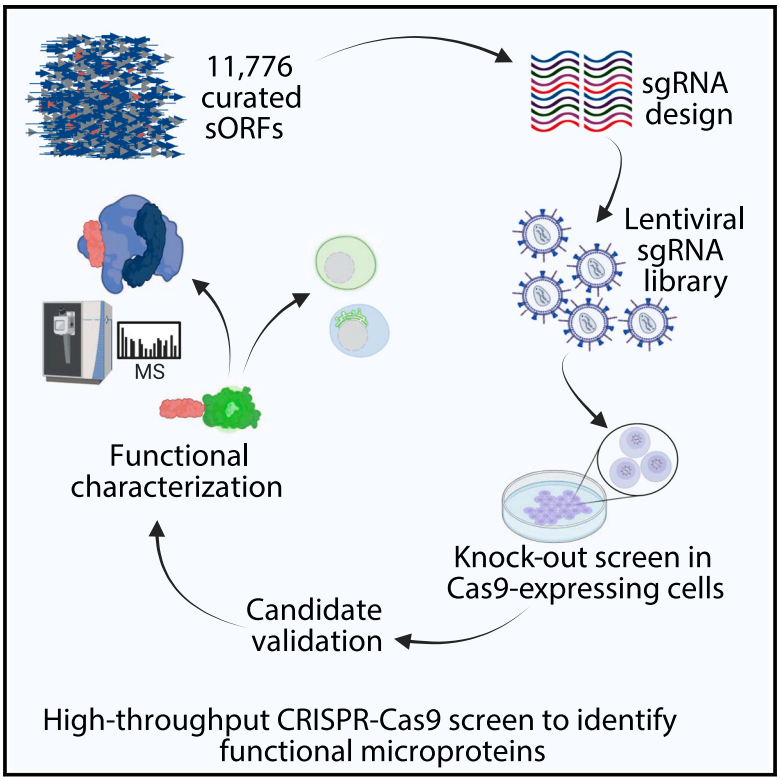
iScience

Article

A large-scale sORF screen identifies putative microproteins involved in cancer cell fitness

Graphical abstract



Highlights

- CRISPR-Cas9 knock-out screening of 11,776 sORFs for role in cancer cell growth
- Gene expression changes upon candidate knock-out could be partially rescued
- Endogenous tagging demonstrated translation of an sORF in the *CENPBD2P* pseudogene
- Microprotein candidates show distinct subcellular localizations and interactors

Authors

Dörte Schlesinger, Christopher Dirks, Carmen Navarro, ..., Jürgen Eirich, Thomas Farid Martinez, Simon Johannes Elsässer

Correspondence

simon.elsasser@scilifelab.se

In brief

molecular genetics, classification of proteins, methodology in biological sciences, cancer, and cell biology

Schlesinger et al., 2025, iScience *28*, 111884
March 21, 2025 © 2025 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2025.111884

CellPress

iScience

CellPress
OPEN ACCESS

Article

A large-scale sORF screen identifies putative microproteins involved in cancer cell fitness

Dörte Schlesinger,[1,2] Christopher Dirks,[1,2] Carmen Navarro,[1,2] Lorenzo Lafranchi,[1,2] Anna Spinner,[1] Glancis Luzeena Raja,[1] Gregory Mun-Sum Tong,[4] Jürgen Eirich,[1,3] Thomas Farid Martinez,[4,5,6] and Simon Johannes Elsässer[1,2,7,*]

[1]Science for Life Laboratory, Karolinska Institutet, Department of Medical Biochemistry and Biophysics, Division of Genome Biology, 17165 Stockholm, Sweden
[2]Ming Wai Lau Centre for Reparative Medicine, Stockholm node, Karolinska Institutet, 17165 Stockholm, Sweden
[3]University of Münster, Institute of Plant Biology and Biotechnology (IBBP), 48143 Münster, Germany
[4]Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA 92617, USA
[5]Department of Biological Chemistry, University of California, Irvine, Irvine, CA 92617, USA
[6]Chao Family Comprehensive Cancer Center, University of California, Irvine, Irvine, CA 92617, USA
[7]Lead contact
*Correspondence: simon.elsasser@scilifelab.se
https://doi.org/10.1016/j.isci.2025.111884

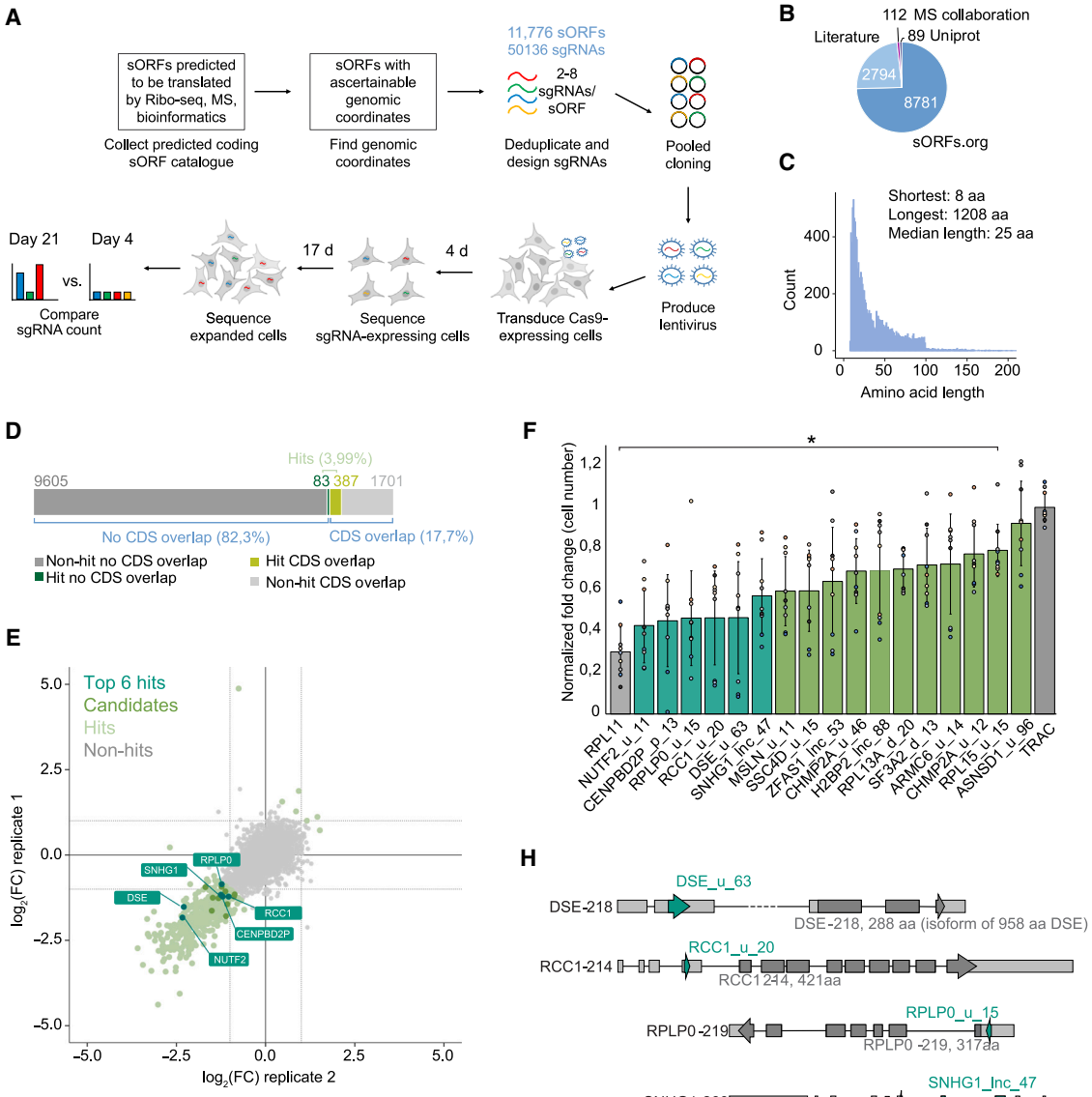
SUMMARY

The human genome contains thousands of potentially coding short open reading frames (sORFs). While a growing set of microproteins translated from these sORFs have been demonstrated to mediate important cellular functions, the majority remains uncharacterized. In our study, we performed a high-throughput CRISPR-Cas9 knock-out screen targeting 11,776 sORFs to identify microproteins essential for cancer cell line growth. We show that the *CENPBD2P* gene encodes a translated sORF and promotes cell fitness. We selected five additional candidate sORFs encoding microproteins between 11 and 63 amino acids in length for further functional assessment. Green fluorescent protein fusion constructs of these microproteins localized to distinct subcellular compartments, and the majority showed reproducible biochemical interaction partners. Studying the fitness and transcriptome of sORF knock-outs and complementation with the corresponding microprotein, we identify rescuable phenotypes while also illustrating the limitations and caveats of our pipeline for sORF functional screening and characterization.

INTRODUCTION

Microproteins are polypeptides originating from short open reading frames (sORF) of less than a hundred codons. For a long time, they have been understudied, as it is difficult to distinguish coding from non-coding sORFs. In recent years, the number of putatively translated sORFs could be narrowed down from hundred-thousands or millions to various thousands, due to the advent of ribosome profiling and advances in bioinformatic and proteomic techniques. Consequently, efforts are now being made to include sORFs with robust translation evidence into databases such as GENCODE.[1,2] The field has since steadily grown, though it is still unclear how many functional coding sORFs exist in the human genome, and relatively few microproteins have been characterized to date. Microproteins arise from a variety of origins, including the following: (1) from canonical protein-coding transcripts where microproteins can be translated for example as so-called upstream open reading frames (uORFs) from the 5' ''untranslated'' region[3]; (2) from sORFs overlapping and out-of-frame with the canonical ORFs[4]; and (3) from transcripts that were previously deemed non-coding,[5] such as long non-coding RNA (lncRNA)[6–9] or microRNAs.[10] Additionally, sORF translation from pseudogenes has been demonstrated.[11] Pseudogenes are often excluded from large scale translation studies due to potential similarity with their canonical origin, and difficulties in scoring them according to conservation criteria.[12] Owing to their small size, microproteins may be ideal for performing fine-tuning tasks.[13] Beyond this function, a number of characterized microproteins have also been shown to carry out versatile, and even essential cellular functions, including roles in signaling,[14–18] metabolism,[19–25] and stress and repair pathways.[26–29] Thus, to better understand cellular function, it is crucial to identify many more bioactive microproteins from existing repositories of predicted coding sORFs. Despite this dearth in literature, only few functional high-throughput screens of sORFs have been performed to-date in eukaryotic cells, each surveying various hundred to a few thousand noncanonical ORFs.[30–38] Here, we outline and demonstrate our strategy to identify and functionally elucidate microproteins required for growth and survival in three human cancer cell lines, utilizing a CRISPR-Cas9 dropout screen with a human sORF-specific library of 11,776 sORF candidates.

This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).



G

Peptide ID	Candidate	Gene of origin	Location	Start	Stop	Length (aa)	Source
PL01811	DSE_u_63	DSE	5′ UTR	ATG	TGA	63	Mackowiak et al., 2015
PS02457	RCC1_u_20	RCC1	5′ UTR	ATG	TAA	20	Olexiouk et al., 2015, Olexiouk et al., 2018. In sORFs.org as tirosh_2015:47449.
PS01753	RPLP0_u_15	RPLP0	5′ UTR	ATG	TGA	15	Olexiouk et al., 2015, Olexiouk et al., 2018. In sORFs.org as liu_Hela_2013:98166.
PL00150	SNHG1_lnc_47	SNHG1	lncRNA	ATG	TGA	47	Bazzini et al., 2014, with data from Guo et al., 2016.
PS05637	CENPBD2P_p_13	CENPBD2P	pseudogene	ATG	TGA	13	Olexiouk et al., 2015, Olexiouk et al., 2018. In sORFs.org as fritsch_2012:37213.
PS04061	NUTF2_u_11	NUTF2	5′ UTR	CTG	TAG	11	Olexiouk et al., 2015, Olexiouk et al., 2018. In sORFs.org as elkon_2015:484930.

(legend on next page)

RESULTS

An sORF-specific high-throughput CRISPR dropout screen to identify sORFs required for cell survival

In order to design a comprehensive CRISPR library targeting human sORF candidates, we curated sORFs (Figure 1A) from published Ribo-seq, mass spectrometry (MS), bioinformatic and combined evidence studies.[11,12,39–50] We further included various publications describing individual microproteins.[9,23,26,27,51–57] Additionally, we reviewed UniProt for human proteins below 60 amino acids in length, as well as candidates from the sORFs.org database,[58,59] applying the filtering criteria outlined in Table S1 (see STAR Methods section for detailed methodology). We used CRISPOR[60,61] to design sgRNAs against the collected sORFs, requiring that each sORF had to be targeted by at least two sgRNAs (Figure S1A). For sORFs with more than eight possible sgRNAs, we retained the eight best guides. Our final screening set included 11,776 sORFs targeted by a total of 50,136 sgRNAs with a median of six sgRNAs per sORF (Figures 1B, 1C, and S1B; Tables S2, S3, and S4). 2,088 (17.73%) of targeted sORFs displayed overlap with canonical protein-coding (CDS) ORFs (Figure 1D). Additionally, we included 292 positive control sgRNA targeting ribosomal genes, as well as 1,000 non-targeting negative control sgRNAs (Figure S1D). Since the curation of this sORF library, additional screens identifying sORFs have been published.[34–38] In assessing the overlap between our library and two published datasets, we identified 1,243 of our sORF genomic regions overlapping to an extent with Chen et al.[34] ORFs, and 317 with Prensner et al.[35] ORFs, albeit only 460 and 57, respectively, were exact matches of the sORF coordinates and sequence (Figure S1C; Table S5).

The total pool of 51,428 sgRNAs generated was then cloned and amplified into a lentiviral library, with which we carried out screens in three different cancer cell lines stably expressing Cas9 (A375 melanoma, HCT116 colon cancer, and K562 leukemia cells). We then performed essentiality screens in either optimal serum (10% FBS in all cell lines), serum starvation conditions (0% or 1% serum in HCT116, and 1% serum in K562), or drug resistance screening with 6-thioguanine (in A375 and K562) (Figure S1D).

The A375 and HCT116 screens displayed a good dynamic range and strong correlation between replicates (correlation coefficients of 0.6–0.66 for sgRNAs and 0.75–0.8 for ORFs) (Figures S1D and S2A). Overall $\log_2$ fold-changes (LFCs) between A375 and HCT116 also corresponded well (R = 0.75), while correlation between A375 and K562 was less clear (R = 0.4), since the K562 screens themselves did not yield a good dynamic range (Figure S2B). The A375 full serum screen showed the largest effect size and yielded a total of 470 hits with an LFC smaller than −1 or larger than 1 (3.99%) (Figures 1D, 1E, and S2A). With few exceptions, the A375 hits were located on transcripts with medium to high expression in the three cell lines (Figure S2C). We used RibORF to analyze published Ribo-seq datasets in the three cell lines and found that ∼25% of hits were supported by good translation evidence (score >0.7) (Figure S2D). Many of the 470 A375 hits were expressed (Figure S2C) and translated (Figure S2D) in HCT116 cells as well, and also showed a negative fold-change in HCT116 cells (R = 0.55) (Figure S2B). Hence, HCT116 cells appeared to largely phenocopy the hits found in A375, although many did not reach statistical significance due to the narrower overall dynamic range of the HCT116 screen. Among few clear exceptions to this weak correlation between A375 and HCT116 phenotypes was the long noncoding RNA SNHG1_lnc_47: it was highly expressed and translated both in A375 and HCT116 cells, but upon targeting showed a negative fold-change in A375, whereas a small positive fold-change in HCT116, indicating that it was beneficial for A375 growth, but neutral or mildly inhibitory to HCT116 (Figures S2C–S2F).

Among the 470 hits in A375, 387 (82.34%) concerned sORFs overlapping to a variable extent with canonical protein-coding ORFs, compared with only 17.73% canonical ORF overlap in the original library (Figure 1D). Hence, targeted regions overlapping with a canonical ORF were much more likely to show a growth phenotype. We considered that in these cases, it was difficult to discern loss-of-function effects of sORF versus canonical protein-coding ORF. Thus, we excluded overlapping hits from further characterization.

This yielded 83 remaining hits from the A375 screen, from which we manually curated 17 candidates for further validation, taking into account (1) supporting evidence of the sORF, (2) strength and reproducibility of the phenotype, (3) quality and

Figure 1. CRISPR screening to identify essential sORFs

(A) Schematic of the sORF-specific CRISPR screen workflow.

(B) Pie chart depicting sORF catalog by source of origin.

(C) Length diagram of ORFs (by amino acid length) in the CRISPR screen catalog, highest 15 values were omitted for visualization purposes.

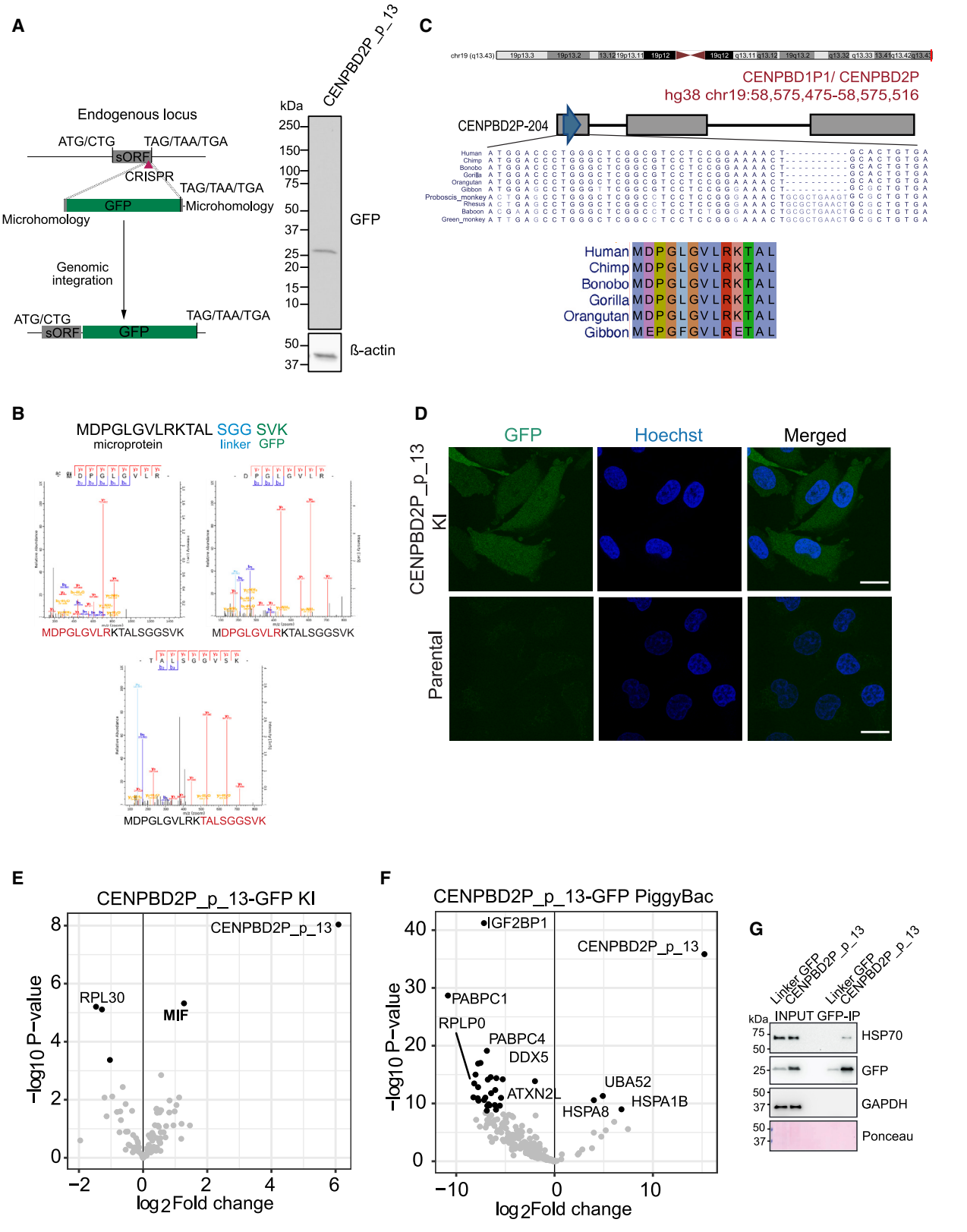
(D) Stacked bar graph illustrating canonical protein-coding (CDS) overlap of hit and non-hit sORFs, for CDS overlap any overlap (partial and full) was considered.

(E) Scatterplot summarizing the A375 10% FBS screen results, depicting log2 fold change (LFC) of the median of all sgRNAs in each replicate. Hits (light green): sORF with LFC (replicate average) ≥ 1 or ≤ -1. Candidates (green): filtered from hits based on genomic mapping, phenotypic evidence, and lack of any overlap with protein-coding regions. Top 6 hits (dark green): selected from candidates based on results in Figure 1F.

(F) Bar graph representing results of the microscopy viability assay. A375 Cas9 cells were treated with the highest scoring screen sgRNA, a multi-guide positive (RPL11, gray) and a multi-guide negative (TRAC, dark gray) control. Fold change (FC) fixed cell number (day 5 or day 6)/live cell number (day 1) was normalized to the average TRAC control FC of the respective experiment. $n = 9$ with three independent experiments. Each data point is colored according to replicate and experiment (yellow, blue, gray spectrum). Top 6 hits: dark green, all other hits: green. Error bars represent standard deviation, * p value <0.05, compared to TRAC control and calculated by two-tailed unpaired Student's t test, no asterisk indicates non-significance.

(G) Table summarizing the top six hits.

(H) Likely Ensembl transcripts of origin for each of the top six candidates. Dark green arrow: sORF, gray arrow: canonical ORF (if present), gray bars: exons. See also Figures S1 and S2 and Tables S1, S2, S3, S4, S5, and S6.



(legend on next page)

number of sgRNAs, and (4) expression and translation evidence (Figures 1E, S2E, and S2F; Table S6). We systematically named the 17 candidates by their gene ID, location with respect to a canonical ORF (if present) or otherwise their transcript biotype, and putative amino acid length. All 17 candidates were also (mildly) downregulated in at least two replicates of any of the K562 viability screens and 15 out of 17 candidates displayed mild downregulation in the HCT116 screens, indicating that our shortlisted candidates displayed a similar phenotypic effect across the cell lines tested, with the exception of SNHG1_lnc_47, already alluded to previously, and ZFAS1_lnc_53, both of which were mildly upregulated in HCT116 (Figures S2E and S2F). We also manually inspected additional hits found in the HCT116 screen (Figure S2E), but this did not yield additional strong candidates. Hence, we chose to characterize the sORF candidates in A375 cells.

To validate the screening phenotype in an independent experiment, we turned to an imaging-based growth assay. To this end, we transfected A375-Cas9 cells with synthetic sgRNAs and followed cellular growth in a 96-well plate for five to six days. We chose the sgRNAs with the highest and most consistent effect size in the original screens to individually validate each of the 17 hits. As a positive control, we used a multi-sgRNA targeting the essential ribosomal gene RPL11, while as a negative control we utilized a multi-sgRNA targeting the TRAC locus, which is not expressed in melanoma cells. For 16 out of 17 candidates, growth was significantly ($p < 0.05$) reduced after transfecting the synthetic sgRNA, as compared to the TRAC control (Figure 1F). We selected the six hits with the strongest phenotype for further studies (Figures 1F–1H). Most were originally reported in sORF.org. Four of the top hits were uORFs (DSE_u_63 located within ENSG00000111817, RCC1_u_20 located within ENSG00000180198, RPLP0_u_15 located within ENSG00000089157, NUTF2_u_11 located within ENSG00000102898), one originated from a pseudogene (CENPBD2P_p_13 located within ENSG00000213753) and one from a lncRNA (SNHG1_lnc_47 located within ENSG00000255717) (Figures 1G and 1H). The candidate sORFs comprised between 11 and 63 codons and started with a canonical ATG start codon apart from NUTF2_u_11 (Figure 1G). RCC1_u_20, RPLP0_u_15, NUTF2_u_11 and SNHG1_lnc_47 were located on transcripts linked to cancer cell proliferation (https://depmap.org/portal).[62,63] The lncRNA *SNHG1* for example, has been implicated in the progression of gastric cancer,[64] bladder cancer,[65] and more recently, breast cancer.[66]

CENPBD2P_p_13 is translated

Our CRISPR experiments suggested that the targeted sORF locations contained some functional element that elicited the reproducible growth defect. As our next step, we wanted to test whether our six predicted coding candidate sORFs were indeed translated. We did not observe any associated MS evidence of a sORF translation product in two tested published A375 datasets (study by Devabhaktuni et al.[67] and Pride PXD016776) and one subcellular localization[68] dataset, although it has to be noted that none of the datasets involved small protein enrichment. Further, we initially attempted to raise rabbit polyclonal antibodies, but limited possibility to select unique and highly immunogenic peptides in the short protein sequences did not allow us to derive selective and sensitive antibody sera. Hence, we resorted to an endogenous CRISPR knock-in strategy, fusing a GFP sequence lacking a start codon to the C-terminal end of the respective endogenous sORF, while retaining the endogenous stop codon (Figure 2A). We hypothesized that if the ribosome translates the sORF, it should also drive translation of the in-frame linker-GFP fusion. To carry out the knock-in, we used the microhomology-mediated end-joining based PITCh strategy, since it allows for very short homology arms (Figure 2A).[69] We co-transfected one plasmid containing the respective template; and a second plasmid containing the sORF-specific sgRNA, a linearizing sgRNA, and the Cas9 nuclease into A375 parental cells. For five out of the six chosen PITCh targets, transfection yielded GFP-expressing cells. We initially sorted the GFP-positive population and expanded the selectants. Despite the polyclonal nature of the resulting GFP-expressing lines, western blotting showed a single distinctive anti-GFP band for the sorted candidates (Figures 2A and S3A). Targeting RCC1_u_20 resulted in an over 100 kDa large product, suggesting a preferential or exclusive off-target integration of the GFP cassette. However, four of the products fell into the expected molecular weight range of the desired sORF-GFP fusion. Encouraged by this, we sorted clones for each of the candidates to assess the respective integrations. The majority proved difficult to validate due to (off-target) insertions or complications in insert amplification (Figure S3B). Correct knock-in and functional

Figure 2. CENPBD2P_p_13 is translated

(A) Schematic of knock-in strategy and western blot of GFP+ sorted polyclonal knock-in A375 cells for CENPBD2P_p_13, with β-actin as loading control. See Figure S11 for uncropped western blots.

(B) MaxQuant MS/MS spectra of the detected unique tryptic peptides for the CENPBD2P_p_13 fusion protein. Amino acid sequence of CENPBD2P_p_13, linker and the first three amino acids of GFP are indicated at the top.

(C) Chromosome location (red line), transcript location (blue arrow) and nucleotide and amino acid level Multiz alignment of CENPBD2P_p_13 in 30 mammals. Chromosome location, nucleotide alignment and phylogeny tree were output from http://genome.ucsc.edu.

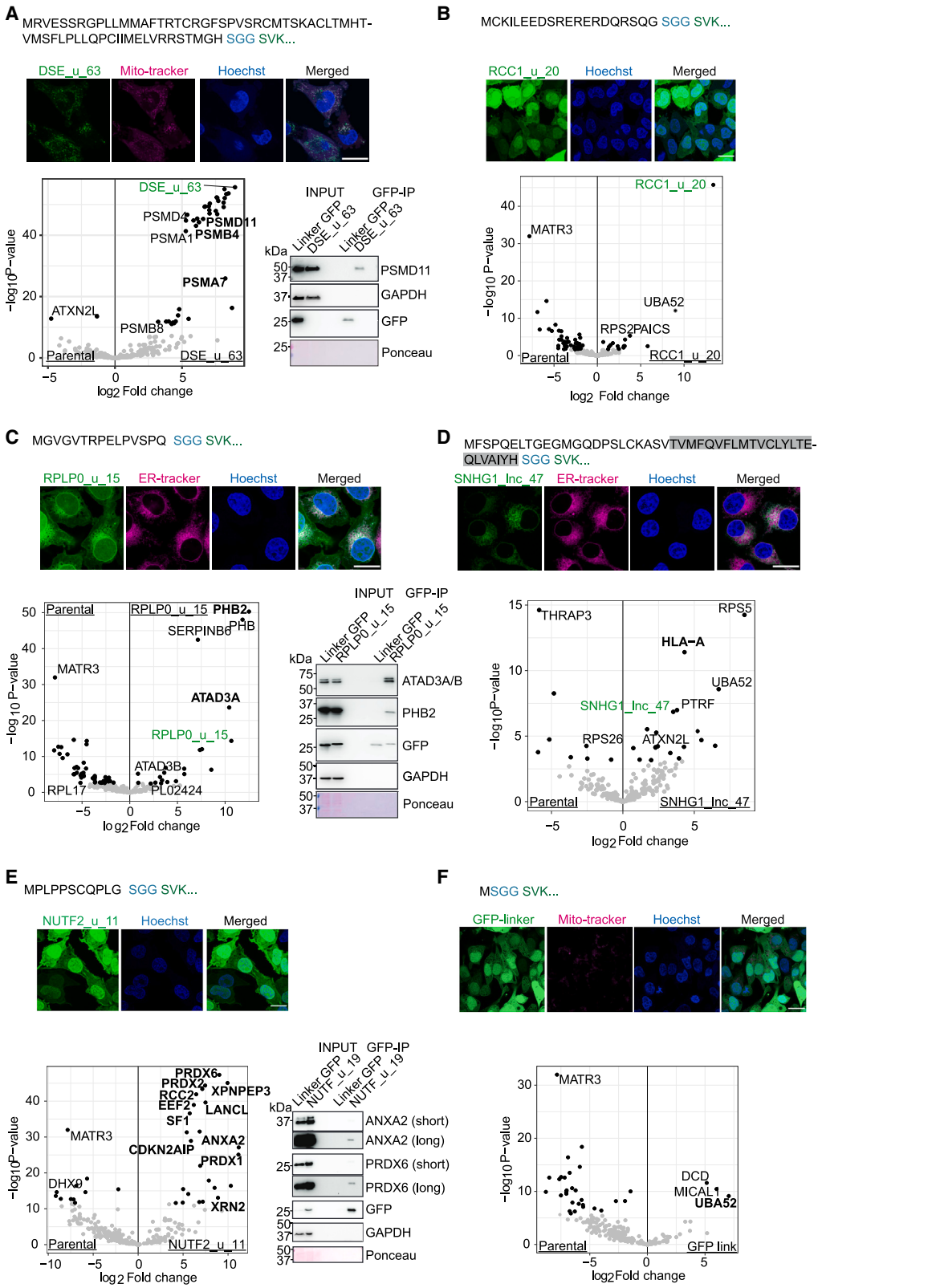
(D) Live cell confocal microscopy of A375 monoclonal CENPBD2P_p_13-GFP knock-in cells and parental controls, with identical laser and Fiji settings. green: CENPBD2P_p_13-GFP, blue: HOECHST 33342. Shown is a single plane from a z-stack. Scale bar represents 20 μm.

(E) Volcano plot showing results of mass-spec coupled GFP-trap coIP of 2 biological and 3 technical replicates from CENPBD2P_p_13 knock-in cells versus A375 parental cell line. Hit threshold: LFC ≥1, DEP adjusted p value <0.05. Additional independent batch replicate shown in Figure S4C.

(F) Volcano plot showing results of mass-spec coupled GFP-trap coIP of 2 batch and 3 technical replicates from CENPBD2P_p_13-GFP expressing A375 Cas9 cell lines versus parental Cas9 cell line. Hit threshold: LFC ≥1, DEP adjusted p value <0.05. Additional independent biological replicate shown in Figure S4C.

(G) Western blot of coIP from CENPBD2P_p_13-GFP expressing A375 Cas9 cell lines and a control cell line expressing only the linker-GFP moiety, with GAPDH as loading control.

See Figure S11 for uncropped western blots. See also Figures S3–S4 and Table S7.



(legend on next page)

translation could be confirmed only for CENPBD2P_p_13. Firstly, western blot (Figure 2A) and flow cytometry (Figure S3C) confirmed expression of the GFP fusion. We performed a pull-down for GFP followed by mass spectrometry and were able to detect peptides tiling the microprotein-GFP-fusion in this assay (Figure 2B). Secondly, both parental and knock-in alleles could be detected via genomic PCR and Sanger sequencing indicating that the knock-in is heterozygous (Figures S3D and S3E). Thus, we could confirm previous predictions that CENPBD2P_p_13 is indeed translated. While originally named as a pseudogene of the CENPB DNA binding domains, we did not find any homology between the 8.3 kb CENPBD2P transcript and CENPB, nor between the CENPBD2P transcript and the pseudogene *CENPBD1P* (Figure S4A). Specifically, no similarities to the sORF encoded in the 5′ region of CENPBD2P were found. A BLAST search against the full human transcriptome yielded only one hit, the ribosomal pseudogene *RPL23AP7*, but the region of similarity was in a repeat-rich 3′ end of CENPBD2P (Figure S4A). Hence, *CENPBD2P* is a bona fide sORF-encoding gene that did not arise through pseudogenization.

CENPBD2P_p_13-GFP displays diffuse localization with discernible protein interactors

A multiple sequence alignment suggested that CENPBD2P_p_13 is a primate-specific sORF (Figure 2C). The endogenous CENPBD2P_p_13-GFP fusion showed diffuse subcellular localization when imaged by confocal live cell microscopy (Figure 2D). A search for linear motifs predicted a potential PDZ-binding and phosphorylation site in the C-terminal LRKTAL hexapeptide (Figure S4B). To address a putative biological activity of the CENPBD2P_p_13 microprotein, we set out to identify possible interaction partners through MS-coupled co-immunoprecipitation (coIP) with anti-GFP nanobody-coupled beads, followed by tryptic digest and liquid chromatography-tandem mass spectrometry (LC-MS/MS). Parental A375 cells served as an experimental control. The bait microprotein was robustly detected in the knock-in cell line, but only one additional protein, macrophage migration inhibitory factor (MIF) was significantly but mildly enriched in two independent batch experiments (Figures 2E and S4C). We considered that we may be able to capture substoichiometric and/or more transient interactors by increasing the bait level, for which we generated a transgenic, CENPBD2P_p_13-GFP-expressing cell line using PiggyBac integration. With this A375 cell line, we once again performed MS-coupled coIP. In this overexpression setting, we pulled down protein chaperones but not MIF (Figures 2F and S4C). An interaction with Hsp70/HSPA1 was also confirmed by western blot (Figure 2G). Hsp70 proteins are known to bind to unfolded proteins, hence we hypothesize that, when overexpressed, CENPBD2P_p_13 may be a client for these chaperones, rather than a regulator.

sORF products display distinct localization

To investigate the localization and interaction partners of the five remaining candidates for which we were not successful in generating a knock-in, we created stable cell lines expressing a GFP-fusion of the respective microprotein using the PiggyBac transposase system in our A375-Cas9 screening cell line, despite our concerns of adding a large molecular weight. Of note, we in parallel attempted to generate stable cell lines with HA-tagged constructs but were not able to detect any signal after antibiotic selection (Figure S5C). We were also unable to generate rabbit polyclonal antibodies against the native sequences that were specific enough for immunoprecipitation. After confirming GFP expression of the transfected cell lines, we performed confocal microscopy (Figures 3A–3F). RCC1_u_20 (Figure 3B) and NUTF2_u_11 (Figure 3E) were diffusely distributed across the cytosol and nucleus, similar to the linker-GFP itself (Figures 3F, S5A, and S5B). In contrast, the remaining three candidates showed specific subcellular distributions: DSE_u_63 resembles mitochondrial localization (Figure 3A), RPLP0_u_15 shows localization in line with the ER/Golgi (Figure 3C), and SNHG1_lnc_47 possesses a vesicular appearance (Figure 3D). These localizations are especially notable since none of the putative microproteins contained predicted specific subcellular targeting motifs (Figure S6A), with the exception of a predicted transmembrane region in SNHG1_lnc_47 (Figure S6B). We again performed MS-coupled coIP with the five stable cell lines and an A375-Cas9 control to find putative interaction partners of the GFP-fusion proteins (Figures 3A–3F and S7A–S7F). Of these, only RCC1_u_20 did not yield interaction partners. The remaining four candidates reproducibly pulled down specific sets of proteins reproducibly across three independent experiments (Table S7).

DSE_u_63 interacts with proteasomal components

CoIP with DSE_u_63-GFP yielded many components of the proteasome core, such as PSMA7, PSMB4, and PSMD11 (Figures 3A and S7A). We tested and confirmed via western blotting the interaction with PSMD11, which was absent in a control pull down from linker-GFP expressing cells. Further studies will have to show whether this is due to DSE_u_63 playing a functional proteasome-related role (given the large portion of proteasomal

Figure 3. GFP-putative microprotein fusions display distinct localizations and interaction partners

(A–F) Microscopy and coIP with A375 Cas9 cells stably expressing the respective codon-altered GFP-fused putative microprotein or GFP control. Amino acid sequence of the respective putative microprotein with predicted transmembrane region indicated for SNHG1_lnc_47 (gray) in (D) (also see Figure S6B). Fixed confocal microscopy images of each cell line. Green: microprotein-GFP, blue: HOECHST 33342. red: ER-Tracker/Mito-Tracker. Shown is a single plane from a z-stack. Scale bars represents 20 μm. coIP/MS: GFP-trap coIP with PBS washing conditions. One of three independent experiments shown (each performed with two biological replicates and three injections). Black dots represent significantly enriched proteins, with LFC $\geq$1.5, DEP adjusted p value <0.05. Bold labels indicate proteins that were significantly enriched in 3/3 independent pulldown experiments (2/2 for linker GFP). Gray dots represent non-significant identified proteins. Additional replicates (independent batches of coIP and MS) shown in Figure S7. CoIP western validation: GFP-trap coIP with PBS washing conditions, followed by SDS-PAGE and western blots with the indicated antibodies.

See Figure S11 for uncropped western blots. See also Figures S5 and S6 and Table S7.

components being pulled down) or it being simply degraded by the proteasome, the latter of which would be in agreement with DSE_u_63-GFP being lower expressed compared to the linker-GFP alone as well as other microprotein fusions (comparatively weaker signal in microscopy of Figure S5A and invisible as a bait in the standard exposure western blot of Figure 3A). Further, its specific subcellular localization overlapping with mitochondria may hint at DSE_u_63 having additional mitochondrial binding partners, which we observed in one out of three pull-down experiments, but did not further verify (Figure S7A; Table S7).

RPLP0_u_15 interacts with ATAD3 and the prohibitin complex

Despite its localization being more diffuse, spanning the ER, nuclear membrane and nucleus, coIP of RPLP0_u_15-GFP yielded two mitochondrial complexes, ATAD3A/B and prohibitin (PHB/PHB2), across three independent experiments (Figures 3C and S7B). ATAD3A/B is known to reside predominantly in the inner mitochondrial membrane, but has also been reported to operate at ER-mitochondrial junctions and in the ER itself.[70] The mitochondrial PHB complex has been shown to interact with the ER and maintain its homeostasis, with depletion of this complex triggering the unfolded protein response.[71] ATAD3A/B binds to PHB, and this complex mediates the organization and maintenance of mitochondrial DNA (mtDNA).[72,73] Our western blot is in agreement with these associations, and comparison of IP and input signals also suggested that RPLP0_u_15 pulled down a larger fraction of cellular ATAD3A/B than PHB2 (Figure 3C). This could indicate that ATAD3A/B is a direct interactor of RPLP0_u_15. In summary, our assessment into the localization and interaction partners of RPLP0_u_15 supports an ER/mitochondrial function for this microprotein.

SNHG1_lnc_47 interactome is consistent with a function in the ER

CoIP of SNHG_lnc_47-GFP confirmed specific pulldown of the bait microprotein alongside one common interactor in three independent experiments: the major histocompatibility complex transmembrane protein HLA-A (Figures 3D and S7E). HLA-A was highly expressed in our A375 RNA sequencing (RNA-seq) data and has been shown to be abundant in A375 cells also on the protein level.[74] ATAD3A was also observed in 2/3 SNHG_lnc_47 pull-down experiments, and various ribosomal proteins were pulled down sporadically. SNHG_lnc_47 contains a predicted transmembrane helix (Figure S6B), supporting both the co-localization with ER-Tracker as well as the putative interaction with the transmembrane protein HLA-A, which is synthesized in the ER and trafficked to the plasma membrane (Figures 3D and S7E).[74]

NUTF2_u_11 interactome

NUTF2_u_11 had a diffuse localization in both cytosolic and nuclear compartments, but also seemed to concentrate at the plasma membrane or cytoskeletal structures near the plasma membrane (Figure 3E). Three reproducible interactors of this extremely short, 11-amino acid microprotein candidate were ANXA2, XPNPEP3, and PRDX6, next to other peroxiredoxin family members (Figure S7C). ANXA2 is a calcium and phospholipid binding protein[75] while PRDX6 is an enzyme with both GSH peroxidase and phospholipase A2 activities.[76] Both ANXA2 and PRDX6 are known to associate with the plasma membrane[77] and, intriguingly, ANXA2 has been shown to interact with peroxiredoxins.[78] Interaction of NUTF2_u_11 with both of these proteins could be confirmed by western blot, suggesting that even our shortest microprotein candidate is capable of having specific interactions in the cell (Figure 3E).

Complementation with GFP-tagged, but not untagged, microprotein partially rescues viability phenotype

Since specific subcellular localization and interactomes support, but do not prove, the functional relevance of our putative sORF-encoded microproteins, we sought to directly test if the microproteins expressed in *trans* could rescue a disruption of their corresponding endogenous sORF. This was particularly important since two of the microprotein candidates with specific, reproducible, interaction partners-DSE_u_63 and NUTF2_u_11-did not score evidence of translation in our RibORF analysis (Figure S2D). Interestingly, four out of six sORF candidate transcripts were implicated in cancer cell proliferation (RCC1_u_20, RPLP0_u_15, NUTF2_u_11, and SNHG1_lnc_47) (https://depmap.org/portal).[62,63] We designed rescue constructs, in which the sORF codons were exchanged to their most common alternative. This made the exogenous sORFs resistant to sgRNAs targeting the endogenous locus and ensured that any rescue would be based on the encoded microprotein and not on a functional RNA sequence element. Placement of a selectable marker (*Neo*) under an IRES guaranteed expression of the sORF-encoding mRNA. After generating knock-in cell lines with the codon-altered constructs in a Cas9 background, we co-cultured the rescue cell lines with A375 Cas9 cells stably expressing GFP, and analyzed the single cell GFP+/GFP− ratio via flow cytometry (Figure S8A). In case of rescue, cells expressing the exogenous microprotein (GFP−) should display a growth advantage over the control cells (GFP+) upon treatment with the matched sgRNA but not any of the mismatched sgRNAs (Figure 4A left panel). Indeed, a growth bias toward the rescue cells could be observed in the RPL11 positive control (Figure 4B, rightmost panel); however, not for any of the top six sORF candidates (Figure 4B). Since we had no means of detecting our untagged microproteins in this assay, the possibility remained that the sORFs were not efficiently translated from the supplied mRNA. Expressing the respective microprotein-GFP fusions would allow us to interrogate a phenotypic rescue while being able to validate for expression of the microprotein via the GFP fluorescence. Thus, we co-cultured GFP-tagged microprotein cell lines with non-GFP A375 Cas9 parental cells, again assessing the single cell GFP+/GFP− ratio via flow cytometry (Figure 4A, right panel; Figure S8B). Here, a rescue effect should be apparent by an increase of the GFP+ population. Again, the positive RPL11 control construct was able to rescue its matched sgRNA condition (Figure 4C, rightmost panel). Additionally, we also observed a significant rescue of the respective sgRNA knock-out phenotypes with GFP-tagged DSE_u_63 as well as NUTF2_u_11, but not any of the other microprotein candidates. Surprisingly both DSE_u_63 and NUTF2_u_11 also showed a rescue trend for the RPL11 sgRNA, possibly because expression of the microprotein-GFP conferred a general

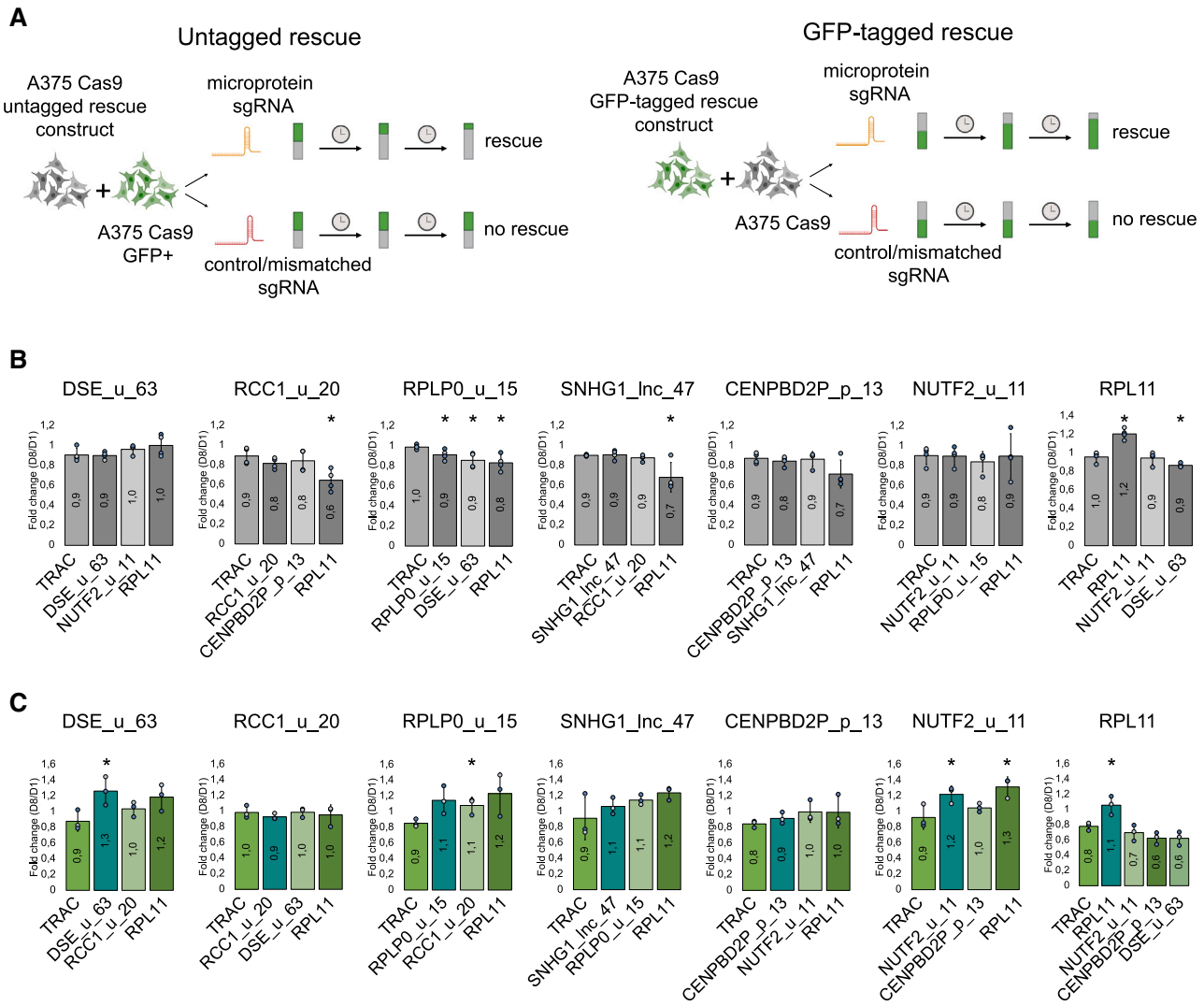


Figure 4. GFP-tagged, but not untagged, sORF products can mildly rescue the viability phenotype

(A) Schematic of untagged (left panel) or tagged (right panel) co-culture rescue experiments.

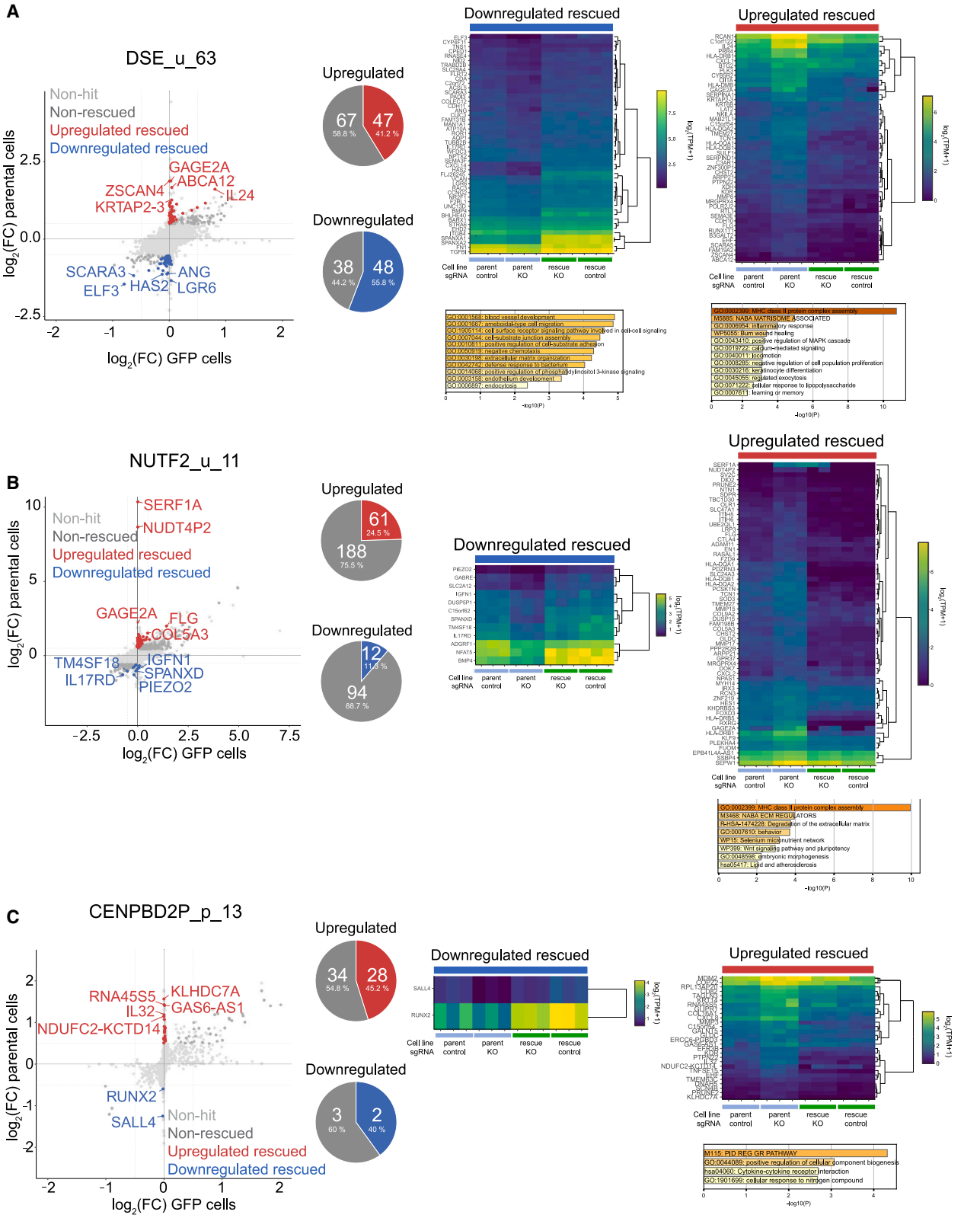
(B) Bar plots showing outcome of the untagged rescue experiment for each of the putative microprotein cell lines and RPL11 control cell line (rightmost panel). A375 Cas9 cells stably expressing the respective untagged sgRNA-resistant microprotein candidate were co-cultured with parental A375 Cas9 cells stably expressing GFP. Each co-culture was treated with the negative control sgRNA (TRAC), the matched microprotein sgRNA and two mismatched sgRNAs (targeting a different sORF candidate and RPL11). Shown is the fold change (day 8/day 1 after sgRNA-transfection) in the GFP-negative fraction measured by flow cytometry analysis. $n = 4$ with 3 independent experiments, data points are colored by experiment.

(C) Bar plots showing outcome of the GFP-tagged rescue experiment for each of the putative microprotein cell lines and RPL11 control cell line (rightmost panel). A375 Cas9 cells stably expressing the respective GFP-tagged sgRNA-resistant microprotein candidate were co-cultured with parental A375 Cas9 cells. Each co-culture was treated with the negative control sgRNA (TRAC), the matched microprotein sgRNA and two mismatched sgRNAs (targeting a different sORF candidate and RPL11). Shown is the fold change (day 8/day 1 after sgRNA-transfection) in the GFP+ fraction measured by flow cytometry analysis. $n = 3$ with 2 independent experiments (for SNHG1_lnc_47), $n = 3$ with 3 independent experiments (for all others), data points are colored by experiment. For (B) and (C), all error bars indicate standard deviation. All p values were calculated by two-tailed unpaired Student's t test compared to TRAC, $^*p < 0.05$, no asterisk indicates non-significance.

proliferation/pro-survival advantage that also helped to ameliorate the strong growth phenotype conferred by RPL11 targeting.

Partial rescue of transcriptional perturbations in rescue cell lines

To further confirm the observed rescue phenotype, and to understand which cellular functions might be altered upon putative microprotein knock-out, we wanted to employ a more sensitive readout, and subsequently turned to RNA-seq experiments. We transfected both parental A375-Cas9 and microprotein-GFP fusion derivative cell lines with the respective sORF-targeting or TRAC control sgRNA in triplicates and carried out RNA-seq analysis three days post-transfection. We subsequently performed DESeq2 differential expression analysis contrasting



(legend on next page)

target and TRAC sgRNA in each cell line background and identified differentially expressed genes (DESeq2 adjusted p value ≤ 0.05, LFC ≥ 0.5 or ≤ -0.5) that were rescued in the respective microprotein-GFP expressing cell line (rescued by $\geq$ 0.5 LFC). We excluded highly variable transcripts (expression variability 10% or more among TRAC sgRNA controls).

We first turned our attention to the DSE_u_63 sORF: DSE_u_63 sgRNA treatment induced a significant upregulation of 114 genes, and downregulation of 86 genes (Figure 5A). Roughly half of these alterations (41.2% in the upregulated, 55.8% in the downregulated fraction) could be diminished in the DSE_u_63 GFP cell line, with rescued downregulated genes being enriched in cell migration and adhesion GO (gene ontology) terms (Figure 5A). Cell adhesion was also enriched in the rescued upregulated fraction, although the top scoring terms were associated with antigen processing and presentation. Hence, the deregulated gene sets did not provide an obvious mechanistic link to the apparent mitochondrial localization of the DSE_u_63 GFP fusion microprotein (Figure 3A). Nonetheless, our RNA-seq analysis corroborated that the DSE_u_63-GFP microprotein fusion supplied in *trans* could phenotypically and transcriptionally partially rescue an endogenous sORF knock-out. We also inspected the gene sets not rescued by the microprotein-GFP fusion: the leading upregulated GO terms were p53 downstream effects (Figure S9A). We attributed this to sgRNA-induced genotoxic effects independent of sORF-targeting and/or off-target effects. Indeed, one of the DSE_u_63 sgRNA's predicted off-targets is the essential gene *RPS5* and upon inspection, we did indeed see downregulation of the RPS5 transcript (Figures S10C and S10D).

Next, we investigated the effects of NUTF2_u_11 knock-out and whether they could be ameliorated in the respective GFP fusion cell line (Figure 5B). Treatment with the NUTF2_u_11 sgRNA induced upregulation of 249, and downregulation of 106 genes. 61 and 12 of these genes respectively were rescued by the NUTF2_u_11-GFP fusion. GO analysis did not yield any enrichment for the rescued downregulated fraction. The rescued upregulated fraction was enriched in ECM-connected GO terms, which could fit with the cell adhesion-associated NUTF2_u_11 interactors (Figure 3D). Additionally, major histocompatibility complex (MHC) class II complex assembly associated genes were enriched, similarly to DSE_u_63 (Figure 5B). Indeed, some overlap existed in the rescuable upregulated genes between DSE_u_63 and NUTF2_u_11, including several in the HLA-D locus (Figure S10A). While this explained the common MHC GO term, the source of this shared set of differential genes remained unclear. We considered that the TRAC sgRNA used as control for all experiments could target and downregulate these loci, but they were not among obvious TRAC off-targets predicted by CRISPOR.[60,61] Conversely, a large number of up- and downregulation events in the NUTF2_u_11 knock-out could not be rescued, and these were enriched in GO terms relating to ribosome biogenesis, various signal, and stress pathways (Figure S9B). Since NUTF2_u_11 is located upstream of a canonical ORF encoding the essential nuclear transport protein NUTF2, we wondered if our sORF-targeting sgRNA also perturbed the canonical NUTF2. Indeed, the NUTF2 RNA was one of the downregulated transcripts, which could not be rescued by the NUTF2_u_11-GFP microprotein fusion (Figure S9B). This suggested that targeting the NUTF2_u_11 uORF caused destabilization of the entire transcript. A reduction of NUTF2 expression may in turn affect nuclear transport and processes that depend on it, such as ribosome biogenesis. This highlights the difficulty in separating uORF function from that of the transcript or the canonical ORF product, especially if the canonical protein is also essential.

Finally, we also tested a rescue of CENPBD2P_p_13 knock-out, given that we observed endogenous production of the CENPBD2P_p_13 microprotein (Figure 2). Sixty-two genes were significantly upregulated, and 5 genes were significantly downregulated upon CENPBD2P_p_13 sgRNA-mediated knock-out. Of these, 28 and 2 transcript levels could be restored, respectively, in the CENPBD2P_p_13-GFP rescue cell lines (Figure 5C). GO terms did not reach high significance, given the small number of differential genes rescued. Upregulated non-rescued genes were enriched in p53 downstream pathways as in the case of DSE_u_63 previously (Figure S9C). Interestingly, the CENPBD2P pseudogene itself was among the non-rescued downregulated genes (Figure S9C), again suggesting that targeting the sORF can affect transcript stability. In the case of canonical ORFs, nonsense-mediated decay is thought to be responsible for a downregulation of the transcript. In the light of this and the fact that CENPBD2P_p_13 microprotein in *trans* cannot rescue the growth defect, we speculate that the translation of CENPBD2P_p_13 sORF may be required to stabilize the CENPBD2P transcript. Interestingly, Chen et al. targeted a different, 79-amino acid sORF downstream of CENPBD2P_p_13 and also observed a mild growth defect,[34] further corroborating that the CENPBD2P pseudogene transcript may possess a function in cell growth. Analysis of published high-resolution Ribo-seq in K562 provides additional evidence for translation of our 13-amino acid sORF candidate, the downstream 79-amino acid sORF from Chen et al., as well as a third, 51-amino acid, sORF with canonical ATG start codon between

Figure 5. RNA-seq rescue experiments with GFP-tagged codon-altered putative microproteins

(A–C) Left panels: scatterplot showing LFC expression levels (sORF-targeting sgRNA vs. TRAC sgRNA) in GFP-tagged rescue A375 Cas9 cells and parental A375 Cas9 cells. Genes that were upregulated or downregulated in A375 Cas9 cells and rescued in microprotein-GFP expressing A375 Cas9 cells are indicated in red and blue, respectively. Their fraction of the total altered genes is illustrated in a pie chart. Middle panels: heatmap of genes that were downregulated upon treatment of A375 Cas9 cells with the respective ORF sgRNA and that could be rescued in microprotein-GFP expressing A375 Cas9 cells. Bar graph of Metascape GO term analysis (where possible). Right panels: heatmap of genes that were upregulated upon treatment of A375 Cas9 cells with the respective ORF sgRNA and that could be rescued in microprotein-GFP expressing A375 Cas9 cells. Bar graph of Metascape GO term analysis. For all panels: hits: genes displaying an average (of 2 replicates) LFC of ≥ 0.5 or ≤ -0.5 between microprotein sgRNA and TRAC sgRNA treated A375 Cas9 cells and showing an average expression variability of $\leq 10\%$ between the TRAC conditions (microprotein-GFP vs. parental cells) and a DeSeq2 adjusted p value of ≤ 0.05, rescued hits: same as hits but additionally showing a difference in LFC of ≥ 0.5 between microprotein-GFP and parental cell lines.

See also Figures S9 and S10.

these two proposed ORFs (Figure S3F).[34,79] Further investigation will be necessary to discern coding versus RNA function of this interesting pseudogene.

Overall, RNA-seq analysis revealed sORF knock-out specific gene perturbations, and expression of the microprotein in *trans* rescued these gene expression changes to a variable extent. Gene sets perturbed downstream of the putative microproteins did not yield an immediate mechanistic hypothesis for how the putative microprotein may work, and also did not have a clear relation with the microprotein localization and protein interactions determined (Figure 3). Even though the different approaches used here to pin down a possible cellular function did not converge on a coherent mechanistic picture, by studying multiple microprotein candidates from a large-scale CRISPR screen targeting sORFs, we find unique cellular localizations, interactomes, and transcriptional perturbations associated with the putative microproteins.

DISCUSSION

In the human genome, sORFs with coding potential exceed canonical protein-coding ORFs by 1–2 orders of magnitude. Ample evidence exists for their pervasive translation but the extent to which sORF-encoded microproteins contribute to cellular function is still unknown. Here, we carried out a sORF-specific large scale functional CRISPR-Cas9 screen targeting 11,776 sORFs with predicted coding potential, from which we selected candidates for detailed characterization. Assaying growth as a readout turned out to be challenging due to the narrow effect size of transformed cell lines, which are highly selected for fast and robust growth. Our attempts to challenge the cells with low serum did not yield additional candidates. We compared our candidate list with two other CRISPR screens targeting non-canonical ORFs.[34,35] Prensner et al. 2021 also surveyed A375 cells, among others, and identified 15 non-canonical ORFs, knock-out of which led to a growth phenotype, with a similar effect size ($\log_2$-fold change ≤ -1) to our screen.[35] The proportion of hits was considerably higher compared to our study, given that their library targeted only 553 ORFs. The ORFs included in the latter study were larger on average (median length 74 amino acids, minimum 23 amino acids). While 57 out of these 553 sORFs were also contained in our library (Figure S1C, Table S5), only one relatively long sORF, ASNSD1_u_96, was a common hit between our studies. In our hands, however, the growth phenotype of ASNSD1_u_96 was not reproduced in A375 cells in a screen-independent growth assay (Figure 1F). The ASNSD1_u_96 microprotein has since been described as a component of the PAQosome,[80] involved in medulloblastoma cell survival.[38] None of our other 16 candidates was included in the 553 ORFs.[35] Conversely, our library contained only one of the other 14 downregulated ORFs in the Prensner screen ([35]: TCONS_l2_00007040, this screen: PL01912), targeting of which did not score any phenotype in our screen. Another CRISPR screen targeting 2353 ORFs was also reported by Chen et al., albeit not in A375 cells.[34] While 460 sORFs overlapped between the Chen screen and our screen (Table S5), none of the respective hits overlapped.

The annotation of sORFs has tremendously improved in the past few years, and a large community effort collected a set of 7264 sORFs with high-confidence Ribo-seq signals for submission to GENCODE.[2] SNHG1_lnc_47 is the only top hit from our screen included in this catalog (now UniProt predicted protein A0A024R548). CENPBD2P_p_13 and RPLP0_u_15 were smaller than the 16-amino acid consortium threshold, but were present respectively in two and three of the published Ribo-seq datasets utilized by the consortium.

We initially derived CENPBD2P_p_13 from the sORFs.org database.[58,59] CRISPR-targeting the CENPBD2P_p_13 reduced the RNA levels of the entire transcript. This, together with the observed inability of the CENPBD2P_p_13 microprotein to rescue a sgRNA-mediated knock-out in *trans* suggested that the growth defect observed in the original screen may relate to a non-coding function of the CENPBD2P RNA and/or derive from one of the alternative ORFs on this transcript. Since we were able to validate CENPBD2P_p_13 synthesis, one could speculate that translation of CENPBD2P_p_13 sORF may influence RNA stability and/or translation of other ORFs. Hence CENPBD2P_p_13 may be a regulatory sORF, and together with potential alternative functional elements of the *CENPBD2P* gene awaits further functional characterization.

The DSE_u_63 sORF originates from a bioinformatics study mining genomes for evolutionary signatures of protein coding sORFs,[12] but in contrast to most of our sORF library, no prior Ribo-seq or mass spectrometry evidence exists for the DSE_u_63 microprotein. While the sORF showed one of the strongest growth defects initially, interpretations of the phenotypic and RNA-seq experiments were complicated by a CRISPR off-target effect toward a ribosomal protein, Rps5. All possible DSE_u_63 sgRNAs share the same off-target region within Rps5 and this is related to the evolutionary origin of this sORF. DSE_u_63 appears to have arisen in the primate lineage after the branching of macaques and gibbons, by a duplication/integration of a stretch of DNA carrying part of the Rps5 coding sequence, reversed with respect to its original orientation in the *RPS5* gene. The full DSE_u_63 sORF including the start and stop codons can already be found in the macaque *RPS5* gene, running antisense to the *RPS5* coding sequence. Hence, the DSE_u_63 sORF appeared immediately through the inverse integration of the *RPS5* fragment into the DSE gene. It has acquired only five sense mutations up to the hominid lineage. Thus, a mitochondrial localization and interactome of DSE_u_63 is unlikely to have evolved since its birth in the primate lineage - instead, we hypothesize that by pure serendipity, a sequence encoded antisense to the *RPS5* coding region was capable of producing a mitochondrial-like localization. Since the human microprotein has acquired five mutations as compared to the gibbon homolog, it would be interesting to experimentally test if the original sequence showed the same subcellular localization and putative phenotype, or both evolved further in the higher hominid lineage.

RCC1_u_20 and RPLP0_15 were originally extracted from the Ribo-seq-based sORF database sORFs.org [58,59] with the latter currently showing some level of evidence in 30 different repository datasets. While RCC1_u_20 did not display any notable subcellular localization nor interaction partners, RPLP0_u_15 expressed as a GFP fusion demonstrated a Golgi/ER localization and interacted with a putative ATAD/PHB complex (Figure 3C),

despite its short 15-amino acid sequence. Another hit with a distinct localization is the putative 47-amino acid microprotein SNHG1_lnc_47, extracted here originally from a Ribo-seq-based study.[39] SNHG1_lnc_47 appeared to localize to vesicles when fused to GFP and contains a predicted transmembrane region. Functional follow-up studies for either of these three putative microproteins proved difficult since neither an untagged protein nor the GFP fusion could rescue the growth phenotype elicited by the respective microprotein sgRNA, and we were not able to independently verify microprotein translation, despite observing good Ribo-seq evidence (Figure S2D).

Finally, the shortest of our candidates is the putative 11-amino acid microprotein NUTF2_u_11, also extracted from sORF.org.[58,59] NUTF2_u_11 lacked RibORF translation evidence, but the short length of the sORF may pose challenges for detecting a significant ribosome footprint and periodicity. NUTF2_u_11 showed specific interactors, likely occurring at the plasma membrane (Figure 3E). The growth defect of the NUTF2_u_11 sgRNA could be significantly rescued with a NUTF2_u_11 microprotein-GFP fusion expressed in *trans,* albeit only a minority of transcriptomic changes could be reverted. As with CENPBD2P_p_13, we observed downregulation of the underlying transcript as a result of targeting the NUTF2_u_11 sORF. This underscores the difficulty of separating sORF activity from other functional elements of the same transcript and the need for suitable rescue experiments.

Limitations of the study

A limitation of this study is the use of GFP fusions for the elucidation of our microprotein candidates. We resorted to this approach because microprotein fusions with smaller tags, specifically HA and FLAG, which we generated in parallel, could not be detected. This could suggest that the native microproteins may be unstable or suboptimally expressed from our plasmids, potentially lacking regulatory sequences that boost translation of the endogenous sORF. It is also likely that a well-folded GFP can protect a microprotein from degradation and hence greatly increase its steady state levels. This may explain why we only observed functional rescue in *trans* when we expressed the transgene microprotein as a GFP fusion. At the same time, we are aware that adding a large tag is not necessarily physiological and may distort or disrupt native microprotein localization and/or microprotein-protein interactions. The lack of strong interaction partners for some of the microproteins may reflect a possible perturbation by GFP. Interesting to us was the fact that several small microproteins could localize to specific compartments, even in the context of a GFP tag. Overall, in the absence of specific antibodies, finding the right tag for microprotein studies remains a major challenge in the field.

Additionally, failure of complementation experiments with the microprotein coding sequence supplied in *trans*, as observed with a number of our candidates in this study, raises a number of questions (assuming sufficient sgRNA activity and expression of the rescue constructs); for example, does sORF disruption lead to a destabilization of the entire RNA or RNA elements (e.g. secondary structures, RNA-binding protein binding sites), hence also directly or indirectly affecting non-coding RNA functions? Or does the targeted genomic region contain regulatory elements (such as enhancers or insulators) unrelated to the RNA species? These questions are beyond the scope of our study but would be exciting to explore in future endeavours. Further, a more fine-grained sgRNA tiling strategy including sgRNAs upstream and downstream of the sORF[34,35] can help to increase the confidence of the hits in the initial screen, though does not substitute rescue experiments. We note, that this may be more difficult for very short sORFs as studied here, which possess limited choice of selective and efficient sgRNAs.

Overall, our study aims to highlight further avenues of research and technical complexities associated with sORF-specific screening. We identified six candidate sORFs, and tested interaction partners and localizations for their encoded microproteins. Rescue experiments implicated two of these putative small proteins in cell proliferation, and RNA-seq experiments demonstrated a partial rescue of gene expression perturbations for three candidates, though for all tested hits, sORF knock-out also induced knockdown of an off-target or the transcript of origin with potential functions of their own. We could demonstrate endogenous translation of the CENPBD2P_p_13 microprotein, and our data suggest that its pseudogene of origin, *CENPBD2P*, is involved in cell proliferation. Our work contributes to the growing number of putative microproteins by adding characterization of various sORF candidates, and translation evidence of a pseudogene-derived microprotein.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents used in this study should be directed to the lead contact Simon Johannes Elsässer (simon.elsasser@scilifelab.se).

Materials availability

Constructs generated in this study are deposited in Addgene. Additional reagents are listed in the key resources table.

Data and code availability

- Mass spectrometry primary data have been deposited on MassIVE under https://doi.org/10.25345/C5DZ03B70 and are publicly available as of the date of publication.
- RNA-seq primary and processed data have been deposited on GEO under GSE232375 and are publicly available as of the date of publication.
- Associate code is available at GitHub (https://github.com/elsasserlab/schlesinger_microprotein_screen).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

D.S. was supported by a Boehringer Ingelheim Fonds PhD fellowship. T.F.M. acknowledges financial support from NIH grant K01CA249038 and the University of California Drug Discovery Consortium. S.J.E. acknowledges funding by the Karolinska Institutet SFO for Molecular Biosciences, Vetenskapsrådet (2015-04815, 2020-04313), H2020 ERC Starting Grant (715024 RAPID), Åke Wibergs Stiftelse (M15-0275), Cancerfonden (22 2354 Pj), the Ming Wai Lau Center for Reparative Medicine, the Ragnar Söderbergs Stiftelse, Stiftelsen för Strategisk Forskning (FFL7), and the Knut och Alice Wallenbergs Stiftelse (2010–0215). CRISPR screening was conducted at the SciLifeLab CRISPR Functional Genomics unit (CFG) at Karolinska Institutet, funded by Science for Life Laboratory. CRISPR and genomics analyses were processed on the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) provided by the National Academic Infrastructure for

Supercomputing in Sweden (NAISS) under projects NAISS 2023/6-19, NAISS 2023/22-84, and SNIC 2022/6-14. NAISS is funded by the Swedish Research Council through grant agreement no. 2022-06725. Graphical Abstract, and parts of Figures 1A and 4A were generated with BioRender (www.biorender.com). We thank Dr. Rui Branca for sharing a list of unannotated peptide candidates from proteogenomics experiments. We thank Prof. Iris Finkemeier and Paulina Heinkow for mass spectrometry service. We thank the BIC facility for enabling usage of their AiryScan microscopy, the Fernandez Capetillo group for allowing us to use their In Cell microscope, the Bartek and Lemmens groups for access to the Nikon Eclipse Ti2. We thank Alberto M. Arenas and Birthe Meineke for proofreading this manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, D.S. and S.J.E.; methodology, D.S., C.N., L.L., J.E., T.F.M., and S.J.E.; investigation, D.S., C.D., C.N., L.L., A.S., G.L.R., G.M.-S.T., J.E., T.F.M., and S.J.E.; formal analysis, D.S., C.N., G.M.-S.T., and S.J.E.; writing – original draft, D.S. and S.J.E.; writing – review and editing, D.S., G.L.R., and S.J.E.; funding acquisition, S.J.E.; resources, S.J.E.; supervision, S.J.E.

DECLARATION OF INTERESTS

T.F.M. is a paid consultant for and holds equity in Velia Therapeutics and is a paid consultant for Ono Pharma USA. S.J.E. holds equity in Epigenica Ab.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - sORF catalogue assembly
  - CRISPR screening
  - Cell culture maintenance
  - Cell line generation
  - Microscopy viability assay
  - Knock-in generation
  - Western blotting
  - Co-immunoprecipitation experiments
  - Genomic location and conservation alignments
  - Confocal microscopy
  - Motif searches
  - Gene ontology analysis
  - Venn diagrams
  - RibORF analysis
  - Rescue assays
  - RNA-sequencing
  - Immunofluorescence microscopy
  - Synthetic sgRNAs, primers and plasmids
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.111884.

Received: May 24, 2023
Revised: October 11, 2024
Accepted: January 21, 2025
Published: January 23, 2025

REFERENCES

1. Wright, J.C., Mudge, J., Weisser, H., Barzine, M.P., Gonzalez, J.M., Brazma, A., Choudhary, J.S., and Harrow, J. (2016). Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Nat. Commun. *7*, 11778. https://doi.org/10.1038/ncomms11778.
2. Mudge, J.M., Ruiz-Orera, J., Prensner, J.R., Brunet, M.A., Calvet, F., Jungreis, I., Gonzalez, J.M., Magrane, M., Martinez, T.F., Schulz, J.F., et al. (2022). Standardized annotation of translated open reading frames. Nat. Biotechnol. *40*, 994–999. https://doi.org/10.1038/s41587-022-01369-0.
3. Delcourt, V., Brunelle, M., Roy, A.V., Jacques, J.F., Salzet, M., Fournier, I., and Roucou, X. (2018). The Protein Coded by a Short Open Reading Frame , Not by the Annotated Coding Sequence , Is the Main Gene Product of the Dual-Coding Gene MIEF1. Mol. Cell. Proteom. *17*, 2402–2411. https://doi.org/10.1074/mcp.RA118.000593.
4. Cao, X., Khitun, A., Luo, Y., Na, Z., Phoodokmai, T., Sappakhaw, K., Olatunji, E., Uttamapinant, C., and Slavoff, S.A. (2021). Alt-RPL36 downregulates the PI3K-AKT-mTOR signaling pathway by interacting with TMEM24. Nat. Commun. *12*, 508. https://doi.org/10.1038/s41467-020-20841-6.
5. Denli, A.M., Narvaiza, I., Kerman, B.E., Pena, M., Benner, C., Marchetto, M.C.N., Diedrich, J.K., Aslanian, A., Ma, J., Moresco, J.J., et al. (2015). Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity. Cell *163*, 583–593. https://doi.org/10.1016/j.cell.2015.09.025.
6. Boix, O., Martinez, M., Vidal, S., Giménez-Alejandre, M., Palenzuela, L., Lorenzo-Sanz, L., Quevedo, L., Moscoso, O., Ruiz-Orera, J., Ximénez-Embún, P., et al. (2022). pTINCR microprotein promotes epithelial differentiation and suppresses tumor growth through CDC42 SUMOylation and activation. Nat. Commun. *13*, 6840. https://doi.org/10.1038/s41467-022-34529-6.
7. Lewandowski, J.P., Dumbović, G., Watson, A.R., Hwang, T., Jacobs-Palmer, E., Chang, N., Much, C., Turner, K., Kirby, C., Rubinstein, N.D., et al. (2020). The Tug1 Locus is Essential for Male Fertility. Genome Biol. *21*, 562066. https://doi.org/10.1101/562066.
8. Zheng, X., Wang, M., Liu, S., Chen, H., Li, Y., Yuan, F., Yang, L., Qiu, S., Wang, H., Xie, Z., and Xiang, M. (2023). A lncRNA-encoded mitochondrial micropeptide exacerbates microglia-mediated neuroinflammation in retinal ischemia/reperfusion injury. Cell Death Dis. *14*, 126. https://doi.org/10.1038/s41419-023-05617-2.
9. Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature *541*, 228–232. https://doi.org/10.1038/nature21034.
10. Lauressergues, D., Couzigou, J.-M., Clemente, H.S., Martinez, Y., Dunand, C., Bécard, G., Combier, J.-P., et al. (2015). Primary transcripts of microRNAs encode regulatory peptides. Nature *520*, 90–93. https://doi.org/10.1038/nature14346.
11. Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs , 5 ' UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife *4*, e08890. https://doi.org/10.7554/eLife.08890.
12. Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., and Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. Genome Biol. *16*, 179. https://doi.org/10.1186/s13059-015-0742-x.
13. Makarewich, C.A., and Olson, E.N. (2017). Mining for Micropeptides. Trends Cell Biol. *27*, 685–696. https://doi.org/10.1016/j.tcb.2017.04.006.
14. Pauli, A., Norris, M.L., Valen, E., Chew, G.-L., Gagnon, J.A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., et al. (2014). Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. Science *343*, 1248636. https://doi.org/10.1126/science.1248636.
15. Chng, S.C., Ho, L., Tian, J., and Reversade, B. (2013). ELABELA: A hormone essential for heart development signals via the apelin receptor. Dev. Cell *27*, 672–680. https://doi.org/10.1016/j.devcel.2013.11.002.

16. Bohère, J., Mancheno-ferris, A., Hayek, S.A., Zanet, J., Valenti, P., Akino, K., Yamabe, Y., Inagaki, S., Chanut-delalande, H., Plaza, S., et al. (2018). Shavenbaby and Yorkie mediate Hippo signaling to protect adult stem cells from apoptosis. Nat. Commun. *9*, 5123. https://doi.org/10.1038/s41467-018-07569-0.

17. Ho, L., Tan, S.Y.X., Wee, S., Wu, Y., Tan, S.J.C., Ramakrishna, N.B., Chng, S.C., Nama, S., Szczerbinska, I., Chan, Y.S., et al. (2015). ELABELA Is an Endogenous Growth Factor that Sustains hESC Self-Renewal via the PI3K/AKT Pathway. Cell Stem Cell *17*, 435–447. https://doi.org/10.1016/j.stem.2015.08.010.

18. Yu, R., Hu, Y., Zhang, S., Li, X., Tang, M., Yang, M., Wu, X., Li, Z., Liao, X., Xu, Y., et al. (2022). LncRNA CTBP1-DT-encoded microprotein DDUP sustains DNA damage response signalling to trigger dual DNA repair mechanisms. Nucleic Acids Res. *50*, 8060–8079. https://doi.org/10.1093/nar/gkac611.

19. Martinez, T.F., Lyons-Abbott, S., Bookout, A.L., De Souza, E.V., Donaldson, C., Vaughan, J.M., Lau, C., Abramov, A., Baquero, A.F., Baquero, K., et al. (2023). Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. Cell Metab. *35*, 166–183.e11. https://doi.org/10.1016/j.cmet.2022.12.004.

20. Chugunova, A., Loseva, E., Mazin, P., Mitina, A., Navalayeu, T., and Bilan, D. (2019). LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism. Proc. Natl. Acad. Sci. USA *116*, 1–6. https://doi.org/10.1073/pnas.1809105116.

21. Stein, C.S., Jadiya, P., Zhang, X., McLendon, J.M., Abouassaly, G.M., Witmer, N.H., Anderson, E.J., Elrod, J.W., and Boudreau, R.L. (2018). Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. Cell Rep. *23*, 3710–3720.e8. https://doi.org/10.1016/j.celrep.2018.06.002.

22. Makarewich, C.A., Baskin, K.K., Munir, A.Z., Bezprozvannaya, S., Sharma, G., Khemtong, C., Shah, A.M., McAnally, J.R., Malloy, C.R., Szweda, L.I., et al. (2018). MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. Cell Rep. *23*, 3701–3709. https://doi.org/10.1016/j.celrep.2018.05.058.

23. Polycarpou-Schwarz, M., Groß, M., Mestdagh, P., Schott, J., Grund, S.E., Hildenbrand, C., Rom, J., Aulmann, S., Jo, H.-P.S., Vandesompele, J., et al. (2018). The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. Oncogene *37*, 4750. https://doi.org/10.1038/s41388-018-0281-5.

24. Zhang, S., Reljić, B., Liang, C., Kerouanton, B., Francisco, J.C., Peh, J.H., Mary, C., Jagannathan, N.S., Olexiouk, V., Tang, C., et al. (2020). Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. Nat. Commun. *11*, 1312. https://doi.org/10.1038/s41467-020-14999-2.

25. Lee, C., Zeng, J., Drew, B.G., Sallam, T., Martin-Montalvo, A., Wan, J., Kim, S.J., Mehta, H., Hevener, A.L., De Cabo, R., and Cohen, P. (2015). The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab. *21*, 443–454. https://doi.org/10.1016/j.cmet.2015.02.009.

26. Arnoult, N., Correia, A., Ma, J., Merlo, A., Garcia-Gomez, S., Maric, M., Tognetti, M., Benner, C.W., Boulton, S.J., Saghatelian, A., and Karlseder, J. (2017). Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. Nature *549*, 548–552. https://doi.org/10.1038/nature24023.

27. Slavoff, S.A., Heo, J., Budnik, B.A., Hanakahi, L.A., and Saghatelian, A. (2014). A human short open reading frame (sORF)-Encoded polypeptide that stimulates DNA end joining. J. Biol. Chem. *289*, 10950–10957. https://doi.org/10.1074/jbc.C113.533968.

28. Hung, P.J., Johnson, B., Chen, B.R., Byrum, A.K., Bredemeyer, A.L., Yewdell, W.T., Johnson, T.E., Lee, B.J., Deivasigamani, S., Hindi, I., et al. (2018). MRI Is a DNA Damage Response Adaptor during Classical Non-homologous End Joining Article MRI Is a DNA Damage Response Adaptor during Classical Non-homologous End Joining. Mol. Cell *71*, 332–342.e8. https://doi.org/10.1016/j.molcel.2018.06.018.

29. Chu, Q., Martinez, T.F., Novak, S.W., Donaldson, C.J., Tan, D., Vaughan, J.M., Chang, T., Diedrich, J.K., Andrade, L., Kim, A., et al. (2019). Regulation of the ER stress response by a mitochondrial microprotein. Nat. Commun. *10*, 4883. https://doi.org/10.1038/s41467-019-12816-z.

30. Schlesinger, D., and Elsässer, S.J. (2022). Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. FEBS J. *289*, 53–74. https://doi.org/10.1111/febs.15769.

31. Hanada, K., Higuchi-Takeuchi, M., Okamoto, M., Yoshizumi, T., Shimizu, M., Nakaminami, K., Nishi, R., Ohashi, C., Iida, K., Tanaka, M., et al. (2013). Small open reading frames associated with morphogenesis are hidden in plant genomes. Proc. Natl. Acad. Sci. USA *110*, 2395–2400. https://doi.org/10.1073/pnas.1213958110.

32. Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.C., Yang, H., Carter, C.D., Wheeler, D., Davis, R.W., Boeke, J.D., et al. (2006). Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res. *16*, 365–373. https://doi.org/10.1101/gr.4355406.7.

33. Guo, X., Chavez, A., Tung, A., Chan, Y., Kaas, C., Yin, Y., Cecchi, R., Garnier, S.L., Kelsic, E.D., Schubert, M., et al. (2018). High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR-Cas9 in yeast. Nat. Biotechnol. *36*, 540–546. https://doi.org/10.1038/nbt.4147.

34. Chen, J., Brunner, A.D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., and Weissman, J.S. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140–1146. https://doi.org/10.1126/science.aay026.

35. Prensner, J.R., Enache, O.M., Luria, V., Krug, K., Clauser, K.R., Dempster, J.M., Karger, A., Wang, L., Stumbraite, K., Wang, V.M., et al. (2021). Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat. Biotechnol. *39*, 697–704. https://doi.org/10.1038/s41587-020-00806-2.

36. Zheng, C., Wei, Y., Zhang, P., Lin, K., He, D., Teng, H., Manyam, G., Zhang, Z., Liu, W., Lee, H.R.L., et al. (2023). CRISPR–Cas9-based functional interrogation of unconventional translatome reveals human cancer dependency on cryptic non-canonical open reading frames. Nat. Struct. Mol. Biol. *30*, 1878–1892. https://doi.org/10.1038/s41594-023-01117-1.

37. Zheng, C., Wei, Y., Zhang, P., Xu, L., Zhang, Z., Lin, K., Hou, J., Lv, X., Ding, Y., Chiu, Y., et al. (2023). CRISPR/Cas9 screen uncovers functional translation of cryptic lncRNA-encoded open reading frames in human cancer. J. Clin. Invest. *133*, e159940. https://doi.org/10.1172/JCI159940.

38. Hofman, D.A., Ruiz-Orera, J., Yannuzzi, I., Murugesan, R., Brown, A., Clauser, K.R., Condurat, A.L., van Dinter, J.T., Engels, S.A.G., Goodale, A., et al. (2024). Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. Mol. Cell *84*, 261–276.e18. https://doi.org/10.1016/j.molcel.2023.12.003.

39. Bazzini, A.A., Johnstone, T.G., Christiano, R., MacKowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., and Giraldez, A.J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. *33*, 981–993. https://doi.org/10.1002/embj.201488411.

40. Gascoigne, D.K., Cheetham, S.W., Cattenoz, P.B., Clark, M.B., Amaral, P.P., Taft, R.J., Wilhelm, D., Dinger, M.E., and Mattick, J.S. (2012). Pinstripe: A suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. Bioinformatics *28*, 3042–3050. https://doi.org/10.1093/bioinformatics/bts582.

41. Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A., Kellis, M., and Saghatelian, A. (2014). Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J. Proteome Res. *13*, 1757–1765. https://doi.org/10.1021/pr401280w.

42. Ma, J., Diedrich, J.K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., Yates, J.R., and Saghatelian, A. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. Anal. Chem. *88*, 3967–3975. https://doi.org/10.1021/acs.analchem.6b00191.

43. Raj, A., Wang, S.H., Shim, H., Harpak, A., Li, Y.I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife *5*, e13328. https://doi.org/10.7554/eLife.13328.

44. Razooky, B.S., Obermayer, B., and Tarakhovsky, A. (2017). Viral Infection Identifies Micropeptides Differentially Regulated in smORF-Containing lncRNAs. Genes *8*, 206. https://doi.org/10.3390/genes8080206.

45. Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. Elife *3*, e03523. https://doi.org/10.7554/eLife.03523.

46. Samandi, S., Roy, A.V., Delcourt, V., Lucier, J.F., Gagnon, J., Beaudoin, M.C., Vanderperre, B., Breton, M.A., Motard, J., Jacques, J.F., et al. (2017). Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. Elife *6*, e27860. https://doi.org/10.7554/eLife.27860.

47. Schwaid, A.G., Shannon, D.A., Ma, J., Slavoff, S.A., Levin, J.Z., Weerapana, E., and Saghatelian, A. (2013). Chemoproteomic discovery of cysteine-containing human short open reading frames. J. Am. Chem. Soc. *135*, 16750–16753. https://doi.org/10.1021/ja406606j.

48. Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64. https://doi.org/10.1038/nchembio.1120.

49. Vanderperre, B., Lucier, J.F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.M., and Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. PLoS One *8*, e70698. https://doi.org/10.1371/journal.pone.0070698.

50. Zhang, P., He, D., Xu, Y., Hou, J., Pan, B.F., Wang, Y., Liu, T., Davis, C.M., Ehli, E.A., Tan, L., et al. (2017). Genome-wide identification and differential analysis of translational initiation. Nat. Commun. *8*, 1749. https://doi.org/10.1038/s41467-017-01981-8.

51. Anderson, D.M., Anderson, K.M., Chang, C.L., Makarewich, C.A., Nelson, B.R., McAnally, J.R., Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R., and Olson, E.N. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell *160*, 595–606. https://doi.org/10.1016/j.cell.2015.01.009.

52. Anderson, D.M., Makarewich, C.A., Anderson, K.M., Shelton, J.M., Bezprozvannaya, S., Bassel-Duby, R., and Olson, E.N. (2016). Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. Sci. Signal. *9*, ra119. https://doi.org/10.1126/scisignal.aaj1460.

53. Quinn, M.E., Goh, Q., Kurosaka, M., Gamage, D.G., Petrany, M.J., Prasad, V., and Millay, D.P. (2017). Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. Nat. Commun. *8*, 15665. https://doi.org/10.1038/ncomms15665.

54. D'Lima, N.G., Ma, J., Winkler, L., Chu, Q., Loh, K.H., Corpuz, E.O., Budnik, B.A., Lykke-Andersen, J., Saghatelian, A., and Slavoff, S.A. (2017). A human microprotein that interacts with the mRNA decapping complex. Nat. Chem. Biol. *13*, 174–180. https://doi.org/10.1038/nchembio.2249.

55. Huang, J.Z., Chen, M., Chen, D., Gao, X.C., Zhu, S., Huang, H., Hu, M., Zhu, H., and Yan, G.R. (2017). A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. Mol. Cell *68*, 171–184.e6. https://doi.org/10.1016/j.molcel.2017.09.015.

56. Toggas, S.M., Krady, J.K., and Billingsley, M.L. (1992). Molecular neurotoxicology of trimethyltin: identification of stannin, a novel protein expressed in trimethyltin-sensitive cells. Mol. Pharmacol. *42*, 44–56.

57. Vanderperre, B., Staskevicius, A.B., Tremblay, G., McCoy, M., O'Neill, M.A., Cashman, N.R., and Roucou, X. (2011). An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. FASEB J. *25*, 2373–2386. https://doi.org/10.1096/fj.10-173815.

58. Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2015). SORFs.org: A repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *44*, D324–D329. https://doi.org/10.1093/nar/gkv1175.

59. Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2018). An update on SORFs.org: A repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *46*, D497–D502. https://doi.org/10.1093/nar/gkv1175.

60. Concordet, J.P., and Haeussler, M. (2018). CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res. *46*, W242–W245. https://doi.org/10.1093/nar/gky354.

61. Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J., Schneider-maunoury, S., Shkumatava, A., Teboul, L., Kent, J., et al. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. *17*, 1–12. https://doi.org/10.1186/s13059-016-1012-2.

62. Thin, K.Z., Tu, J.C., and Raveendran, S. (2019). Long non-coding SNHG1 in cancer. Clin. Chim. Acta *494*, 38–47. https://doi.org/10.1016/j.cca.2019.03.002.

63. DepMap Broad (2024). Current DepMap Release data, including CRISPR Screens, PRISM Drug Screens, Copy Number, Mutation, Expression, and Fusions. DepMap 24Q2 Public. Figshare+. Dataset. https://doi.org/10.25452/figshare.plus.25880521.v1.

64. Wang, S., Han, H., Meng, J., Yang, W., Lv, Y., and Wen, X. (2021). Long non-coding RNA SNHG1 suppresses cell migration and invasion and upregulates SOCS2 in human gastric carcinoma. Biochem. Biophys. Rep. *27*, 101052. https://doi.org/10.1016/j.bbrep.2021.101052.

65. Cai, H., Xu, H., Lu, H., Xu, W., Liu, H., Wang, X., Zhou, G., and Yang, X. (2022). LncRNA SNHG1 Facilitates Tumor Proliferation and Represses Apoptosis by Regulating PPARγ Ubiquitination in Bladder Cancer. Cancers *14*, 4740. https://doi.org/10.3390/cancers14194740.

66. Deng, L., Wang, J., Song, J., Wu, Q., Gong, Z., Song, J., and Hou, L. (2024). Long noncoding RNA SNHG1 promotes breast cancer progression by regulating the miR-641/RRS1 axis. Sci. Rep. *14*, 3265. https://doi.org/10.1038/s41598-024-52953-0.

67. Devabhaktuni, A., Lin, S., Zhang, L., Swaminathan, K., Gonzalez, C.G., Olsson, N., Pearlman, S.M., Rawson, K., and Elias, J.E. (2019). TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. Nat. Biotechnol. *37*, 469–479. https://doi.org/10.1038/s41587-019-0067-5.

68. Orre, L.M., Vesterlund, M., Pan, Y., Arslan, T., Zhu, Y., Fernandez Woodbridge, A., Frings, O., Fredlund, E., and Lehtiö, J. (2019). SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. Mol. Cell *73*, 166–182.e7. https://doi.org/10.1016/j.molcel.2018.11.035.

69. Sakuma, T., Nakade, S., Sakane, Y., Suzuki, K.I.T., and Yamamoto, T. (2016). MMEJ-Assisted gene knock-in using TALENs and CRISPR-Cas9 with the PITCh systems. Nat. Protoc. *11*, 118–133. https://doi.org/10.1038/nprot.2015.140.

70. Arguello, T., Peralta, S., Antonicka, H., Gaidosh, G., Diaz, F., Tu, Y.-T., Garcia, S., Shiekhattar, R., Barrientos, A., and Moraes, C.T. (2021). ATAD3A has a scaffolding role regulating mitochondria inner membrane structure and protein assembly. Cell Rep. *37*, 110139. https://doi.org/10.1016/j.celrep.2021.110139.

71. Lourenço, A.B., Rodríguez-Palero, M.J., Doherty, M.K., Cabrerizo Granados, D., Hernando-Rodríguez, B., Salas, J.J., Venegas-Calerón, M., Whitfield, P.D., and Artal-Sanz, M. (2021). The Mitochondrial PHB Complex Determines Lipid Composition and Interacts With the Endoplasmic Reticulum to Regulate Ageing. Front. Physiol. *12*, 696275. https://doi.org/10.3389/fphys.2021.696275.

72. Thuaud, F., Ribeiro, N., Nebigil, C.G., and Désaubry, L. (2013). Prohibitin Ligands in Cell Death and Survival: Mode of Action and Therapeutic Potential. Chem. Biol. *20*, 316–331. https://doi.org/10.1016/j.chembiol.2013.02.006.

73. Kasashima, K., Sumitani, M., Satoh, M., and Endo, H. (2008). Human prohibitin 1 maintains the organization and stability of the mitochondrial nucleoids. Exp. Cell Res. *314*, 988–996. https://doi.org/10.1016/j.yexcr.2008.01.005.

74. Peppicelli, S., Ruzzolini, J., Andreucci, E., Bianchini, F., Kontos, F., Yamada, T., Ferrone, S., and Calorini, L. (2019). Potential Role of HLA Class I Antigens in the Glycolytic Metabolism and Motility of Melanoma Cells. Cancers *11*, 1249. https://doi.org/10.3390/cancers11091249.

75. Bharadwaj, A., Bydoun, M., Holloway, R., and Waisman, D. (2013). Annexin A2 Heterotetramer: Structure and Function. Int. J. Mol. Sci. *14*, 6259–6305. https://doi.org/10.3390/ijms14036259.

76. Fisher, A.B. (2011). Peroxiredoxin 6: A Bifunctional Enzyme with Glutathione Peroxidase and Phospholipase A $_2$ Activities. Antioxid. Redox Signal. *15*, 831–844. https://doi.org/10.1089/ars.2010.3412.

77. Varyukhina, S., Lamazière, A., Delaunay, J.L., de Wreede, A., and Ayala-Sanmartin, J. (2022). The Ca2+- and phospholipid-binding protein Annexin A2 is able to increase and decrease plasma membrane order. Biochim. Biophys. Acta. Biomembr. *1864*, 183810. https://doi.org/10.1016/j.bbamem.2021.183810.

78. Talwar, D., Messens, J., and Dick, T.P. (2020). A role for annexin A2 in scaffolding the peroxiredoxin 2–STAT3 redox relay complex. Nat. Commun. *11*, 4512. https://doi.org/10.1038/s41467-020-18324-9.

79. Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Nat. Chem. Biol. *16*, 458–468. https://doi.org/10.1038/s41589-019-0425-0.

80. Cloutier, P., Poitras, C., Faubert, D., Bouchard, A., Blanchette, M., Gauthier, M.S., and Coulombe, B. (2020). Upstream ORF-Encoded ASDURF Is a Novel Prefoldin-like Subunit of the PAQosome. J. Proteome Res. *19*, 18–27. https://doi.org/10.1021/acs.jproteome.9b00599.

81. Keller, O., Odronitz, F., Stanke, M., Kollmar, M., and Waack, S. (2008). Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinform. *9*, 278. https://doi.org/10.1186/1471-2105-9-278.

82. Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M., and Liu, X.S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. *15*, 554. https://doi.org/10.1186/s13059-014-0554-4.

83. Pujar, S., O'Leary, N.A., Farrell, C.M., Loveland, J.E., Mudge, J.M., Wallin, C., Girón, C.G., Diekhans, M., Barnes, I., Bennett, R., et al. (2018). Consensus coding sequence (CCDS) database: A standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res. *46*, D221–D228. https://doi.org/10.1093/nar/gkx1031.

84. McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B.A., Karhohs, K.W., Doan, M., Ding, L., Rafelski, S.M., Thirstrup, D., et al. (2018). CellProfiler 3.0: Next-generation image processing for biology. PLoS Biol. *16*, e2005970. https://doi.org/10.1371/journal.pbio.2005970.

85. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372. https://doi.org/10.1038/nbt.1511.

86. Zhang, X., Smits, A.H., van Tilburg, G.B., Ovaa, H., Huber, W., and Vermeulen, M. (2018). Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat. Protoc. *13*, 530–550. https://doi.org/10.1038/nprot.2017.147.

87. Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., and Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. Nucleic Acids Res. *49*, W216–W227. https://doi.org/10.1093/nar/gkab225.

88. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. *50*, W276–W279. https://doi.org/10.1093/nar/gkac240.

89. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

90. Lee, B.T., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M., et al. (2022). The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. *50*, D1115–D1122. https://doi.org/10.1093/nar/gkab959.

91. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198. https://doi.org/10.1093/nar/gkz369.

92. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

93. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646. https://doi.org/10.1093/nar/gkac1000.

94. Heberle, H., Meirelles, G.V., da Silva, F.R., Telles, G.P., and Minghim, R. (2015). InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. BMC Bioinform. *16*, 169. https://doi.org/10.1186/s12859-015-0611-3.

95. Hannon, G.J. (2010). FASTX-Toolkit. https://www.encodeproject.org/software/fastx_toolkit/.

96. Cao, K., Hajy Heydary, Y., Tong, G., and Martinez, T.F. (2023). Integrated workflow for discovery of microprotein-coding small open reading frames. STAR Protoc. *4*, 102649. https://doi.org/10.1016/j.xpro.2023.102649.

97. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

98. Bergeron, D., Lapointe, C., Bissonnette, C., Tremblay, G., Motard, J., and Roucou, X. (2013). An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. J. Biol. Chem. *288*, 21824–21835. https://doi.org/10.1074/jbc.M113.472654.

99. Nelson, B.R., Makarewich, C.A., Anderson, D.M., Winders, B.R., Troupes, C.D., Wu, F., Reese, A.L., McAnally, J.R., Chen, X., Kavalali, E.T., et al. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science *351*, 271–275. https://doi.org/10.1126/science.aad4076.

100. Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J.R., Shelton, J.M., Sánchez-Ortiz, E., Bassel-Duby, R., and Olson, E.N. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. Science *356*, 323–327. https://doi.org/10.1126/science.aam9361.

101. Zhang, Q., Vashisht, A.A., O'Rourke, J., Corbel, S.Y., Moran, R., Romero, A., Miraglia, L., Zhang, J., Durrant, E., Schmedt, C., et al. (2017). The microprotein Minion controls cell fusion and muscle formation. Nat. Commun. *8*, 15664. https://doi.org/10.1038/ncomms15664.

102. Zhu, Y., Orre, L.M., Johansson, H.J., Huss, M., Boekel, J., Vesterlund, M., Fernandez-Woodbridge, A., Branca, R.M.M., and Lehtiö, J. (2018). Discovery of coding regions in the human genome by integrated

proteogenomics analysis workflow. Nat. Commun. *9*, 903. https://doi.org/10.1038/s41467-018-03311-y.

103. Branca, R.M.M., Orre, L.M., Johansson, H.J., Granholm, V., Huss, M., Pérez-Bercoff, Å., Forshed, J., Käll, L., and Lehtiö, J. (2014). HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat. Methods *11*, 59–62. https://doi.org/10.1038/nmeth.2732.

104. Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., et al. (2017). UniProt: The universal protein knowledgebase. Nucleic Acids Res. *45*, D158–D169. https://doi.org/10.1093/nar/gkw1099.

105. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191. https://doi.org/10.1038/nbt.3437.

106. Cross, B.C.S., Lawo, S., Archer, C.R., Hunt, J.R., Yarker, J.L., Riccombeni, A., Little, A.S., Mccarthy, N.J., and Moore, J.D. (2016). Increasing the performance of pooled CRISPR-Cas9 drop-out screening. Sci. Rep. *6*, 31782. https://doi.org/10.1038/srep31782.

107. Schmierer, B., Botla, S.K., Zhang, J., Turunen, M., Kivioja, T., and Taipale, J. (2017). CRISPR/Cas9 screening using unique molecular identifiers. Mol. Syst. Biol. *13*, 1–8. https://doi.org/10.1101/114355.

108. Alexaki, A., Kames, J., Holcomb, D.D., Athey, J., Santana-Quintero, L.V., Lam, P.V.N., Hamasaki-Katagiri, N., Osipova, E., Simonyan, V., Bar, H., et al. (2019). Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design. J. Mol. Biol. *431*, 2434–2441. https://doi.org/10.1016/j.jmb.2019.04.021.

109. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinf. *13*, 134.

110. Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.P., St-Denis, N.A., Li, T., Miteva, Y.V., Hauri, S., Sardiu, M.E., Low, T.Y., et al. (2013). The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. Nat. Methods *10*, 730–736. https://doi.org/10.1038/nmeth.2557.

111. Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J.A., Bukirova, D., Čalyševa, J., et al. (2020). ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res. *48*, D296–D306. https://doi.org/10.1093/nar/gkz1030.

112. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

113. Parrish, N., Hormozdiari, F., and Eskin, E. (2014). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform. *12*, 21–40. https://doi.org/10.1201/b16589.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-GFP, B-2, mouse monoclonal	Santa Cruz	sc-9996; RRID:AB_627695
Anti- β-Actin, 13E5, rabbit monoclonal	Cell Signaling	4970; RRID:AB_2223172
Secondary anti-mouse HRP	BioRad	1721011; RRID:AB_2617113
Secondary anti-rabbit HRP	BioRad	1721019; RRID:AB_11125143
Anti-PSMD11/Rpn6/S9, Rabbit pAb	Proteintech	14786-1-AP; RRID:AB_2268979
Anti-PHB2/BAP/REA, Rabbit pAb	Proteintech	12295-1-AP; RRID:AB_2164779
Anti-ATAD3A/ATAD3B/TOB3, Rabbit pAb	Proteintech	16610-1-AP; RRID:AB_2878288
Anti-Antioxidant protein 2/1 Cys PRX, Mouse mAb	Proteintech	67499-1-Ig; RRID:AB_2882723
Anti-Annexin A2, Mouse mAb	Proteintech	66035-1-Ig; RRID:AB_11045659
Anti-HSP70 (D69), Rabbit pAb	Cell Signaling	4876S; RRID:AB_2119693
Anti-GAPDH (G-9), Mouse mAb	Santa Cruz	sc-365062; RRID:AB_10847862
Anti-HA, F-7 mouse monoclonal	Santa Cruz	sc-7392; RRID:AB_627809
Anti-mouse Alexa 647	Thermo Fisher Scientific	A-21236; RRID:AB_2535805
Critical commercial assays		
QIAamp DNA Blood Maxi Kit	Qiagen	51192
In-Fusion HD cloning kit	Takara Bio	639650
T4 DNA ligase kit	Thermo Fisher Scientific	EL0016
Lipofectamine™ LTX reagents	Thermo Fisher Scientific	15338100
Phusion Flash High-Fidelity PCR Master Mix	Thermo Fisher Scientific	F-548L
GeneJET Genomic DNA Purification Kit	Thermo Fisher Scientific	K0721
RNeasy Plus Mini Kit	Qiagen	74136
Qubit RNA HS assay kit	Life Technologies	Q322852
Deposited data		
Mass spectrometry primary data	MassIVE	https://doi.org/10.25345/C5DZ03B70
RNA-Seq primary and processed data	GEO	GSE232375
Associated code	GitHub	https://github.com/elsasserlab/schlesinger_microprotein_screen
Experimental models: Cell lines		
HEK293T	ATCC	CRL-3216
A375	ATCC	CRL-1619
HCT116	ATCC	CCL-247
K562	ATCC	CCL-243
Software and algorithms		
Scipio	Max-Planck-Institute for Biophysical Chemistry	Keller et al.[81]
CRISPOR	UC Santa Cruz	Concordet et al.,[60] Haeussler et al.[61]
MAGeCK version 0.5.8 89	Dana-Farber Cancer Institute	Li et al.[82]
CCDS database version 22	NCBI	Pujar et al.[83]
CellProfiler 3.1.9	Broad Institute (MIT and Harvard)	McQuin et al.[84]
MaxQuant Version 1.6.3.4	Max Planck Institute of Biochemistry	Cox et al.[85]
DEP package 1.22.0	RStudio (Bioconductor)	Zheng et al.[86]
Expasy	SIB Swiss Bioinformatics Resource Portal	Duvaud et al.[87]

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
EMBL-EBI Clustal Omega	EMBL-EBI	Madeira et al.[88]
JalView	University of Dundee	Waterhouse et al.[89]
UCSC Genome Browser	UCSC	Lee et al.[90]
Fiji/ImageJ	https://imagej.net/software/fiji/	N/A
gProfiler version 20230223	http://biit.cs.ut.ee/gprofiler/	Raudvere et al.[91]
MetaScape version v3.5.20230501	https://metascape.org/gp/index.html#/main/step1	Zhou et al.[92]
String database version 11.5	SIB Swiss Bioinformatics Resource, CPR- Novo Nordisk Foundation Center Protein Research, EMBL	Szklarczyk et al.[93]
InterActiVenn	https://www.interactivenn.net/	Heberle et al.[94]
FASTX-toolkit	https://www.encodeproject.org/software/fastx_toolkit/	Hannon et al.[95]
RibORF v0.1	https://github.com/zhejilab/RibORF	Cao et al.[96]
DeSeq2	RStudio (Bioconductor)	Love et al.[97]

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study does not involve human samples/subjects. Human cell lines used in the study: A375 (ATCC, CRL-1619, female donor), HCT116 (ATCC, CCL-247, male donor), HEK293T (ATCC, CRL-3216, female donor), and K562 (ATCC, CCL-243, female donor) parental cells, were authenticated by the supplier and mycoplasma tested regularly using the MycoAlert™ Mycoplasma Detection Kit per manufacturer instructions (LT07-218 / 11650261, BioNordika).

METHOD DETAILS

sORF catalogue assembly

Microprotein candidates were curated from literature,[5,9,11,12,14,15,23,26,27,39–57,98–101] collaborative mass spectrometry data,[102,103] Uniprot,[104] and the sORF.org database[58,59] with the criteria described in Table S1. Each sORF was assigned a number with the first two letters indicating the peptide source (PL: literature, PM: mass spectrometry, PS: sORFs.org database, PU: Uniprot). The resulting sORF catalogue was deduplicated by amino acid sequence. Since genomic coordinates were absent in some publications and differently reported between studies, we mapped all candidates to the human genome (GRCh38) using Scipio[81] based on the amino acid sequence supplied in the original source. Since we noticed that searching for a start codon led to errors (partially due to some sORFs containing non-canonical start codons), we first removed the starting methionine, carried out the mapping and then expanded the mapped interval again by three nucleotides on the 5' end. In cases where both amino acid sequence and genomic coordinates were provided, conflicts were resolved by keeping Scipio results when at least 80% of the Scipio prediction overlapped with the original annotation, and retaining genomic coordinates supplied in the original source otherwise. The mappable sORFs were then deduplicated once more and subsequently carried through to sgRNA design (Tables S2, S3, and S4).

CRISPR screening

sgRNA design

To design sgRNAs against each of the sORFs in the catalogue, we utilised the CRISPOR design tool[60,61] specifying the sORF coordinates with an additional 20-nucleotide flanking region. We excluded sgRNAs within the lowest 20% of all cutting frequency determination (CFD) scores, as well as sgRNAs displaying more than four thymidines in a row (Figure S1A). If the sORF possessed only one targeting sgRNA, the sgRNA and sORF were dismissed from the library, while all sgRNAs were kept if the sORF could be targeted by two to eight sgRNAs. The CFD specificity score and the Doench et al. efficiency score[105] supplied by CRISPOR were each normalised to a 0 to 1 value range each (x - min(scores)) / (max(scores) - min(scores)) and subsequently added to create an aggregated score (values ranging between 0 and 2). For cases in which the microprotein could be targeted by more than eight sgRNAs, this aggregation score was used for sgRNA ranking. If various sgRNAs were less than four nucleotides apart, only the better ranked sgRNA was kept. Subsequently the best-ranking eight sgRNAs were included in the library. This resulted in a total of 50,136 sgRNAs targeting 11,776 unique microprotein candidates. Of note, since some sORFs displayed overlap, a number of sORFs with <2 (461) or >8 (1091) sgRNAs were retained, since these sgRNAs were included in the 2-8 sgRNAs targeting another ORF. Additionally, 1000 non-targeting negative control sgRNAs and 292 positive control sgRNA targeting ribosomal genes were included in the library.

Library cloning

The customised gRNA library was synthesised as oligo nucleotides by Twist Bioscience (Table S4). The array oligos were double-stranded and amplified via PCRs. The resulting PCR product included an A-U flip in the tracrRNA[106] 10-nucleotide random sequence labels (RSLs), and an i7 sequencing primer binding site.[107]

ggctttatatat**cttgtggaaaggacgaaacaccgnnnnnnnnnnnnnnnnnnnnngtttaagagctagaaatagcaagtttaaataaggct**ccgttatcaacttagtg aaaaagtggcaccgagtcggtgctttttt<u>GATCGGAAGAGCACACGTCTGAACTCCAGTCAC</u>NNNNNNNNNNaagcttggcgtaactagatcttgagac aaa

(bold: array oligo, n: sgRNA-sequence, N: RSL sequence, underlined: i7 sequence).

This construct was then cloned into the pLenti-Puro-AU-flip-3xBsmBI (Addgene #196709)[107] by Gibson assembly. After input sequencing to confirm library representation (same as gDNA sequencing described below but without PCR1), the plasmid library was packaged into the lentivirus utilising plasmids psPAX2 (a gift from Didier Trono, Addgene #12260) and pCMV-VSV-G (a gift from Bob Weinberg, Addgene #8454) in HEK-293T (ATCC). The lentivirus-containing supernatant was concentrated around 40-fold using Lenti-X concentrator (Takara), aliquoted and stored in liquid nitrogen.

Cas9 cell line generation

Cells stably expressing a codon optimized, WT SpCas9 flanked by two nuclear localisation signals and coupled via a self-cleaving peptide to a blasticidin-S-deaminase-mTagBFP fusion protein (herein called ''A375 Cas9'', ''HCT116 Cas9'' or ''K562 Cas9'') were generated by lentiviral transduction of the lenti.Cas9.BFP.Blast plasmid (Addgene #196714) and selected with 5 μg/ml blasticidin (Invivogen, ant-bl-10p). Subsequently, the cells were bulk sorted using a BD FACSAria Fusion (BD BioSciences) flow cytometer until a stable BFP-positive population was reached. Cas9 expression was additionally confirmed by Western blot.

CRISPR screening

The functional virus titre was estimated through assessment of cell survival rates after transduction of the respective cancer cell lines with different concentrations of virus and puromycin selection. For the screens, A375 Cas9, HCT116 Cas9 or K562 Cas9 cells (see ''Cas9 cell line generation'') were subjected to lentiviral transduction in duplicate at a MOI of around 0.3 and a coverage of around 1,000 cells per guide in presence of 2 μg/ml polybrene (Sigma-Aldrich, TR-1003-G). Two days after transduction, cells were selected with 2 μg/ml puromycin (VWR, CAYM13884) for five days and subsequently split into the different screening condition arms. For the essentiality screens, cells were grown for 21 days (A375 and HCT116 essentiality screens) or 28 days (K562 screens) after transduction and split every two to three days. A cell number of at least 60 million cells (around 1000 cells/sgRNA) per replicate was maintained at all times to ensure full library coverage. Cell pellets of 60 million cells were collected on day four and day 21 (A375, HCT116) or day seven and day 28 (K562) post–transduction for sequencing. For the 6-thioguanine screens, cells were treated with 30 μM 6-thioguanine (Sigma-Aldrich, A4882) in NaOH (stock solution 30mM 6-thioguanine in 1M NaOH) starting on day 10 post-transduction. The medium was subsequently changed every two to three days and cells were collected on day 28 post-transfection. Genomic DNA was extracted from the pellets with the QIAamp DNA Blood Maxi Kit (Qiagen). sgRNA and RSL sequences were amplified by three PCRs as described in[107] but with altered PCR2_FW and PCR3_fw primers. The amplicon was sequenced on Illumina NovaSeq, reading 20 cycles Read 1 with the CRISPRSEQ primer, 10 cycles index read i7 to read the RSL, and six cycles index read i5 for the sample barcode.

CRISPR screen analysis

The NGS data from the CRISPR screens was analysed using the MAGeCK software,[82] version 0.5.8.[105] Log fold change (LFC) was calculated between sequencing reads of D4/D21 (A375, HCT116 essentiality), D7/D28 (K562 essentiality) or D28 control/D28 6-TG treatment (A375, K562 6-TG screens) with sORFs displaying a LFC $\leq$ -1 or LFC $\geq$ 1 considered to be hits. Full results can be found in Tables S2 and S3. For summary statistics, e.g. in Figures 1B–1D, the sORFs were further deduplicated, retaining only unique sequences/genomic coordinates. To test for protein-coding (CDS) overlap, we intersected the sORF coordinates with the CCDS database version 22.[83]

Cell culture maintenance

Human cell lines A375 (ATCC, CRL-1619, female donor), HCT116 (ATCC, CCL-247, male donor), HEK293T (ATCC, CRL-3216, female donor), and K562 (ATCC, CCL-243, female donor) parental cells were authenticated by the supplier and mycoplasma tested regularly using the MycoAlert™ Mycoplasma Detection Kit per manufacturer instructions (LT07-218 / 11650261, BioNordika). A375, HEK293T, and HCT116 were grown in Gibco DMEM (Thermo Fisher Scientific, 10569010) supplemented with 10% FBS (Sigma-Aldrich, F7524). K562 cells were cultured in RPMI-1640 (Sigma-Aldrich, R8758) with 10% FBS (fetal bovine serum) (Sigma-Aldrich, F7524). During the CRISPR screening, the cells were maintained in 1x Pen-Strep (Sigma-Aldrich, P4333) at all times.

Cell line generation

To generate A375 Cas9 cells (see ''Cas9 cell line generation'') that additionally stably express GFP, we transduced the cells with lentivirus generated from pCDH_EF1_GFP_IRES_Puro. For viral packaging and production, the transfer plasmid was co-transfected with the envelope plasmid pMD2.G (Addgene #12259) and the second-generation packaging plasmid psPAX2 (Addgene #12260) into HEK293T cells (ATCC) using transIT®-LT1 (Mirus Bio). The growth media was replenished one day later. Viral particles were harvested 48- and 72-hours post-transfection and filtered through a 0.45 μm mixed cellulose esters syringe filter (Millipore). The first transduction was carried out on 40% confluent A375 Cas9 cells, seeded one day prior, by adding equivalent amounts of fresh

DMEM and viral supernatant. 2 μg/ml polybrene (Sigma-Aldrich) was added to the growth media to boost efficiency of transduction. This was repeated with the second viral harvest and eight hours after the second transduction, cells were selected for five days in presence of 3 μg/ml puromycin (VWR, CAYM13884). A near 100% GFP-positive population was confirmed via microscopy (ZOE Fluorescent Cell Imager, BioRad), and flow cytometry (Navios flow cytometer, Beckman Coulter).

To generate microprotein rescue cells lines, the endogenous sORF nucleotide sequence was extracted, and every codon changed to the most common (using the codon efficiency table from[108]). ATG was used as the start codon and TAG TAA were used as two subsequent stop codons for each of the constructs. The corresponding oligos were ordered from Twist BioScience and cloned into an untagged (System BioSciences, PB533A-2), GFP-tagged (pPB_EF1_MCS-EGFP_IRES_Puro), or HA-tag backbone (pPB_EF1_MCS-HA_IRES_Puro) plasmid. Stable cell lines were generated using the PiggyBac Transposase system, in which A375 Cas9 cells were co-transfected with the rescue plasmids as well as the Super PiggyBac Transposase Expression Vector (System Bioscience Inc, PB200PA) at a 4:1 ratio utilising Lipofectamine™ LTX reagents (Thermo Fisher Scientific, 15338100) according to the manufacturer's instructions. After 48 hours, cells were selected with 1-2 mg/ml G418 (Sigma-Aldrich, G8168) (untagged rescue plasmids) or 10 μg/ml puromycin (VWR, CAYM13884) (tagged rescue plasmids) for 5-7 days.

Microscopy viability assay

Viability assays were carried out as three independent experiments with triplicate wells each (9 replicates total). A375 Cas9 cells were seeded at low confluency (2000 cells per well) into a 96-well plate (Falcon, 353219). One day after seeding, cells were transfected with the respective synthetic sgRNAs (Synthego) using Lipofectamine™ RNAiMAX transfection reagent (Thermo Fisher Scientific, 13778075) according to the manufacturers' instructions. The medium was changed one day after transfection and live cells were imaged on an IN Cell Analyzer 2200 (GE Healthcare) at 4x magnification before the plate was returned to the incubator and cells were allowed to expand. Once they reached near confluency (day 4-6), cells were fixed and stained in a 4% formaldehyde (Thermo Fisher Scientific, 28908) and 2 μM Hoechst 33342 (Thermo Fisher Scientific, 62249) in PBS solution for 20 minutes, washed three times with PBS, and then imaged in PBS with the IN Cell Analyzer 2200 (GE Healthcare) at 4x magnification. Cell numbers were counted via CellProfiler 3.1.9.[84] Briefly, cells were segmented on the basis of Cas9-BFP fluorescence (Day 1) or Hoechst fluorescence (Day 4-6), using the MaskImage function with customised masks, Gaussian filter smoothing, division illumination correction, speckle enhancement, and adaptive three class thresholding with the Otsu method. Fold changes D4-6/Day 1 were normalised to the average triplicate TRAC fold change of each experiment. The average normalised fold change of the 9 replicates was then plotted in the figure, and significance compared to TRAC control tested using a two-tailed unpaired student's t-test.

Knock-in generation

Plasmid generation

Knock-ins were performed using the PITCh method as previously described.[69] Briefly sORF-specific template plasmids were created by AccuPrime Supermix PCR (Thermo Fisher Scientific, 12344040) using the pPB_EF1_MCS-EGFP_IRES_Puro as a PCR template and utilising primers with the respective knock-in-overhangs (ordered from Sigma-Aldrich). Gel-extracted PCR products (Thermo Fisher Scientific, K0692) were then cloned into the MluI (Thermo Fisher Scientific, FD0564)-digested pCRIS-PITChv2-FBL plasmid (Addgene #63672) using the In-Fusion HD cloning kit (Takara Bio, 639650). To create the sgRNA-plasmid backbone, PITCh sgRNA from pX330S-2-PITCh (Addgene #63670) was cut-out via Eco31I (Thermo Fisher Scientific, FD0293) and inserted into the pX330A-1x2 backbone (Addgene #58766) via a Golden gate assembly (T4 DNA ligase, NEB, M0202). sORF-specific sgRNAs were designed using the CRISPOR tool,[60,61] and ordered as DNA oligos (forward and reverse) from Thermo Fisher Scientific. The oligos (1:1 ratio) were phosphorylated and annealed in one step, using T4 PNK (Thermo Fisher Scientific, EK0032) according to the manufacturer's recommendations but altering the incubation conditions (37°C for 20 minutes, 75°C for 10 minutes, 95°C for 5 minutes, 95°C - 25°C decrease at 6°C per minute). The thus-generated inserts were then ligated into the BpiI (Thermo Fisher Scientific, FD1014)-digested px330A-1x2_ PITCh_sgRNA plasmid using the T4 DNA ligase kit (Thermo Fisher Scientific, EL0016). For SNHG1_lnc_47, we did not find any suitable sgRNA that was U6-compatible. Thus, we ordered a synthetic sgRNA (Synthego) for this knock-in.

Transfection

1.6 million A375 parental cells were seeded. Two days after seeding, cells were double transfected with sgRNA and template vector plasmids at a 2:1 ratio using the Lipofectamine™ LTX reagents (Thermo Fisher Scientific, 15338100) per manufacturer's recommendation. Cells were expanded and bulk-sorted with the Sony SH8000 sorter two or three times to enrich for GFP+ cells (protein extracts of these polyclonal cells are shown in Figure 2A S3A), before being sorted as single cells into 96-well plates. Clones were trial-screened for positive PCR products by making mirror plates and subjecting one of the plates to cell lysis (50mM Tris-HCl pH 8, 1mM EDTA, 0.5% Tween-20, 50-80 μg/ml Proteinase K Merck-3115801001) at 37°C overnight. The resulting crude gDNA was transferred to a PCR plate, heated to 95°C for 10 min to inactivate the PK and performing genomic PCR with Phusion Mix (Thermo Fisher Scientific, F-548L) as described below but at 35 cycles and visualised with GelGreen (VWR, 730-1535) on a 1% agarose gel (run for 30 minutes at 135 V) using an ImageQuant™ LAS 500 (GE Lifesciences). Positive clones were then expanded from the remaining plate, with genomic PCR and sequencing performed as outlined below.

For the SNHG1_lnc_47 knock-in, transfections of template (DNA) and sgRNA (RNA) were done in a sequential manner: Seeding was as above and two days after seeding, we transfected the template plasmid using Lipofectamine™ LTX reagents (Thermo Fisher Scientific, 15338100). The medium was changed and synthetic sgRNA transfected using Lipofectamine™ RNAiMAX transfection

reagent (Thermo Fisher Scientific, 13778075) one day later. Expanding and sorting was done as described above (though no GFP+ cells were observed).

Genomic PCR and sequencing

Genomic DNA was extracted with the GeneJET Genomic DNA Purification Kit (Thermo Scientific, K0721) according to the manufacturer's recommendation. Genomic PCR primers were designed using the Primer-BLAST tool[109] and ordered from Thermo Fisher Scientific. Genomic PCR of 20 cycles was then performed with the respective primer pair on the extracted DNA utilising the Phusion Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific, F-548L) as described by the manufacturer. PCR products were visualised on a 1% agarose gel (run for 30 minutes at 135 V) with GelGreen (VWR, 730-1535) using an ImageQuant™ LAS 500 (GE Lifesciences), excised and DNA extracted using the GeneJet Gel extraction kit (Thermo Fisher Scientific, K0692). PCR products were subsequently sequenced by using the primers that were also utilized in the genomic PCR and Eurofins Genomics sequencing service.

Western blotting

Cells were pelleted at 300 x g for 5 minutes, pellets washed once with PBS (Sigma-Aldrich, D8537), spun again and then resuspended in ice-cold RIPA-SDS buffer (50 mM Tris-HCl pH 7.6, 150 mM NaCl, 0.25 % sodium deoxycholate, 1mM EDTA, 1% NP-40, 0.1% SDS), supplemented with Complete Protease inhibitor (Sigma-Aldrich, 5056489001). To isolate the soluble protein fraction, protein extracts were centrifuged at maximum speed for ten min at 4°C in a table-top centrifuge and the supernatant was extracted. SDS buffer (final concentration: 62.5 mM Tris pH 6.8, 2% SDS, 10% glycerol, 0.1M DTT, 0.01% bromophenol blue) was added, the sample boiled for five minutes at 95°C and size-separated by SDS-PAGE using a 4–20% Mini-PROTEAN® TGX™ Gel (Biorad). Proteins were transferred onto a nitrocellulose membrane (Biorad, 1704270) and Ponceau staining (Sigma-Aldrich, P7170) was performed. Blocking and antibody incubations were done in 4% non-fat milk powder in TBS-T under mild shaking. The membrane was blocked for one hour at room temperature, antibody incubations were done overnight at 4°C, and secondary antibody incubations were done for one hour at room temperature. Proteins were visualised using Immobilon Classico Western HRP substrate (Sigma-Aldrich, WBLUC0100). Primary antibody dilutions are as follow: anti-GFP 1:5000 (B-2, mouse monoclonal, Santa Cruz sc-9996), anti-β-Actin 1:4000 (13E5, Cell Signaling 4970, rabbit monoclonal). Secondary antibodies as follows: anti-mouse 1:10,000 (BioRad 1721011), anti-rabbit 1:10,000 (BioRad 1721019).

Co-immunoprecipitation experiments

Sample preparation

For each of the conditions, we prepared two biological replicates. For each replicate, a confluent T75 flask of cells was harvested using TrypLE Express (Thermo Fisher Scientific, 12605010) and pelleted for five minutes at 300 x g. Pellets were washed with PBS (Sigma-Aldrich, D8537), the suspension centrifuged for five minutes at 300 x g again and the resulting pellet flash frozen in dry ice and Methanol (VWR, 20847.307). Pellets were thawed on ice, resuspended in ice-cold RIPA buffer (50 mM Tris-HCl pH 7.6, 150 mM NaCl, 0.25 % sodium deoxycholate, 1mM EDTA, 1% NP-40) supplemented with Complete Protease inhibitor (Sigma-Aldrich, 5056489001) and incubated for five minutes on ice. The resulting protein extracts were centrifuged at maximum speed on a table-top centrifuge for 15 minutes at 4°C and the supernatant (soluble fraction) incubated with 25 μl GFP-trap magnetic beads (ChromoTek, gtma) for four hours at 4°C under constant rotation. The protein-bound beads were then washed once with RIPA buffer, twice with PBS or wash buffer (0.5 % NP-40, 0.1 mM EDTA, 20 mM Tris–HCl pH 7.4, 500 mM NaCl) and once with ddH20. Bound proteins were subsequently eluted twice for five minutes with 15 μl 1% acetic acid. Protein eluates were reduced and alkylated in the same step with 5 mM TCEP (Thermo Fisher Scientific, PG82080) and 20 mM chloroacetamide (Sigma-Aldrich, C0267) in 250 mM Tris buffer pH 8 for 30 minutes at room temperature. Sera-Mag magnetic bead mix (1:1 ratio of Sigma-Aldrich, GE45152105050250 and GE65152105050250) was added in a 25:1 protein to bead volume ratio and proteins were bound onto the beads by adding an equal volume (protein plus beads) of absolute ethanol. After a five-minute incubation, the supernatant was removed, and beads were washed three times with 80% ethanol. To perform on-bead digestion, protein-bound beads were resuspended in 50mM TEAB pH 8 (Sigma-Aldrich, T7408) containing 1 μg trypsin (Thermo Fisher Scientific, 90057) and incubated overnight gently shaking at 37°C. Afterwards, beads were allowed to settle on the magnetic rack and the supernatant was taken and kept as the first protein eluate. Subsequently, another elution was performed by adding 50 μl 2% acetonitrile (Sigma-Aldrich, 34851) in 50 mM TEAB. The second elution was combined with the first one and the eluate was dried in a SpeedVac.

LC-MS/MS

Two proteomics experiments (Figures 2E, 2F, 3A-F and S7A–S7F, bottom) were performed on an Exploris 120 mass spectrometer (Thermo Fisher Scientific) with UltiMate 3000 (Thermo Fisher Scientific) UHPLC system. Separation was performed on PepMap EASYSpray columns at 50°C. One of the LC-MS/MS analyses (Figures S7A–S7F, top, and S4C) was performed on an EASY-nLC 1200 (Thermo Fisher Scientific) coupled to an Exploris 480 mass spectrometer (Thermo Fisher Scientific). Separation of peptides was performed on 20 cm frit-less silica emitters (CoAnn Technologies, 0.75 μm inner diameter), packed in-house with reversed-phase ReproSil-Pur C18 AQ 1.9 μm resin (Dr. Maisch). Peptides were eluted in 115 min applying a segmented linear gradient of 0 % to 98 % solvent B (solvent A 0 % ACN, 0.1 % FA; solvent B 80 % ACN, 0.1 % FA) at a flow rate of 300 nL/min. Mass spectra were acquired in data-dependent acquisition mode. MS1 scans were acquired at an Orbitrap Resolution of 120,000 with a Scan Range (m/z) of 380-1500, a maximum injection time of 100 ms and a Normalised AGC Target of 300 %. For fragmentation only

precursors with charge states 2-6 were considered. Up to 20 Dependent Scans were taken. For dynamic exclusion, the exclusion duration was set to 40 sec and a mass tolerance of +/- 10 ppm. The Isolation Window was set to 1.6 m/z with no offset. A normalised collision energy of 30 was used. MS2 scans were taken at an Orbitrap Resolution of 15,000, with a fixed First Mass (m/z) = 120. Maximum injection time was 22 ms and the normalised AGC Target 50 %.

Data analysis

The raw data was analysed using MaxQuant Version 1.6.3.4[85] and searched against the Uniprot protein database[104] (Human all 2017/11), as well as our microprotein library, a list of GFP-fused bait candidates and the CRAPome database[110] for contaminants. MQ default settings were used, besides the following adjustments: Match between runs and LFQ intensity reporting were activated and MS/MS identification for both peptides in a pairwise LFQ intensity comparison was not required. The resulting proteinGroups file was analysed using the DEP package 1.22.0.[86] Common contaminants were filtered out and imputation was performed according to the "min" method. No DEP normalisation was carried out. For a protein to be considered in the analysis, we required its presence in at least three replicates. Additional requirements were a LFC enrichment of $\geq$ 1 (CENPBD2P_p_13 knock-in) or $\geq$ 1.5 (all others) over the parental control and an adjusted DEP p-value of $\leq$ 0.05 Volcano plots of the resulting data were generated using R (https://www.R-project.org/, 4.2.1) and Rstudio (http://www.rstudio.com, 2022.07.1) and further processed with Adobe Illustrator. MS/MS spectra plots shown for the CENPBD2P_p_13 knock-in were generated via MaxQuant Version 1.6.3.4 94.

Western validation

Validation of protein interactors was performed through western blot as described above. Protein sample aliquots were collected throughout the co-IP for SDS-PAGE. SDS sample buffer (recipe described above) was added to the samples and was boiled for five minutes at 95°C, followed by SDS-PAGE using a 4–20% Mini-PROTEAN® TGX™ Gel (Biorad). Proteins were transferred onto a nitrocellulose membrane (Biorad, 1704270) and Ponceau staining (Sigma-Aldrich, P7170) was performed. Blocking and antibody incubations were done in 4% TBS-T under mild shaking, using the antibodies listed in the key resources table. The membrane was blocked for one hour at room temperature, antibody incubations were done overnight at 4°C and secondary antibody incubations were done for one hour at room temperature. Proteins were visualised using Immobilon Classico Western HRP substrate (Sigma-Aldrich, WBLUC0100). Secondary antibodies used were as follows: anti-mouse 1:10,000 (BioRad 1721011), anti-rabbit 1:10,000 (BioRad 1721019).

Genomic location and conservation alignments

The genomic location (GRCh38 assembly) and phylogeny tree were output from the UCSC Genome Browser.[90] Genomic alignment was performed using the Cons 30 Primates track and alignment subsequently manually confirmed. The respective genomic sequences were then translated using the Expasy translation tool,[87] and the resulting amino acid sequence pasted into the EMBL-EBI Clustal Omega tool,[88] and from there transferred into JalView[89] for colouring.

Confocal microscopy

For live cell imaging 10,000 cells/well were seeded in an 18-well glass bottom plate (Ibidi, 81817). 24 hours after seeding, cells were washed once with PBS and subsequently incubated with prewarmed live cell imaging solution (Thermo Fisher Scientific, A14291DJ) containing 2 μM Hoechst 33342 (Life Technologies, 62249) for 20 minutes at 37°C. The staining solution was removed, cells were washed twice with PBS and then z-stack imaged in fresh pre-warmed live cell imaging solution. For fixed cell imaging, cells were seeded as above. 24 hours after seeding, wells were washed twice with PBS, fixed and stained using 4% formaldehyde (Thermo Fisher Scientific, 28908) and 2 μM Hoechst 33342 (Thermo Fisher Scientific, 62249) in PBS for fifteen minutes. Three PBS washes were performed, and cells were subsequently z-stack imaged in PBS. Imaging was performed using a Zeiss LSM800-Airy Scan laser scanning microscope with a 60x oil immersion objective. For representative images, we use a single plane from the z-stack image. Images were analysed using Fiji/ImageJ and annotated in Adobe Illustrator.

Motif searches

Motif searches were performed using the ELM tool.[111]

Gene ontology analysis

Gene ontology analysis was performed using gProfiler version 20230223,[91] MetaScape version v3.5.20230501,[92] and String database version 11.5 tools[93] with default parameters.

Venn diagrams

Venn Diagrams were made with InterActiVenn.[94]

RibORF analysis

Published Ribo-seq datasets used for this study were downloaded from the Gene Expression Omnibus repository: The Sequence Read Archive (SRA) IDs for the Ribo-seq datasets are as follows: K562 Rep 1 – SRR8449579, K562 Rep 2 – SRR8449580, K562 Rep 3 – SRR8449581, A375 – SRR10846532, and HCT116 – SRR1333393. The Ribo-seq reads were processed and trimmed of 3' adapter sequences (AGATCGGAAGAGCACACGTCT) using the FASTX-toolkit. Reads aligning to rRNA and tRNA sequences

were filtered out using STAR with parameters -outReadsUnmappedFastx. The remaining reads were aligned to the GENCODE hg38 version 39 assembly with the following settings –outFilterMismatchNmax 2 –outFilterMultimapNmax 4 –chimScoreSeparation 10 –chimScoreMin 20 –chimSegmentMin 15 –outSAMattributes All –outSAMtype BAM SortedByCoordinate. Afterwards, multimappers were filtered using samtools with setting -bq 255. Ribosome A-site metagene plots were created using RibORFv0.1's readDist.pl script and offset corrections for each ribosome protected fragment read length were determined for each dataset. Offset corrected reads in SAM format were generated using RibORFv0.1's offsetCorrect.pl script. Finally, for each dataset the ribORF.pl function was run to score each individual sORF for translation using the offset corrected SAM file and the set of 11,776 candidate sORFs in refFlat format with orf length and read coverage cutoffs set to 1. sORFs were considered translated if the RibORF P-value $\geq$ 0.7 as per the original developers' suggestions.

Rescue assays

The generated rescue cell lines and corresponding A375 Cas9/A375 Cas9 GFP cells were seeded at a 1:1 (80,000 cells/well total) ratio into a 24-well plate. One day after seeding, cells were transfected with synthetic sgRNAs (Synthego) utilizing the Lipofectamine™ RNAiMAX transfection reagent (Thermo Fisher Scientific, 13778075) according to the manufacturer's instructions. One day after transfection, cells were transferred into 12-well plates and grown for another seven days. Cells were split on days 1, 4, 6 and harvested on day 8 post-transfection. A small fraction of cells was separated for flow cytometry analysis on each splitting day, and all cells were harvested on day 8 post-transfection. For all flow cytometry analysis, cells were harvested, pelleted for five minutes at 300 x g, cell pellets resuspended in 5% FBS (Sigma-Aldrich, F7524) in PBS (Sigma-Aldrich, D8537) and then analysed on a Navios flow cytometer (Beckman Coulter). Flow cytometry data analysis was performed with FlowJo™ v10.8 Software (BD Life Sciences), assessing single cell GFP+ and GFP- fractions.

RNA-sequencing

Sample preparation

80,000 A375 Cas9 cells per well were seeded in a 24- well plate and transfected with the respective synthetic sgRNA (Synthego) in triplicate wells one day after seeding using Lipofectamine™ RNAiMAX (Thermo Fisher Scientific, 13778075) according to the manufacturer's protocol. Cells were transferred into a 6-well plate one day post-transfection and expanded until three days post-transfection. One million cells were then harvested by trypsinization (TrypLE Express, Thermo Fisher Scientific, 12605010) and pelleted by centrifugation at 500 x g for three minutes. Pellets were washed once with PBS (Sigma-Aldrich, D8537), flash frozen and stored at -80°C until use. The pellet was resuspended in 350µl RLT buffer (Qiagen, 74106) supplemented with 1% β-mercaptoethanol. The lysate was homogenised using the Qiashredder column (Qiagen, 79654) and RNA extracted utilising the RNeasy Plus Mini Kit (Qiagen, 74136) according to manufacturer's instructions. RNA concentration was measured with the Qubit RNA HS assay kit (Life Technologies, Q32852) according to manufacturer's instructions and methanol flash frozen. RNA-sequencing service was provided through BGI services (www.bgi.com), performing strand-specific RNA-seq with poly(A) selection (DNBseq Eukaryotic Transcriptome *De novo* Sequencing).

Analysis

RNA-seq FASTQ files were processed using the nf-core RNA-seq pipeline version 3.5 (https://nf-co.re/rnaseq/3.5) with star_rsem parameters,[112,113] hg38 as reference and RefSeq as gene annotation. The resulting counts were processed using the DeSeq2 software[97] with default parameters to calculate corresponding LFC and significance values. TPM values displayed as calculated by the pipeline. For a gene to qualify as a hit, we required an average LFC of $\geq$ 0.5 or $\leq$ –0.5 between microprotein sgRNA and TRAC sgRNA treated A375 Cas9 cells, a DeSeq2 adjusted p-value of $\leq$ 0.05 and the average gene expression variability of the two TRAC conditions (microprotein-GFP vs. parental cells) to not exceed 10%. Subsequently we scored rescued hits, by testing for genes that additionally displayed a LFC difference of $\geq$ 0.5 between the microprotein knock-out conditions in microprotein-GFP and parental cell lines.

Immunofluorescence microscopy

10,000 cells/well were seeded in a 96-well glass bottom plate (VWR, GREI655891_16). 24 hours after seeding, the cells were washed once with PBS (Sigma-Aldrich, D8537), fixed with 4% formaldehyde (Thermo Fisher Scientific, 28908) for fifteen minutes at room temperature, and subsequently washed twice time with PBS. Cells were permeabilized using 0.1% Triton X-100 (Sigma-Aldrich, T9284) in PBS (Sigma-Aldrich, D8537) for 15 minutes and washed twice with TBS-T. The samples were blocked for one hour at room temperature with 0.1% BSA in 0.05% Triton X-100 (Sigma-Aldrich, T9284-100ML) in TBS-T. For staining, primary antibodies were diluted in the blocking solution, and incubated with the sample for one hour at room temperature or overnight at 4°C. Subsequently, three washes with TBS-T were performed, and the samples were incubated with secondary antibody diluted in blocking solution, for one hour at room temperature, followed by three washes with TBS-T. To stain nuclei, the cells were then incubated with 1µg/ml DAPI (Sigma-Aldrich, D9542) in PBS for five minutes and washed twice with PBS. The sample was imaged in PBS on a Nikon Eclipse Ti2 inverted widefield microscope. Primary and secondary antibody dilutions as follows: anti-HA 1:500 (F-7, mouse monoclonal, Santa-Cruz sc-7392), anti-mouse Alexa 647 1:1000 (Thermo Fisher Scientific, A-21236).

Synthetic sgRNAs, primers and plasmids

All Synthetic sgRNAs were ordered from Synthego (Table S4). Primers used in this study are listed in Table S8. Plasmids used in this study are described in Table S9.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data from a minimum of three experiments are presented as the mean ± SD. A two-tailed unpaired t-test was used for comparisons between two groups. A p-value of less than 0.05 was considered statistically significant and denoted with *.