Science
AAAS

Supplementary Materials for

Pervasive functional translation of non-canonical open reading frames

Jin Chen, Andreas-David Brunner, J. Zachery Cogan, James K. Nuñez, Alexander P. Fields, Britt Adamson, Daniel N. Itzhak, Jason Y. Li, Matthias Mann, Manuel D. Leonetti, Jonathan S. Weissman

Correspondence to: jonathan.weissman@ucsf.edu

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

Materials and Methods

Ribosome profiling

WTC iPSCs (Coriell Biorepository #GM25256) were maintained under feeder-free conditions on growth factor-reduced Matrigel (Corning) in mTeSR medium (STEMCELL Technologies). Accutase (STEMCELL Technologies) was used to enzymatically dissociate iPSCs into single cells to passage by incubating the cells at 37°C for 5 minutes. To promote cell survival during enzymatic passaging, cells were passaged with 10 μM p160-Rho-associated coiled-coil kinase (ROCK) inhibitor Y-27632 (Selleckchem). iPSCs were differentiated into iPS-derived cardiomyocytes using the WNT modulation-differentiation method following previously published protocols(*46, 47*).

Harringtonine-treated cells were treated with 2 μg/mL harringtonine in DMSO at 37ºC for 2 minutes. Cells were not pre-treated with the elongation inhibitor cycloheximide before harvest. Ribosome-protected footprints from harringtonine-treated and no-drug samples were prepared for sequencing as described in a recently updated protocol of ribosome profiling(*48*). Briefly, cells were rapidly harvested and lysed. Clarified cell lysates were treated with RNase I (Invitrogen) to digest RNA not protected by ribosomes. 80S ribosomes were isolated by centrifuging lysates through a 34% sucrose cushion at 100,000×g for 1 hour at 4°C. RNA was then purified from the ribosome pellet using the Direct-zol RNA kit (Zymo Research). The RNA was then resolved by electrophoresis through a denaturing gel, and the fragments corresponding to 28 to 34 bp were extracted from the gel. There were recent reports demonstrating that much smaller ribosome footprints spanning 17 to 20 bp are in fact active ribosomes(*49*). At the time of doing the ribosome profiling experiments, we reasoned that since the smaller footprints were not well characterized, and since the 28 – 34 bp footprints are still the predominant population of footprints, we still chose to go with the canonical footprint sizes. Though, recent protocols have also included the smaller 17 – 20 bp footprints(*48*).

The 3' ends of the ribosome footprint RNA fragments were then treated with T4 polynucleotide kinase (NEB) to allow ligation of a pre-adenylated DNA linker with T4 Rnl2(tr) K227Q (NEB). The DNA linker incorporates sample barcodes to enable library multiplexing, as well as unique molecular identifiers (UMIs) to enable removal of duplicated sequences. To separate ligated RNA fragments from unligated DNA linkers, 5'-deadenylase (Epicentre) was used to deadenylate the pre-adenylated linkers, which were then degraded by the 5'-3' ssDNA exonuclease RecJ (NEB). After rRNA reduction using the Ribo-Zero Gold rRNA removal kit (Illumina), The RNA-DNA hybrid was used as a template for reverse transcription, followed by circularization with CircLigase (Epicentre). Finally, PCR of the cDNA circles attached suitable adapters and indices for Illumina Sequencing. The library was sequenced on an Illumina HiSeq 4000 sequencer with a single-end 50 base pair run. The corresponding RNA-seq samples were prepared as described previously(*27*).

Ribosome profiling analysis

The genome assembly used throughout this manuscript is hg19/GRCh37. Custom transcriptome annotations was merged from Gencode Gene V24lift37 with a custom lncRNA annotation assembled as described in a previous publication(*46*). Briefly, lncRNA annotations were retrieved from Ensembl build 75 (using the biotypes lincRNA, antisense, 3 prime overlapping ncRNA, processed transcript, sense intronic, sense overlapping), the Broad human lincRNA catalog(*50*), and the MiTranscriptome(*51*). Annotations were merged using the cuffmerge command in Cufflinks v2.2.1.

For processing of ribosome profiling data, linker sequences were removed from sequencing reads and samples were de-multiplexed using FASTX-clipper and FASTX-barcode splitter (FASTX-Toolkit). Unique molecular identifiers and sample barcodes were then removed from reads using a custom Python script. Bowtie v1.1.2 was used to filter out reads aligning to rRNAs and contaminants, and all surviving reads were aligned to the custom transcriptome described above with Tophat v2.1.1 using the --b2-very-sensitive --transcriptome-only --no-novel-juncs --max-multihits=64 flags. These alignments were assigned a specific P-site nucleotide using a 12-nt offset from the 3' end of reads.

The ORF-RATER pipeline (https://github.com/alexfields/ORF-RATER) was run as previously described(*27*) (see Supplemental Detailed Protocol in the referenced manuscript for explanation), starting with the BAM files from the alignments described above. Note that ORF-RATER, compared with other algorithms, is tuned to indicate the highest-confidence sites of translation, at the expense of an increase false negative rate, so it is possible that a translated ORF may be assigned a low score. PhyloCSF analysis was also performed as previously described(*27, 33*). For each non-canonical CDS, only those with at least ten codons for which no nucleotides overlapped annotated coding regions were analyzed, and the non-overlapped sequences were identified in a set of ten mammals spanning the Euarchontoglires: human, chimpanzee, rhesus macaque, bushbaby, mouse, rat, guinea pig, squirrel, rabbit, and pika. Aligned sequences were retrieved using the 100-way multispecies alignment available from the UCSC genome browser. PhyloCSF was ran on the aligned sequences if they could be identified in at least five of the species(*33*). Additional analysis and plotting were performed in Python 2.7 using a combination of plastid(*52*), Biopython, Numpy (v1.12.1), Pandas (v0.17.1), and Scipy (v0.17.0). The human fibroblast ribosome profiling dataset was previously published(*28*). The output of ORF-RATER is included in Table S1.

Sample preparation for MS analysis

For the deep proteome analysis, a 15 cm, 80% confluent, dish of iPSCs was detached from the plate with 0.5 mM EDTA solution at room temperature (RT) for 5 min. Cells were then washed extensively in Phosphate Buffer Solution (PBS), and then resuspended in 1 mL PBS, transferred into a 2 mL Eppendorf tube, and spun down at 500g RT for 5 min. Supernatant was removed and the cell pellet was flash-frozen in liquid nitrogen.

Cells were then resuspended in 1 mL SDC lysis buffer(*53*) and boiled for 20 min at 95°C, 1500 rpm to denature and reduce and alkylate cysteins, followed by sonication in a Branson sonicator (3x45sec). The suspension was boiled again for 20 min at 95°C, 1500 rpm, cooled down to room temperature and diluted 1:1 with ddH2O. Protein concentration was estimated by Nanodrop measurement and 500 µg were further processed for overnight digestion by adding LysC and Trypsin in a 1:50 ratio (µg of enzyme to µg of protein) at 37°C and 1500 rpm. Next day, samples were sonicated in a Branson sonicator (3x45sec) and further digested for 4 hours with LysC and Trypsin (1:50 ratio) at 37°C and 1500 rpm. Peptides were acidified by adding 1% TFA 99% Isopropanol in a 1:1 ratio and vortexed, followed by centrifugation at 22,000g at RT to pellet residual particles. The supernatant was transferred into a fresh tube and subjected to stage-tip clean-up via SDB-RPS. 20 µg of peptides were loaded on two 14-gauge stage-tip plugs. Peptides were washed two times with 200 µL 1% TFA 99% Isopropanol followed 200 µL 1%

TFA 99% Isopropanol in an in-house-made Stage-tip centrifuge at 2,000 g, followed by elution with 100 µL of 1% Ammonia, 80% ACN, 19% ddH2O into PCR tubes and dried at 60°C in a SpeedVac centrifuge (Eppendorf, Concentrator plus). Peptides were resuspended in 0.1% TFA, 2% ACN, 97.9% ddH2O. 100 µg of peptides were subjected to fractionation into 24 concatenated fractions by high-pH reversed-phase fractionation with a "loss-less" nano-fractionator(*54*).

Liquid chromatography MS-analysis

LC-MS/MS analysis was performed on a quadrupole Orbitrap mass spectrometer (Q Exactive HFX, Thermo Fisher Scientific) coupled to an EASYnLC 1200 system via nano-electrospray ion source. 500 ng of peptides were loaded on a 50 cm in-house packed HPLC-column (75µm inner diameter packed with 1.9 µm ReproSil-Pur C18-AQ silica beads, Dr. Maisch GmbH, Germany). Sample analytes were separated using a linear 120 min gradient from 5-30% B in 95 min followed by an increase to 60% in 5 min, and by a 5 min wash at 95% B at 300 nl/min (Buffer A: 0.1% Formic Acid, 99.9% ddH2O; Buffer B: 0.1% Formic Acid, 80% ACN, 19.9% ddH2O).

The column temperature was kept at 60°C by an in-house manufactured oven. Mass spectrometry analysis was performed in a data dependent scan mode. For full proteome measurements, MS1 spectra were acquired at a 60,000 resolution and a m/z range of 315-1715 with an automatic gain control (AGC) target of 3E6 ions and a maximum injection time of 60 ms. The top 12 most intense ions with a charge of two to eight from each MS1 scan were isolated with a width of 1.4 m/z, followed by higher-energy collisional dissociation (HCD) with a normalized collision energy of 27% and a scan range of 200 – 2000 m/z. MS/MS spectra were acquired at 15,000 resolution with an AGC target of 1E5, a minimum AGC target of 2.5E3, and a maximum injection time of 120 ms. Dynamic exclusion of precursors was set to 40 sec.

MS Proteomics and HLA Peptidomics data analysis

Raw-files were searched against the human Uniprot databases (UP000005640_9606.fa, UP000005640_9606_additional.fa) using the MaxQuant version 1.6.5.0 either with a 1% FDR both on PSM and protein level, or 1% FDR on PSM and 100% FDR on protein level. Peptides with a minimum length of seven amino acids were considered for the search including N-terminal acetylation and methionine oxidation as variable modifications and cysteine carbamidomethylation as fixed modification. Enzyme specificity was set to trypsin cleaving C-terminal to arginine and lysine. A maximum of two missed cleavages were allowed. Maximum precursor and fragment ion mass tolerance was set to 4.5 and 20 ppm. For the deep proteome and HLA peptidome analysis, raw-files were searched additionally against a custom ribosome sequencing fasta file (Table S2). In Figure 1F, the Andromeda score is defined as a measure of how well an acquired spectrum matches with the theoretical fragment masses.

HLA peptidome analysis was performed on a previously published HLA class I dataset, describing the HLA class I peptidomes of six allotype-resolved cell lines (Fibroblasts: HLA-A*03:01, A*23:01, B*08:01, B*15:18, Cw*07:02, Cw*07:04; HCC1143: HLA-A*31:01, B*35:08, B*37:01, Cw*04:01, Cw*06:02; HCC1937: HLA-A*23:01, A*24:02, B*07:02, B*40:01, Cw*03:04, Cw*07:02; SupB15: HLA-A*03, A*11, B*51, B*52, Cw*12:04, Cw*14:02; HCT116: HLA-A*01:01, A*02:01, B*45:01, B*18:01, Cw*05:01, Cw*07:01; JY:

HLA-A*02:01, B*07:02, Cw*07:02)(*30*). Proteomic analysis of HLA-I complexes rely on immunoprecipitation of HLA complexes with bound peptides, serving as an enrichment step to enhance novel peptide detection. HLA peptidome data was only searched with a 1% PSM FDR since we are only interested in HLA-I peptide identifications. Protease specificity was set to unspecific, possible peptide identifications were restricted from 8 to 15 amino acids, maximum peptide mass was set to 1500 Da, and modification was set to without fixed modifications. Bioinformatics analysis was performed with the Perseus software (version 1.6.5.0) and GraphPad Prism (version 7.04). Proteins identified only by site modification or in the decoy reverse database and potential contaminant were excluded from the downstream analysis.

HLA binding motif analysis was performed with the GibbsCluster-2.0 server and default settings only taking into account HLA class I 9-mers, which served as the basis for the known allotype assignment(*55*). MHC class I 9-mer peptide binding prediction to distinct allotypes revealed by the GibbsCluster analysis and compared to the known allotype background HLA-I clustering results of each particular cell line, was performed with the NetMHC 4.0 server and default settings(*56*). Strong binders are reported with a ≤50 nM binding affinity and weak binders are reported with a binding affinity of >50 nM and ≤500 nM.

Explanation of MS proteomic data analysis

The assignment of tryptic peptides, which we identify via mass spectrometry, occurs in two ways. Either these peptides are assigned uniquely to a single protein, or to several proteins depending on the shared amino acid sequence proportion. Keeping this in mind, we report protein groups, which can contain several protein identifications, but are not distinguishable based on the identified peptide for identification. Here, we filtered the MaxQuant protein groups output table only for protein group identifications, which contain one protein and only peptides mapped to a single protein (unique peptides) to be as stringent as possible for reporting non-canonical CDS identifications. We noticed that the overall predicted non-canonical CDS protein length is small, which decreases the likelihood of yielding tryptic peptides with a high score that subsequently converts into protein identification. Also, we reasoned that the abundance of these non-canonical CDS proteins might be low, which again contributes negatively to the identification score. Therefore, we re-analyzed the deep proteome experiment with a 1% PSM and 100% Protein group FDR (Q-Value) and compared the identifications to the 1% PSM and 1% Protein group FDR output. This identified 11 non-canonical CDS peptides in the former and 45 in the latter case.

CRISPR library design and cloning

From the non-canonical CDSs identified by ribosome profiling and ORF-RATER, the ORFs satisfying the following categories were selected: ORF-RATER score greater than 0.8, length at least 10 amino acids, and the ORF type was either new, upstream, downstream, Giso, new_iso, start_overlap, or extension. These are the ORF types that can be specifically targeted without affecting annotated coding regions, so ORF types such as truncations and internal out of frame ORFs are excluded. All sequences 150 bp upstream of the ORF start to the end of the ORF containing 19 bp followed by an NGG PAM were extracted as potential sgRNAs. All sequences were prepended with a 5' G to enable robust transcription from the U6 promoter, whether or not this base was present in the genomic sequence. All potential sgRNAs were scored for predicted on-target activity using the SSC score(*57*) as well as the Doench v2 score(*58*). In addition, the sgRNAs were scored for off-target sites using weighted Bowtie, as previously described(*59*), and using GuideScan(*32*). Briefly, sgRNAs were scored by uniqueness in the genome, as determined by an empirically derived and experimentally verified scoring metric: PAM G1 = 40, PAM G2 = 19, PAM N = 0, the next 7 bases from the PAM = 28, the next 5 bases = 19, and the last 7 bases = 10. A mismatch score was then calculated by the sum of the mismatches with the scoring metric. This mismatch score was implemented using the Phred score threshold feature of Bowtie using the --nomaqround, -n 3, -l 15, -a, and --best flags. For the most stringent threshold, sgRNAs were required to have no more than 1 alignment (the sgRNA target site itself) in the genome with a mismatch score of 39. For each ORF, up to 10 sgRNAs targeting within the ORF and up to 5 sgRNAs targeting the upstream genomic region as controls were selected without predicted off-targets and with the highest guide SSC scores. We have found the SSC score to be a slightly better measure of on-target activity than the Doench v2 score. For ORFs that cannot be targeted by 10 sgRNAs at the most stringent threshold, the threshold was relaxed, in descending order: 1 alignment under 30, 1 alignment under 20, 1 alignment under 11, 1 alignment under 1, 2 alignments under 39, and 3 alignments under 39. Control non-targeting sgRNAs were extracted from a previously tested list of control sgRNAs(*60*). The sgRNA library composition is included in Table S3.

Oligonucleotide pools were designed with flanking PCR and restriction sites (BstXI and BlpI), synthesized by Agilent Technologies, and cloned into the sgRNA expression vector pCRISPRia-v2 (Addgene #84832), as described previously(*59*). The expression vector contains a U6 promoter driving the sgRNA expression, as well as an EF1α promoter driving puromycin-T2A-BFP.

CRISPR screen and analysis

iPSC expressing Cas9 (WTC CRISPRn Gen1C) and K562 expressing Cas9 were obtained from previous publications(*46, 52*). WTC CRISPRn Gen1C iPS cell line was a gift from Bruce R. Conklin (Gladstone, UCSF). iPSC-Cas9 cells were cultured and passaged as described above. K562-Cas9 cells were grown in RPMI 1640 (GIBCO) with 25 mM HEPES, 2 mM L-glutamine, 2 g/L NaHCO3 and supplemented with 10% (v/v) fetal bovine serum (FBS), 100 units/mL penicillin, 100 mg/mL streptomycin, 2 mM L-glutamine, and passaged daily between 0.5×10^6 cells/mL and 1×10^6 cells/mL. The sgRNA library described above was packaged into a lentivirus library with TransIT-LT1 (Mirus) transfection in HEK293T cells. K562 and iPSC cell lines expressing Cas9 were infected in duplicate with the lentivirus library at an initial infection rate of 30% (1000x coverage of the library). Cells were cultured for two days following infection, and then treated for two days with 0.75 μg/mL puromycin (GoldBio) for K562, and six days with 3 μg/mL puromycin for iPSCs. The cells were allowed to recover for two days, and then cultured at a minimum coverage of 1000x for 10 doublings starting from this "T0". iPSCs were then treated daily with 2 μM doxycycline (Sigma) starting from this T0, and were split on alternate days. K562 cells were passaged daily. At the endpoint, cells were harvested, and sgRNA-encoding regions were enriched and then amplified by PCR, and then sequenced on the Illumina HiSeq 4000 with a single-end 50 bp run, as previously described(*47, 52*).

Sequencing counts from CRISPR screens were processed using the Python-based ScreenProcessing pipeline (https://github.com/mhorlbeck/ScreenProcessing), as previously described(*47, 52*). sgRNA phenotype scores (γ) were calculated, as defined in Fig. 2. ORF phenotypes were scored based on the average phenotype of the 3 strongest sgRNAs (by absolute value) targeting it. Mann-Whitney test p-values were calculated by comparing all sgRNAs targeting a given ORF to the full set of negative control sgRNAs. In order to call hit ORFs from screens, a "screen score" was defined as | γ z-score from negative control gene distribution | × –log10 p-value. Two criteria were used to determine the threshold for the screen score: that the false-discovery rate (FDR) be less than 0.05, and the weakest phenotype score be at least -0.02 (because the phenotype due to Cas9 cutting was determined to be on average -0.015). It is important to note that due to the small sizes of the ORFs targeted, unlike targeting canonical proteins, there are cases in which not all the sgRNAs targeting a single ORF have high on-target scores. Thus, this is the reasoning behind calculating the average phenotype from the top 3 sgRNAs. Furthermore, this is the reason why in Figure 2C, even though the difference is significant between sgRNAs targeting within the ORF and immediately upstream, there are still sgRNAs with low phenotype scores targeting within the ORF. The results from the screens are summarized in Table S4 and S5.

Furthermore, in this current manuscript, we focused on uORFs and lncRNA CDSs. On the other hand, extensions and start overlaps are also very interesting categories of ORFs worth following up on. However, due to the possible short distance from the non-canonical start site to the canonical start codon, it is potentially difficult to precisely disrupt only the extended region without affecting the canonical CDS. Thus, the results should be interpreted with care.

All additional CRISPR screen data analyses and plotting were performed in Python 2.7 using a combination of Numpy (v1.12.1), Pandas (v0.17.1), Scipy (v0.17.0), and scikit-learn(v0.19.1). Gene ontology analysis was conducted using DAVID 6.8. RNA min-free energy (MFE) is calculated using the ViennaRNA package. Additional datasets used in the analysis include previously published K562 CRISPRn screen(*26*), K562 CRISPRi screen(*47*) and iPSC CRISPRi lncRNA screen(*46*). The m6A dataset is from (*61*). Comparison of the different screen results is summarized in Table S6. Kozak context score was calculated by a scoring metric: 3×(G at position -6) + (C at position -5) + (C at position -4) + 3×(A or G a position -3) + (C at position -2) + (C at position -1) + 3×(G at position 3), where position 0 is the first base of the start codon.

Perturb-Seq screen and analysis

A smaller Perturb-Seq library was designed to screen 83 uORFs and 80 lncRNA CDSs, chosen manually based on conservation, phenotype from the screen, and ORF-RATER score. 2 of the most active sgRNAs from the CRISPR screen were chosen for each ORF, as well as 6 non-targeting control sgRNAs and 6 control sgRNAs targeting intergenic regions of the genome. The Perturb-Seq library is included in Table S7.

Similar to above, oligonucleotide pools were designed with flanking PCR and restriction sites, synthesized by Twist Bioscience, and cloned into a modified CROP-seq vector backbone(*35*). The CROP-seq backbone was a gift from Christoph Bock (Addgene plasmid #86708) and modified to match the vector used in the CRISPR screen, pCRISPRia-v2, by cloning in the mouse U6 promoter, BstXI and BlpI restriction sites, as well as the optimized sgRNA constant region. The puromycin resistance cassette driven by the EF1α promoter was changed to a BFP.

The Perturb-Seq sgRNA library was packed into a lentivirus library with TransIT-LT1 (Mirus) transfection in HEK293T cells. iPSC cell lines expressing Cas9 were infected with the

lentivirus library at an initial infection rate of 10% (1000x coverage of the library). iPSCs were then treated daily with 2 μM doxycycline (Sigma) and cultured for two days following infection, and then FACS sorted (BD FACS Aria2) for the BFP+ population. Seven days post-infection, cells were prepared for single-cell RNA-seq using the 10X Chromium Controller and Chromium Single Cell 3' Library & Gel Bead Kit v2 (10x Genomics). From the final library, the sgRNA sequences are specifically amplified with PCR, as described before(*34*), with the following primers:
5'-AATGATACGGCGACCACCGAGATCTACAC-3'
5'-
CAAGCAGAAGACGGCATACGAGAT**CAGCCTCG**GTCTCGTGGGCTCGGAGATGTGTA
TAAGAGACAGGTGTTTGAGACTATAAGTATCCCTTGGAGAACCACCTTGTTG-3'
The sequence is bold is the i7 index. PCR cycling was performed according to the following protocol: (1) 95°C for 3 min, (2) 98°C for 15 s, then 70°C for 10 s (15 cycles) (3) 72 C for 1 min. The resulting sgRNA barcode library was purified via a 0.8X SPRI selection, and then further size selected using BluePippin (Sage Science). sgRNA barcode libraries were sequenced as spike-ins alongside the parent RNA-seq libraries (5% spike-in) on a NovaSeq 6000, following manufacturer's recommendations. An auxiliary, follow-up Perturb-Seq experiment was performed similarly, but with the Chromium Single Cell 3' Reagent Kits (v3 Chemistry) with Feature Barcoding technology for CRISPR Screening (10X Genomics), following manufacturer's instructions.

Cell Ranger (version 2.1.1, 10X Genomics) with default parameters was used to align reads and generate digital expression matrices from single-cell sequencing data. Cell Ranger (version 3.0.0, 10X Genomics) was used to align and analyze the follow-up Perturb-Seq screen. Cell sgRNA identity assignments were processed using custom Python scripts described previously(*34*). Cells with only one assigned sgRNA were retained for further analysis. Custom Python scripts described previously(*34*) were used to analyze the digital expression matrices including normalization, quality control, and filtering. To normalize for differences in sequencing capture and coverage across emulsion droplets, we rescaled all cells to have the median number of total UMIs. sgRNAs with lower than 10 cells were removed from analysis. Expression of each gene was then z-normalized with respect to the mean and standard deviation of that gene in the control population, and in our case the population with sgRNA "intergenic_chr10_120424177_+". To analyze the differences in gene expression between populations of perturbed cells, a random forest classifier was used, motivated by the idea that a gene is likely important for a given perturbation of its expression level can be used to accurately predict that perturbation's identity. See the Methods section in reference (*34*) for more detail. Because sgRNAs targeting the same ORF were found to produce similar expressions, differentially expressed genes were determined on the ORF level using a random forest classifier for each ORF compared against the intergenic control sgRNAs. Cell cycle analysis was performed as described previously(*34*). Gene ontology analysis of the gene expression was performed using Gene Set Enrichment Analysis (GSEA) with GSEAPY in Python 2.7. Gene expression analysis was performed with genes chosen either from Gene Ontology gene sets (Molecular Signatures Databse), STRING interactions (http://string-db.org) or GeneMANIA(*62*). The results from the Perturb-Seq analysis are included in Table S7.

Competition assays and validation experiments

sgRNAs for individual validation were cloned by annealing complementary synthetic oligonucleotide pairs (Integrated DNA Technologies) with flanking BstXI and BlpI restriction sites and ligating the resulting double-stranded segment into BstXI/BlpI-digested pCRISPRia-v2 (same as the one used in the CRISPR library, marked with a puromycin resistance cassette and BFP). The resulting sgRNA expression vectors were individually packaged into lentivirus with TransIT-LT1 (Mirus) transfection in HEK293T cells. Internally controlled growth assays to evaluate sgRNA phenotypes were performed by transducing Cas9-expressing K562 cells with sgRNA expression constructs at MOI < 1 (15 – 30% infected cells), and measuring the fraction of sgRNA-expressing cells as BFP-positive cells by flow cytometry on an LSR-II (BD Biosciences) over the course of 7-10 days. All sgRNA sequences used, as well as the backbone sequence, are included in Table S8.

To characterize the indels resulting from Cas9 cutting, Cas9-expressing K562 cells were infected as described above. BFP-positive cells were sorted by FACS (SH800S, Sony), and 5 days after infection, cells were harvested. Genomic DNA was isolated from frozen K562 cell pellets with QuickExtract (Lucigen). The target site was amplified by nested-PCRs to make sequencing libraries, as described in (*63*), and sequenced on the MiSeq (Illumina). Sequencing data was processed with CRISPResso (https://github.com/lucapinello/CRISPResso), and then further analyzed using Python 2.7.

For the rescue experiments, sgRNAs were cloned by annealing complementary synthetic oligonucleotide pairs (Integrated DNA Technologies) with flanking BstXI and BlpI restriction sites and ligating the resulting double-stranded segment into BstXI/BlpI-digested pCRISPRia-v2, same as the competition assay above. The corresponding peptide for the rescue of each sgRNA was cloned into the same vector by digesting with NsiI and EcoRI. DNA fragments corresponding to the SFFV promoter, the native context of the peptide CDS (the entire transcript sequence 5' of the CDS starting from the transcript start site to the stop codon of the CDS), and IRES-mCherry were either obtained by PCR or ordered as a gBlock (Integrated DNA Technologies). Sense mutations were introduced to the CDS sequence to prevent sgRNA targeting. The DNA fragments and the NsiI/EcoRI digested sgRNA vector were then assembled by Gibson Assembly (NEBuilder HiFi DNA assembly kit, New England Biolabs). The resulting plasmid expresses both the sgRNA and the peptide in the same vector. For the Δstart codon plasmids, the initiating start codon was deleted. For the knockout controls with no rescue, the same plasmid was used except the peptide sequence was replaced by a HA tag. The resulting vectors were individually packaged into lentivirus with TransIT-LT1 (Mirus) transfection in HEK293T cells. Internally controlled competition assays were performed as above with Cas9-expressing K562s. The mCherry+ population was measured by flow cytometry on a LSR-II (BD Biosciences) over the course of 7-14 days. See Table S8.

Western blot analysis

Transcripts containing 3xFLAG tagged uORF and mCherry tagged main CDS was ordered as gBlocks (Integrated DNA Technologies) and cloned into the pHR-SFFV-HA-IRES-mCherry vector (from a previous publication(*64*), expressing a HA tag sequence) using Gibson Assembly (NEBuilder HiFi DNA assembly kit, New England Biolabs), between the SFFV promoter and WPRE sequence (taking out the HA-IRES-mCherry portions). Transcripts containing 3xFLAG tagged uORF and 3xFLAG tagged main CDS are cloned similarly. Two days after transfection (TransIT-LT1, Mirus) into HEK293T cells, the cells were collected and flash frozen for Western blot analysis. The cells were lysed in RIPA buffer (Thermo Fisher) containing 0.5 mM EDTA and 1X Halt Protease inhibitor cocktail (Thermo Fisher). Lysate was centrifuged, supernatant was isolated and protein content was assessed using BCA assay (Pierce). 20 ug of protein samples were loaded onto a NuPage 4-12% Bis-Tris gel and ran with NuPage MES SDS running buffer (Thermo Fisher). For Western blots, primary antibodies were diluted as follows: anti-FLAG (Sigma, F1804, 1/2000), anti-mCherry (Abcam, ab125096, 1/1000), and anti-GAPDH (Abcam, ab8245, 1/2000). Quantification of Western blots are performed using Image Studio Lite (LI-COR Biosciences).

Immunoprecipitations experiments were conducted using RFP-Trap (ChromoTek) following manufacturer's protocol. Immunoprecipitated proteins were eluted from beads by adding NuPage LDS sample buffer with reducing agent, and boiling at 95ºC for 10 minutes.

Ectopic expression of tagged peptide for microscopy and co-immunoprecipitation

To express the tagged peptide for microscopy and co-immunoprecipitation, the native context of the peptide CDS (the entire transcript sequence 5' of the ORF to the stop codon of the ORF), with a short linker (GGTGGCGGC) and GFP11(*36*) or mNeonGreen11(*37*) tag inserted before the stop codon, was ordered as a gBlock (Integrated DNA Technologies). For difficult to synthesize sequences, the exon sequence was amplified by PCR from cDNA generated with iPS cells. The DNA fragments were then assembled into the pHR-SFFV-HA-IRES-mCherry vector (from a previous publication(*64*), expressing a HA tag sequence) using Gibson Assembly (NEBuilder HiFi DNA assembly kit, New England Biolabs), between the SFFV promoter and WPRE sequence (taking out the HA-IRES-mCherry portions). See Table S8 for backbone sequence as well as insert sequences. The expression vectors were packaged into lentivirus and infected into HEK293T cells expressing GFP1-10 or mNeonGreen1-10. GFP+ populations were FACS sorted (SH800S, Sony) 5 days after infection. We took care during the sort to only sort cells with fluorescence intensities ~10 to 50 fold above background, consistent with the expression of many well-expressed endogenous genes, to prevent artifacts from massive overexpression. The cells were then expanded and analyzed by microscopy or co-immunoprecipitation (as described below). HEK293T cells were grown in Dulbecco's modified eagle medium (DMEM, GIBCO) with 25 mM D-glucose, 3.7 g/L NaHCO3, 4 mM L-glutamine and supplemented with 10% (v/v) FBS, 100 units/mL penicillin, and 100 mg/mL streptomycin. To estimate magnitude of overexpression, total RNA was isolated from frozen cell samples using the Direct-zol RNA MiniPrep kit (Zymo Research). Reverse-transcription was carried using SuperScript VILO Master Mix (Thermo Fisher Scientific). Quantitative PCR (qPCR) was performed with Kappa Sybr Fast qPCR 2x Mix (Roche), according to the manufacturer's instructions on a LightCycler 480 Instrument (Roche). Experiments were performed in technical triplicates, and the qPCR primers were designed using the Roche Universal ProbeLibrary Assay Design Center.

Endogenous tagging

Cas9/sgRNA ribonucleoprotein (RNP) complexes were prepared following previously published protocols(*36*). Cas9 protein, synthetic crRNAs and tracrRNAs were purchased from Integrated DNA Technologies. HEK293T cells expressing GFP1-10(*36*) or mNeonGreen1-10(*37*) were treated with 200 ng/mL nocodazole (Sigma) for 15 hours before electroporation to increase gene editing efficiency. RNP complexes were assembled with 50 pmol Cas9 protein, 50 pmol crRNA, 130 pmol tracrRNA just prior to electroporation, and combined with 120 pmol 200 base-pair single-stranded oligonucleotide HDR (homology-directed repair) template (Integrated DNA Technologies). See Table S8 for sgRNA and HDR oligo sequences.

Electroporation was carried out in Amaxa 96-well shuttle Nucleofector device (Lonza) using SF-cell line reagents (Lonza) following the manufacturer's instructions. Cells were extensively washed with PBS and resuspended to 100 cells/µL in SF solution immediately prior to electroporation. For each sample, 20 µL of cells were added to 10 µL RNP/template mixture. Cells were immediately electroporated using CM130 program and transferred to 1 mL pre-warmed culture media in a 24-well plate. Electroporated cells are cultured for > 5 days prior to analysis with microscopy or the LSRII (BD). The top 1% of the GFP+ population was then FACS sorted (SH800S, Sony) to generate a population that is enriched in the correct insert. Note that the final population is still polyclonal, so the cells might have a mixture of different repair and HDR outcomes. For clones with well-defined alleles, the cells were electroporated as before, but are then single-cell FACS sorted (SH800S, Sony) into 96-well plates. After the cells were grown up and expanded, genomic DNA was isolated with QuickExtract (Lucigen). The target site was amplified by nested-PCRs to make sequencing libraries, as described in (*63*), and sequenced on the MiSeq (Illumina). Sequencing data was processed with CRISPResso (https://github.com/lucapinello/CRISPResso), and then further analyzed using Python 2.7.

Immunoprecipitations experiments of the tagged clonal lines were conducted using mNeonGreen-Trap (ChromoTek) following manufacturer's protocol. Immunoprecipitated proteins were eluted from beads by adding NuPage LDS sample buffer with reducing agent, and boiling at 95ºC for 10 minutes. For Western blots, primary antibodies were diluted as follows: anti-MIEF1 (Proteintech, 20164-1-AP, 1/1000), anti-HAUS6 (Thermo Fisher, PA5-31257, 1/500), and anti-GAPDH (Abcam, ab8245, 1/2000).

Microscopy

Tagged HEK293T cells were plated in 8-well ibiTreat µSlides (ibidi 80826) at 25-50,000 cells/well. For fixed cell imaging, on the following day, cells were fixed with cold methanol by removing the media and incubating in cold methanol for 3 min at -20°C. Cells were then washed with PHEM buffer (60 mM PIPES, 25 mM HEPES, 10 mM EGTA, 2 mM MgCl2, pH 6.9) twice. Cells were then incubated with 5 µg/mL DAPI at room temperature for 10 minutes in the dark to stain for the nucleus, and finally washed again with PHEM.

For live cell imaging of organelle localizations, tagged HEK293T cells were plated in 8-well ibiTreat µSlides (ibidi 80826) at 25-50,000 cells/well the day before. 4 hours after the cells have been seeded, CellLight BacMam 2.0 Reagents (Thermo Fisher) were added to the cells and incubated at 37°C overnight according to manufacturer's instructions. The CellLight localization reagents used include Mitochondria-RFP, Plasma Membrane-RFP, Golgi-RFP, and ER-RFP, and are indicated in the figures. The next day, the media was removed, and incubated with 1:2000 dilution of 10 mg/mL Hoescht 33342 solution (Thermo Fisher) in DMEM media at 37°C for 30 minutes. The staining media was then removed and replaced with imaging buffer (FluoroBrite DMEM media, Thermo Fisher).

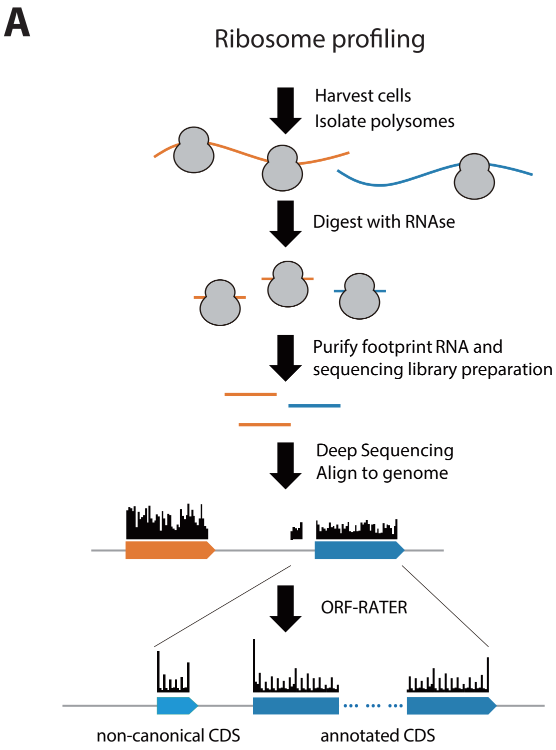
For live cell imaging of mitochondrial morphology, HEK293T MIEF1 uORF knockout cells were generated by TransIT-LT1 (Mirus) transfection of PX458 (pSpCas9(BB)-2A-GFP (PX458) was a gift from Feng Zhang (Addgene plasmid # 48138)) cloned with the sgRNA sequence GGCCCCGTGGAGCCGAGAGG. Two days after transfection, the GFP+ cells were then single-cell FACS sorted (SH800S, Sony) into 96-well plates. After the cells were grown up and expanded, genomic DNA was isolated with QuickExtract (Lucigen). The target site was amplified by nested-PCRs to make sequencing libraries, as described in (*63*), and sequenced on the MiSeq (Illumina). Sequencing data was processed with CRISPResso and then further analyzed using Python 2.7. A clone with a 8 bp deletion was chosen for further experiments (see fig. S14). For the overexpression cell line, the MIEF1 uORF sequence was ordered as a gblock and then assembled a pHR-SFFV-HA-IRES-GFP vector using Gibson Assembly (NEBuilder HiFi DNA assembly kit, New England Biolabs), between the SFFV promoter and IRES sequence (taking out the HA portion). The vector was transfected to either wild-type HEK293T cells (for overexpression studies) or MIEF1 uORF knockout HEK293T cells described above (for knockout and rescue studies) using TransIT-LT1 (Mirus). For controls, the original pHR-SFFV-HA-IRES-GFP vector was transfected as mock. For microscopy, the different cell lines were plated in 8-well ibiTreat µSlides (ibidi 80826) at 25-50,000 cells/well the day before. The next day, the media was removed, and incubated with 1:2000 dilution of 10 mg/mL Hoescht 33342 solution (Thermo Fisher) and 1:5000 of 1 mM MitoTracker Deep Red FM (Thermo Fisher) in DMEM media at 37°C for 30 minutes. The staining media was then removed and replaced with imaging buffer (FluoroBrite DMEM media, Thermo Fisher). Mitochondria morphology was analyzed as previously described (*44, 65*). 100 cells were counted for each condition.

Slides were imaged on a spinning disk confocal with Yokogawa CSUX A1 scan head, Andor iXon EMCCD camera and 100x ApoTIRF objective NA 1.49 (Nikon).

Pull-downs and mass spectrometry

Immunoprecipitation and mass spectrometry was performed essentially as previously described(*66*). Roughly 10^7 GFP11- or mNeonGreen11-tagged HEK293 cells, co-expressing GFP1-10 or mNeonGreen1-10 respectively, were washed with PBS and collected. The cells were lysed, clarified, and immunoprecipitated using anti-GFP beads or anti-mNeonGreen beads (Chromotek). Bound proteins were then digested on the beads with trypsin and prepared for mass spectrometry. Each construct was prepared and measured in triplicates.

Processing of raw mass spectrometry files was done with MaxQuant (1.6.3.4), using Label-free quantification and setting the LFQ min ratio count to 1, and searched against the UniProt complete human proteome sequence, as well as a custom fasta file with the bait peptide sequences (see Table S7). Data normalization, filtering, and imputation were performed using Perseus (1.6.5.0), and the enrichment analysis was performed using the "Hawaii plot" function. Missing values were imputed from a normal distribution with a downshift of 1.8 and a width of 0.15. The enrichment values and p-values were then exported to Python 2.7 for plotting the volcano plots.



B

ORF type	Num in iPSC	Num in fibroblasts
Annotated	9490	10452
New	818	225
Internal	282	232
Downstream	13	4
Upstream	2342	1621
Extensions	438	620
Truncations	1489	514
Isoforms	539	502

Fig. S1. Ribosome profiling and ORF-RATER to identify and annotate novel CDSs.

(**A**) Schematic of ribosome profiling. Ribosome profiling provides a global snapshot of ribosome occupancy and active translation. ORF-RATER is used to identify novel open reading frames from the ribosome densities outside of annotated regions. (**B**) Comparison of the number of identified ORFs in iPSCs (induced pluripotent stem cells) and HFFs (human foreskin fibroblasts). Similar number of ORFs is identified.

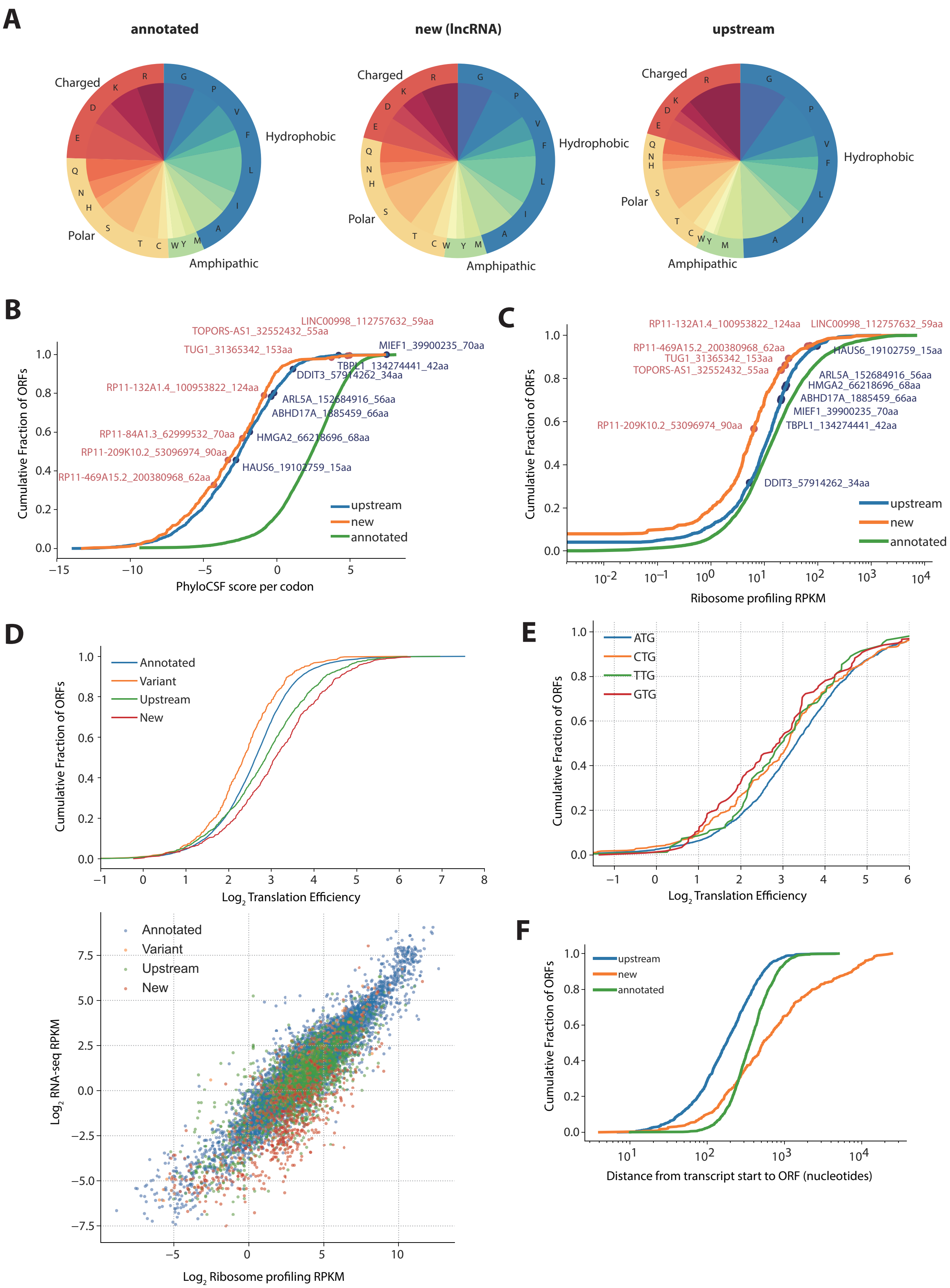

A
annotated
new (lncRNA)
upstream
Charged
Hydrophobic
Polar
Amphipathic
R
G
P
V
F
L
I
A
M
Y
W
C
T
S
H
N
Q
E
D
K
B
Cumulative Fraction of ORFs
PhyloCSF score per codon
LINC00998_112757632_59aa
TOPORS-AS1_32552432_55aa
TUG1_31365342_153aa
MIEF1_39900235_70aa
TBPL1_134274441_42aa
DDIT3_57914262_34aa
RP11-132A1.4_100953822_124aa
ARL5A_152684916_56aa
ABHD17A_1885459_66aa
RP11-84A1.3_62999532_70aa
HMGA2_66218696_68aa
RP11-209K10.2_53096974_90aa
HAUS6_19102759_15aa
RP11-469A15.2_200380968_62aa
upstream
new
annotated
–15
–10
–5
0
5
0.0
0.2
0.4
0.6
0.8
1.0
C
Cumulative Fraction of ORFs
Ribosome profiling RPKM
RP11-132A1.4_100953822_124aa
LINC00998_112757632_59aa
RP11-469A15.2_200380968_62aa
TUG1_31365342_153aa
TOPORS-AS1_32552432_55aa
HAUS6_19102759_15aa
ARL5A_152684916_56aa
HMGA2_66218696_68aa
ABHD17A_1885459_66aa
MIEF1_39900235_70aa
RP11-209K10.2_53096974_90aa
TBPL1_134274441_42aa
DDIT3_57914262_34aa
upstream
new
annotated
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
D
Cumulative Fraction of ORFs
Annotated
Variant
Upstream
New
Log$_2$ Translation Efficiency
–1
0
1
2
3
4
5
6
7
8
E
Cumulative Fraction of ORFs
ATG
CTG
TTG
GTG
Log$_2$ Translation Efficiency
Log$_2$ RNA-seq RPKM
Log$_2$ Ribosome profiling RPKM
Annotated
Variant
Upstream
New
7.5
5.0
2.5
0.0
–2.5
–5.0
–7.5
–5
0
5
10
F
Cumulative Fraction of ORFs
upstream
new
annotated
Distance from transcript start to ORF (nucleotides)

Fig. S2. Further characterizations of non-canonical CDSs identified by ribosome profiling.

(**A**) Amino acid composition of annotated coding regions, compared with identified “new” (lncRNA) CDSs and uORFs. Interestingly, the identified lncRNA ORFs have similar amino acid composition as annotated CDSs, but uORFs tend to encode more hydrophobic amino acids than annotated CDSs. (**B**) Cumulative distribution of the PhyloCSF score per codon (conservation score) for lncRNA CDSs (“new”), uORFs, and annotated coding regions, with a few examples highlighted. (**C**) Cumulative distribution of translation rates (RPKM, Reads Per Kilobase of transcript, per Million mapped reads) for lncRNA CDSs (“new”), uORFs, and annotated coding regions, with a few examples highlighted. (**D**) Top: cumulative distribution of translation efficiency (TE) for the different ORF types. Bottom: comparison of the RNA-seq RPKM and ribosome profiling RPKM for the four different ORF categories. Ribosome profiling RPKM divided by RNA-seq RPKM is the TE. (**E**) The TE for non-canonical CDSs (for “new” CDSs and uORFs) initiated with the four different start codons. CDSs initiated with the canonical ATG start codon have slightly higher TEs, followed by CTG, TTG, and finally GTG. (**F**) Cumulative distribution of the distance from the transcript start site to the ORF (in nucleotides, nt). Interestingly, some lncRNA CDSs are located much farther down the transcript than annotated CDSs, implicating possible cap-independent initiation mechanisms, though this hypothesis requires further studies and verifications.

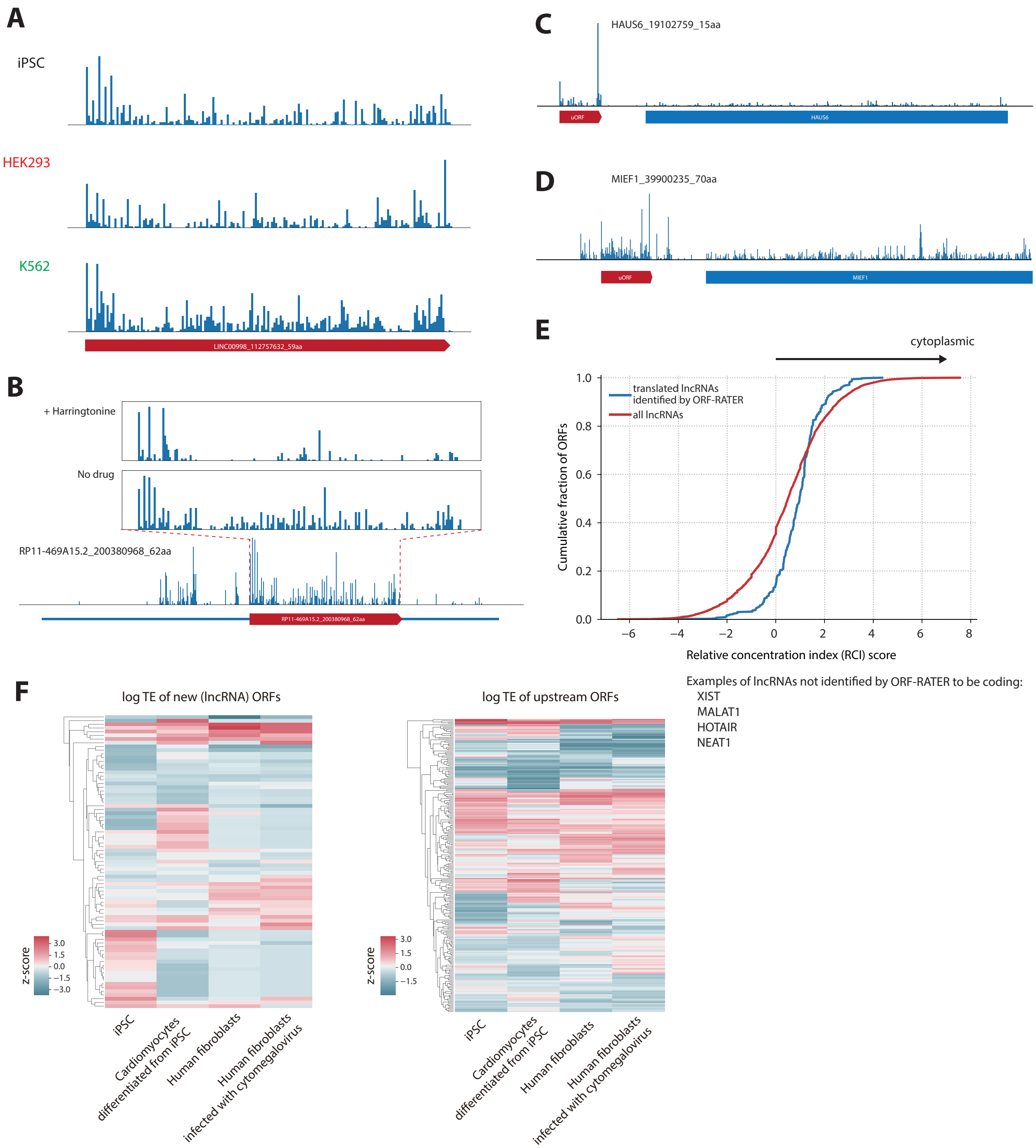


Fig. S3. Further lines of evidence suggesting the non-canonical CDSs identified by ribosome profiling are actively translated.

(**A**) Example ribosome profiling traces of the 59 amino acid lncRNA peptide on *LINC00998* in three cell types, K562, HEK293, and iPSC, showing the peptide is expressed in multiple cell types. (**B**) Example ribosome trace of the lncRNA peptide on *RP11-469A15*, encoding a 62 amino acid microprotein. (**C**) Example ribosome profiling trace of a uORF upstream of the HAUS6 protein, encoding a 15 amino acid peptide. (**D**) Example ribosome profiling trace of a uORF upstream of the MIEF1 protein, encoding a 70 amino acid peptide. (**E**) The cumulative distribution of relative concentration index (RCI) of all lncRNAs versus lncRNAs identified to be peptide-coding by ORF-RATER. RCI is a comparison of the concentration of a gene, per unit mass of RNA, between the nucleus and cytoplasm, with values obtained from lncATLAS (*67*). A positive RCI score suggests predominant localization in the cytoplasm. (**F**) Heatmap of z-score changes in expression (TE, translational efficiency), upon iPSC-differentiation into cardiomyocytes and infection of HFFs with cytomegalovirus (CMV). TE is defined by ribosome profiling RPKM / RNA-seq RPKM. RPKM is defined as Reads Per Kilobase of transcript, per Million mapped reads.

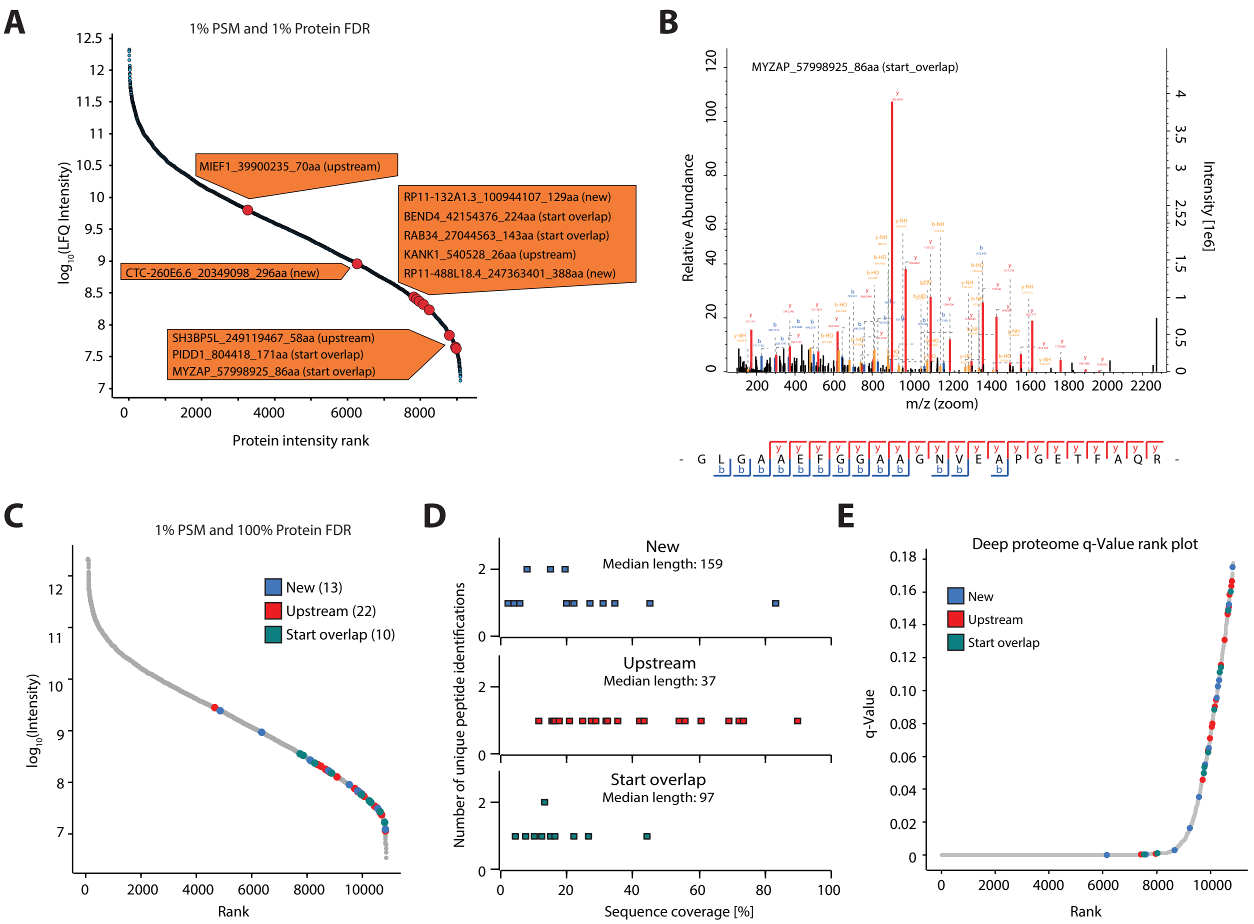
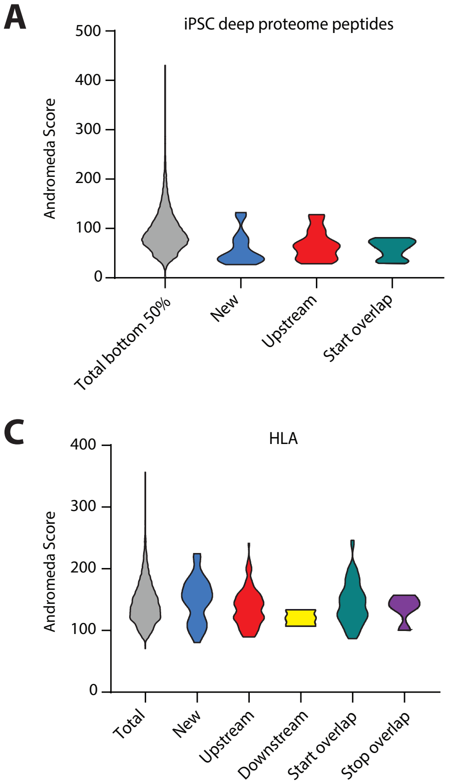


Fig. S4. Mass spectrometry-based proteomics in iPSCs confirm non-canonical proteins identified by ribosome profiling.

(**A**) Full proteome abundance rank plot of the iPS cell line for 1% PSM (peptide spectral match) FDR and 1% protein FDR. To investigate if non-canonical CDS proteins predicted by ribosome profiling can be identified and quantified by MS-based proteomics, a deep representative tryptic proteome of the iPSC line was acquired, identifying more than 9,100 protein groups and more than 78,000 peptides when searching against the human proteome fasta file appended with the custom ORF fasta file from ribosome profiling to keep the power for the FDR calculation. All peptide identifications mapping to non-canonical CDSs were manually inspected for amino acid overlap to proteins in the human proteome fasta file and with regard to their spectral quality in terms of b- and y-ion series. A total of 10 high-confidence, non-canonical CDS proteins resulted from this analysis with a median Andromeda score of 105, which is in agreement with the median score of 106 for human proteome annotation. Interestingly, all non-canonical CDS identifications belong to the lower abundance range of the deep iPSC proteome. This supports the need of complementary techniques in addition to MS-based proteomics to define the full protein-coding capacity of the genome. (**B**) Exemplary MS/MS spectrum of a peptide belonging to the least abundant non-canonical CDS identification MYZAP_57998925_86aa is shown yielding a full b- and y-series, confirming its existence and identification. (**C**) Full proteome abundance rank plot of the iPS cell line, similar to (A), but for 1% PSM FDR and 100% protein FDR. We identified a total of 45 non-canonical CDS peptides in this analysis. (**D**) Sequence coverage of identified non-canonical CDS peptide classes versus numbers of unique identified peptides mapping to these classes. We identified "new" ORFs with a median length of 159 amino acids (aa), "upstream" with 37 aa, and "start overlap" with 97 aa. We identified most of the CDSs with a single unique peptide, but with a very high sequence coverage due naturally small protein size, compared to a median sequence coverage per protein in the deep proteome of 16.4%. (**E**) Full proteome Q-value rank plot of the iPS cell line. Due to the small size of the non-canonical CDSs naturally yielding only one tryptic peptide per protein and occupying the lower abundance range within the total proteome, we searched the raw data with a 1% PSM and 100% Protein FDR to increase the chance of identifying biologically meaningful non-canonical CDS identifications.



B

Category	1% PSM and 1% Protein FDR			1% PSM and 100% Protein FDR		
	Numbers of Protein identifications	Median protein weight (kDa)	Median Peptide Score	1% Peptide, 100% Protein group FDR	Median protein weight (kDa)	Median Peptide Score
New	3	35.131	50.61	13	8.342	42.59
Upstream	3	4.433	128.05	22	3.546	61.96
Start overlap	5	19.818	66.23	10	9.949	62.49
Total	9,124	52.451	106.6	10,847	50.47	105.17

D

Category	Number of identifications	Median Score
New	59	152.79
Upstream	110	134.79
Downstream	3	120.45
Start overlap	80	139.68
Stop overlap	5	138.85
Total	23,613	139.68

Fig. S5. Comparison of proteomic Andromeda scores for the different non-canonical CDS types.

(**A**) Violin plots of the peptide Andromeda scores for the different categories of non-canonical CDSs for the deep iPSC proteome analysis, as well as the score for the bottom 50% abundance range of the total peptides residing from the deep proteome (which represents the cleanest background comparison, because all non-canonical CDS identifications fall into this abundance range). (**B**) Non-canonical CDS identifications within the deep proteome with 1% PSM and 1% Protein FDR versus 1% PSM and 100% Protein FDR. We are aware of the fact that the median overall non-canonical CDS score on the peptide level is lower than the median score of the overall data set. We attribute this to the fact that the score is calculated as the product of each individual peptide posterior error probability of each protein group and it includes a factor that takes the number of peptide identifications per protein group into account(*68*). (**C**) Violin plots of the peptide Andromeda scores of non-canonical CDSs and the total background identifications for the HLA peptidome analysis. Scores for all categories are on the same level. (**D**) HLA class I non-canonical CDS peptide identifications. Interestingly, the score distribution for the HLA peptidome data set is uniform across all categories of non-canonical CDSs. We attribute this to the fact that HLA peptides are of non-tryptic origin, which results in different ionization and fragmentation properties compared to tryptic peptides. Furthermore, they represent a distinct molecule class, without the need for protein inference.

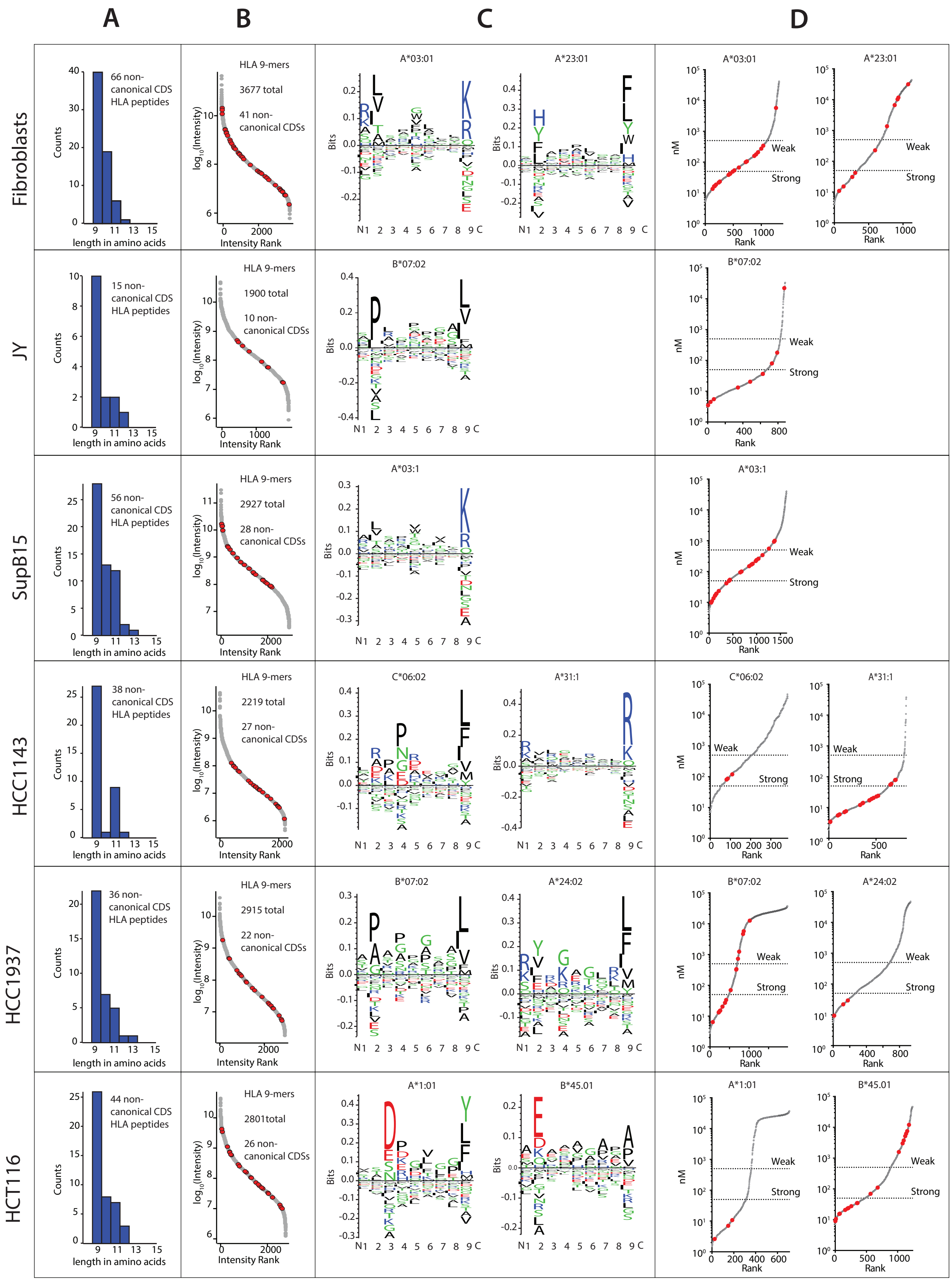

A
B
C
D
Fibroblasts
66 non-canonical CDS HLA peptides
Counts
length in amino acids
HLA 9-mers
3677 total
41 non-canonical CDSs
log$_{10}$(Intensity)
Intensity Rank
A*03:01
A*23:01
Bits
N1 2 3 4 5 6 7 8 9 C
nM
Weak
Strong
Rank
JY
15 non-canonical CDS HLA peptides
HLA 9-mers
1900 total
10 non-canonical CDSs
B*07:02
SupB15
56 non-canonical CDS HLA peptides
HLA 9-mers
2927 total
28 non-canonical CDSs
A*03:1
HCC1143
38 non-canonical CDS HLA peptides
HLA 9-mers
2219 total
27 non-canonical CDSs
C*06:02
A*31:1
HCC1937
36 non-canonical CDS HLA peptides
HLA 9-mers
2915 total
22 non-canonical CDSs
B*07:02
A*24:02
HCT116
44 non-canonical CDS HLA peptides
HLA 9-mers
2801total
26 non-canonical CDSs
A*1:01
B*45.01

Fig S6. Cross-validation, quality assessment and high-confidence identification of non-canonical CDS HLA-I peptides.

(**A**) Non-canonical CDS HLA-I peptides matching the ORF type categories “new”, “upstream”, “downstream”, “start overlap”, and “stop overlap” in six different cell lines (Fibroblasts, JY, SupB15, HCC1143, HCC1937, HCT116) with known HLA allotypes from reference (*30*). Length distribution of all identified non-canonical CDS derived HLA class I peptides within the six cell lines, confirming a pronounced 9-mer HLA length distribution. (**B**) Intensity distribution of all 9-mer HLA class I peptides found in each cell line. Non-canonical CDS derived HLA class I peptides identified in each cell line are distributed across the entire abundance range and highlighted in red. (**C**) Gibbs-clustering of all 9-mer non-canonical CDS HLA-I peptide identifications based on the highest Kullbach Leibler distance resembling sequence logos matching the HLA allotypes in each particular cell line. (**D**) Binding affinity prediction with NetMHCcon of all 9-mer HLA class I peptides from each cell line matching the HLA peptide anchor point resolved allotype. The non-canonical CDS derived HLA class I peptides are highlighted in red. More than 75% of all non-canonical CDS HLA-I peptides are predicted to be either strong binders (Predicted affinity of ≤50 nM) or weak binders (Predicted affinity of >50 nM and ≤500 nM), confirming their affinity to their particular HLA-I allotype.

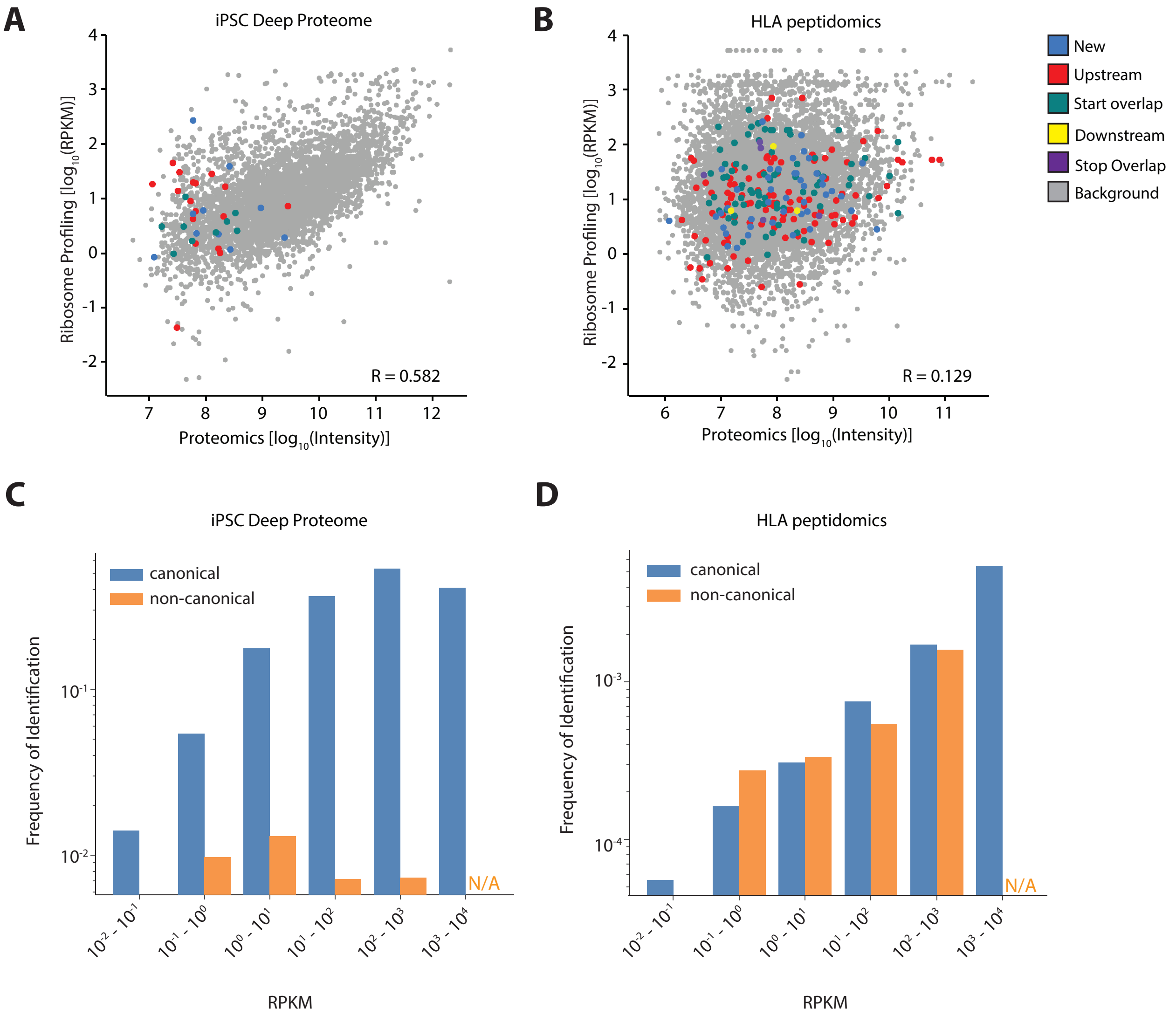


Fig. S7. Comparison of translation rates from ribosome profiling with peptide abundance from proteomics.

(**A**) Comparison of RPKM (Reads Per Kilobase of transcript, per Million mapped reads) values obtained from the ribosome profiling with the intensity values obtained from the deep proteome iPSC and the HLA peptidome analysis. Interestingly, the iPSC deep proteome expression levels correlate well with the RPKM values from ribosome profiling with a Pearson correlation of *R* = 0.582, indicating that the active protein translation correlates with the final proteome abundance and homeostasis. (**B**) In the HLA case, we do not see any correlation (*R* = 0.129), which makes biologically sense, since the cellular HLA peptidome processing pathway and MHC presentation are independent of direct protein translation. (**C**) We are interested in examining why peptides from non-canonical CDSs are not well detected by proteomic mass spectrometry (MS). Possible reasons are: (1) the lower abundance of non-canonical CDSs and (2) their shorter length, hence fewer tryptic peptides per non-canonical CDSs. As a back-of-the-envelope calculation, we can (1) bin the non-canonical and canonical CDSs based on RPKM values, (2) for each bin, compute in silico the number of all tryptic peptides that can be detected from non-canonical and canonical CDSs, and (3) for each bin, divide the number of tryptic peptides detected by MS by the total number of possible tryptic peptides. For the iPSC deep proteome dataset, the ratios for the non-canonical CDSs are lower than the ratios for canonical CDSs. The result is consistent with the hypothesis that the short lengths of the non-canonical CDSs result in not only fewer tryptic peptides, but also tryptic peptides that are not well detected by the MS instrument. Indeed, for the non-canonical CDSs detected, the Andromeda score is on the lower end of the distribution (see fig. S5). Thus, the short length, amino acid composition, and the ionization and fragmentation properties of the non-canonical CDSs do not yield optimal tryptic peptides for detection by mass spectrometry-based bottom-up proteomics. (**D**) We calculate the ratios similar to (C), except for the HLA peptidomics dataset rather than the iPSC deep proteome dataset. Furthermore, since HLA class I peptides are of non-tryptic origin and are processed endogenously through the MHC presentation machinery to yield peptides 8 – 15 bp in length (predominantly 9 bp), we do not calculate the number of tryptic peptides as in (C). Instead, we estimate the total number of possible peptides by calculating the total number of 8 – 15 bp fragments that tile across the entire CDS. Here, the ratios for the non-canonical CDSs and canonical CDSs are similar. In this case, since both the canonical and non-canonical CDSs are processed into fragments of 8 – 15 bp length, detection by the MS instrument will be comparable (unlike the trypsin digestion case in (C)). Indeed, the Andromeda score distribution for the non-canonical CDSs are evenly distributed across the entire range of canonical CDSs (see Fig. 1F and fig. S5). The similar ratios suggest that the non-canonical peptides are indeed produced in the cell and processed by the HLA class I peptide presentation machinery in the same frequency and nature as canonical proteins.

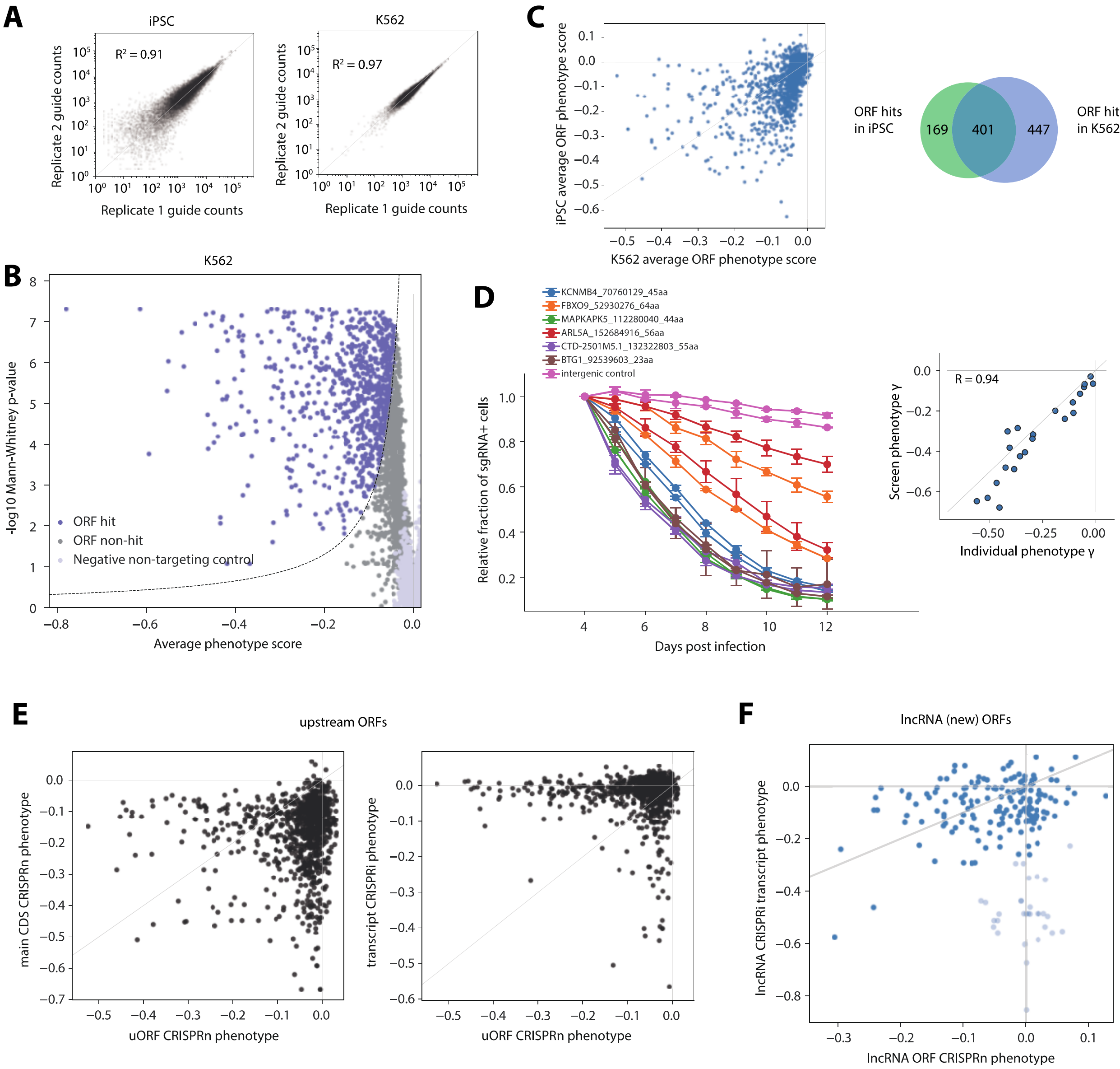


Fig. S8. Validation of CRISPR screens.

(**A**) Correlation of read counts from next-generation sequencing of individual sgRNAs between experimental replicates in K562 and iPSC. (**B**) Volcano plot, similar to Fig. 2, summarizing knockout phenotypes and statistical significance (Mann-Whitney U test) for ORFs targeted in the K562 screen. Each dot on the graph is an ORF targeted, and ORF hits are labeled in purple, with a more negative phenotype score indicating a stronger growth defect. (**C**) Comparison of phenotypes between iPSCs and K562s. A large overlap (401) of hits is shared between the two cell types. (**D**) Individual validation of sgRNAs in internally controlled, competitive growth assays performed with sgRNAs targeting ORF hits in K562 cells. Cells were infected with lentivirus of the sgRNA expression vector (including a blue fluorescent protein (BFP) marker gene) and passaged. The fraction of sgRNA-containing cells was measured as the fraction of high BFP expressing cells by flow cytometry, and expressed relative to the fraction at 4 days after infection. The lines with the same color represent two different sgRNAs targeting the same

ORF. Points represent the mean and standard deviation of replicates. The phenotype scores found in the individual validations correlate with the scores found in the pooled CRISPR screen (R = 0.94). (**E**) Comparison of the uORF knockout phenotype with the knockout of the downstream protein by CRISPRn (nuclease) and the knockdown of the entire transcript by CRISPRi. The uORF knockout phenotype with the downstream, main CDS knockout phenotype is not correlated, suggesting that the hits are not due to accidentally disrupting the main CDS. A fraction of the uORF hits do not have main, canonical CDSs with fitness defects upon knockout, suggesting an independent function of the uORFs or that disruption of the uORFs result in increases in main CDS expression that results in the growth phenotype. Here, we include the comparison with CRISPRi for completeness, but want to note that it might not be appropriate to compare CRISPRn and CRISPRi phenotype values since the mechanism of silencing is different. (**F**) Comparison of the lncRNA CDS knockout versus the transcript knockdown by CRISPRi. We can see for some CDSs, the knockdown and knockout phenotypes seem to be weakly correlated (R = 0.31, labeled in dark blue), suggesting that the lncRNA functions either on the peptide level or both on the RNA and peptide level. On the other hand, for some lncRNA CDSs, there is only a knockdown phenotype, with little to no phenotype when the CDS is knocked out (labeled in light blue), suggesting these may be true lncRNAs. Though, as emphasized above, it might not be appropriate to compare CRISPRn and CRISPRi phenotype values.

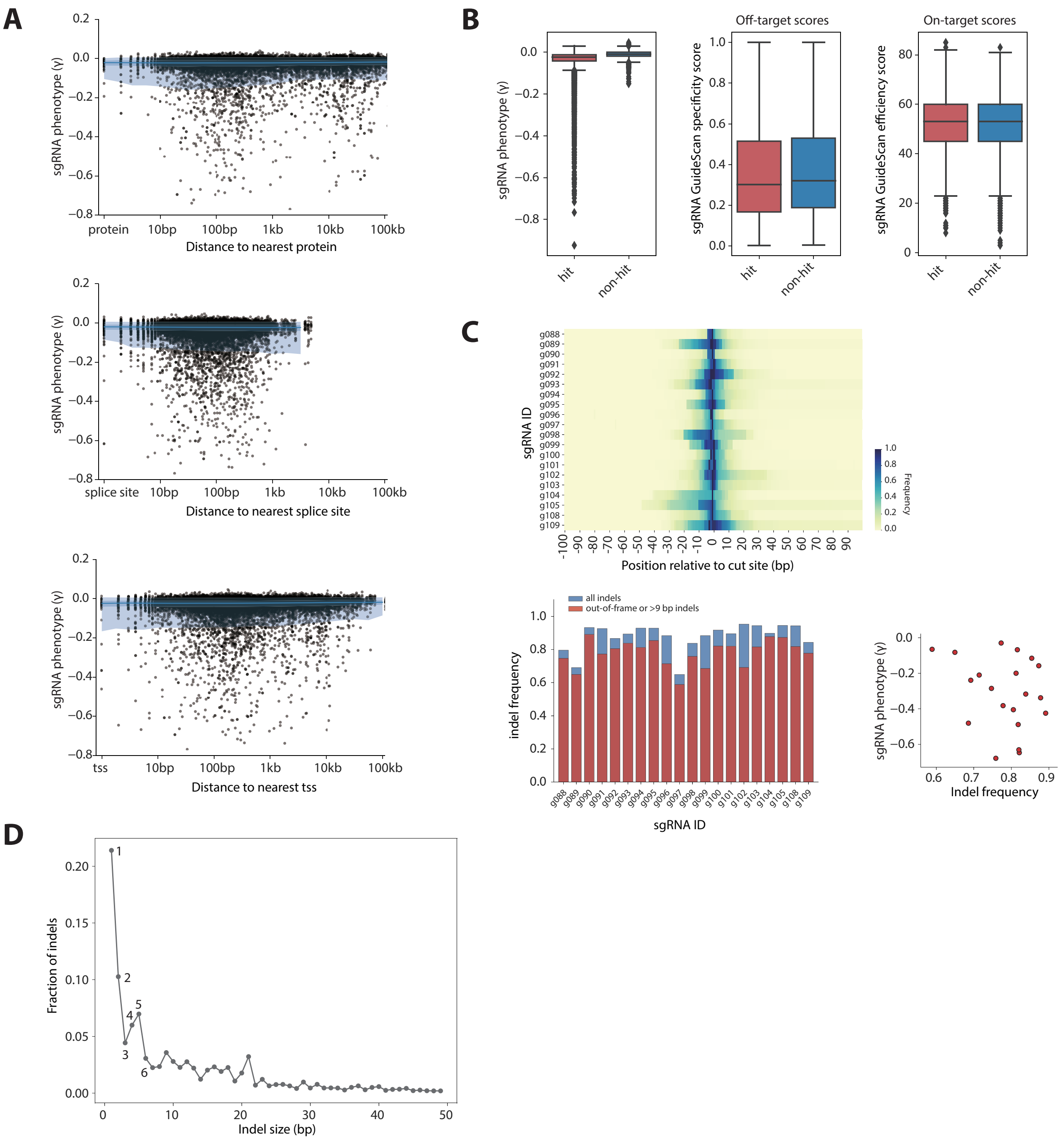


Fig. S9. Follow-up validation of CRISPR screens: analysis of on-targets, off-targets, and indels.

(**A**) Distribution of sgRNAs relative to distance to protein-coding gene, splice site, or transcript start site (TSS) (in base pairs, bp). Points indicate individual sgRNA phenotypes and blue shaded regions represent 5th, 25th, 50th, 75th, and 95th percentiles. There is not an observed bias of stronger sgRNA phenotypes closer to these sites. (**B**) The off-target efficiencies and on-target efficiencies with the GuideScan algorithm(*32*) of sgRNAs targeting ORF hits and ORF non-hits show no observable difference. This indicates the phenotypes we observed from the screen are not due to differences in off-target or on-target efficiencies. (**C**) Top: Heatmap of indel sizes for select sgRNAs. The sgRNA sequences are included in Table S8. Most indels created are less than 25 nucleotides, so the short ORFs can be precisely targeted without affecting nearby transcript elements. Bottom: The frequency of indels for select sgRNAs, showing greater than 90% indel frequencies for most sgRNAs, as well as their correlation with the guide phenotype score. (**D**) Histogram of the different indel sizes. The indels are predominantly 1 or 2 bp insertions or deletions.

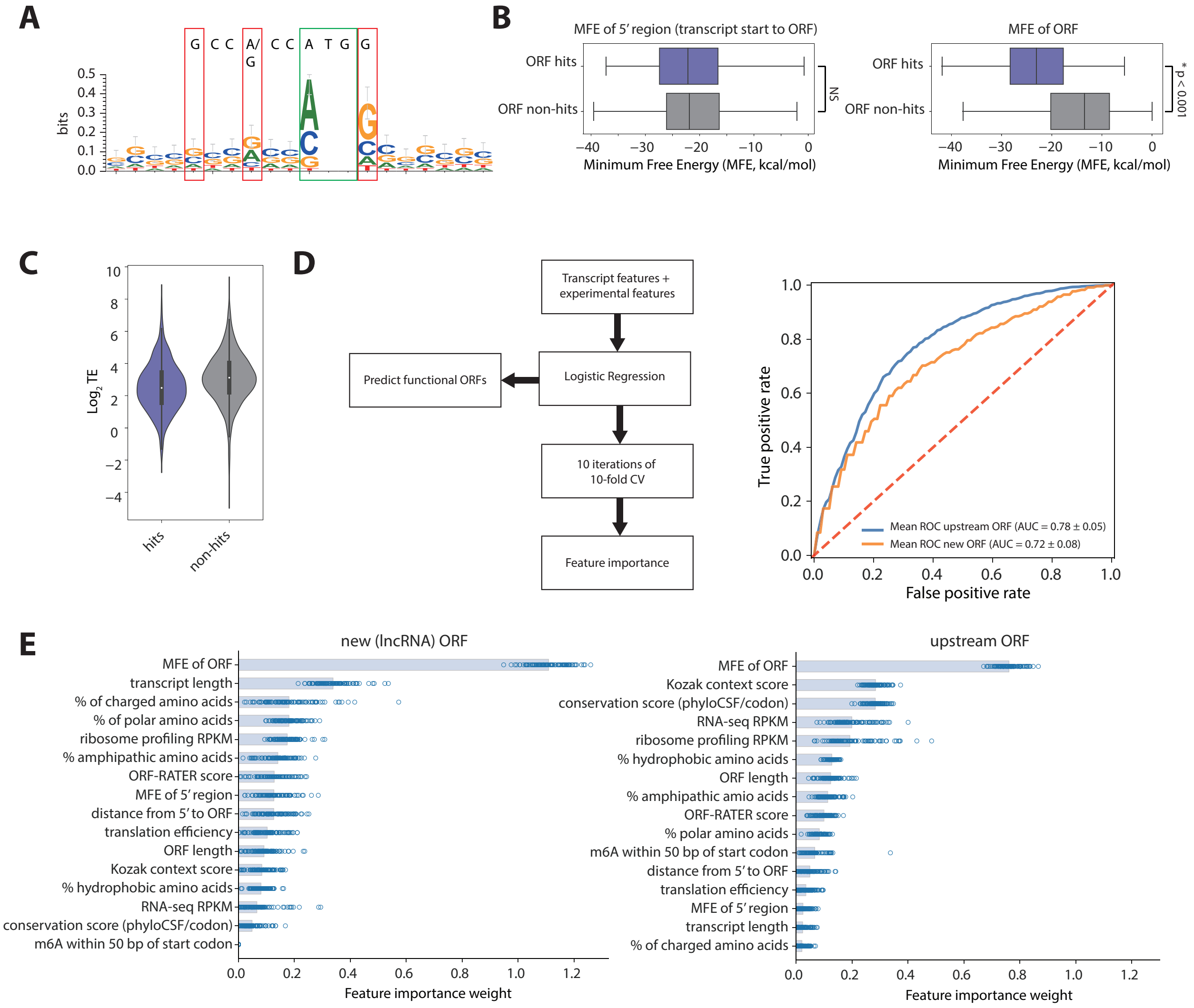


Fig. S10. Features that distinguish ORF hits and non-hits.

(**A**) The start codons of ORF hits are enriched with the Kozak motif that plays a role in ribosome initiation. (**B**) The sequences of the ORF hits tend to have a more negative min-free energy (MFE) compared to the non-hits ($P < 10^{-20}$, Mann-Whitney test), suggesting possible secondary structures. The sequences 5' of the ORF have no significant difference between ORF hits vs. non-hits. (**C**) Comparison of translation efficiency (TE) for hits vs. non-hits. (**D**) To incorporate multiple features and determine the top distinguishers, we can train a machine learning classifier (logistic regression), followed by 10 iterations of 10-fold cross validation. Receiver operating characteristic (ROC) curves of the lncRNA ORF model and the uORF model shows AUC (area under the curve) is 0.72 and 0.78 respectively. (**E**) Results from logistic regression models and 10-fold cross-validation. These top distinguishing features between hits vs. non-hits suggest possible implications for the mechanism of alternative start site selection.

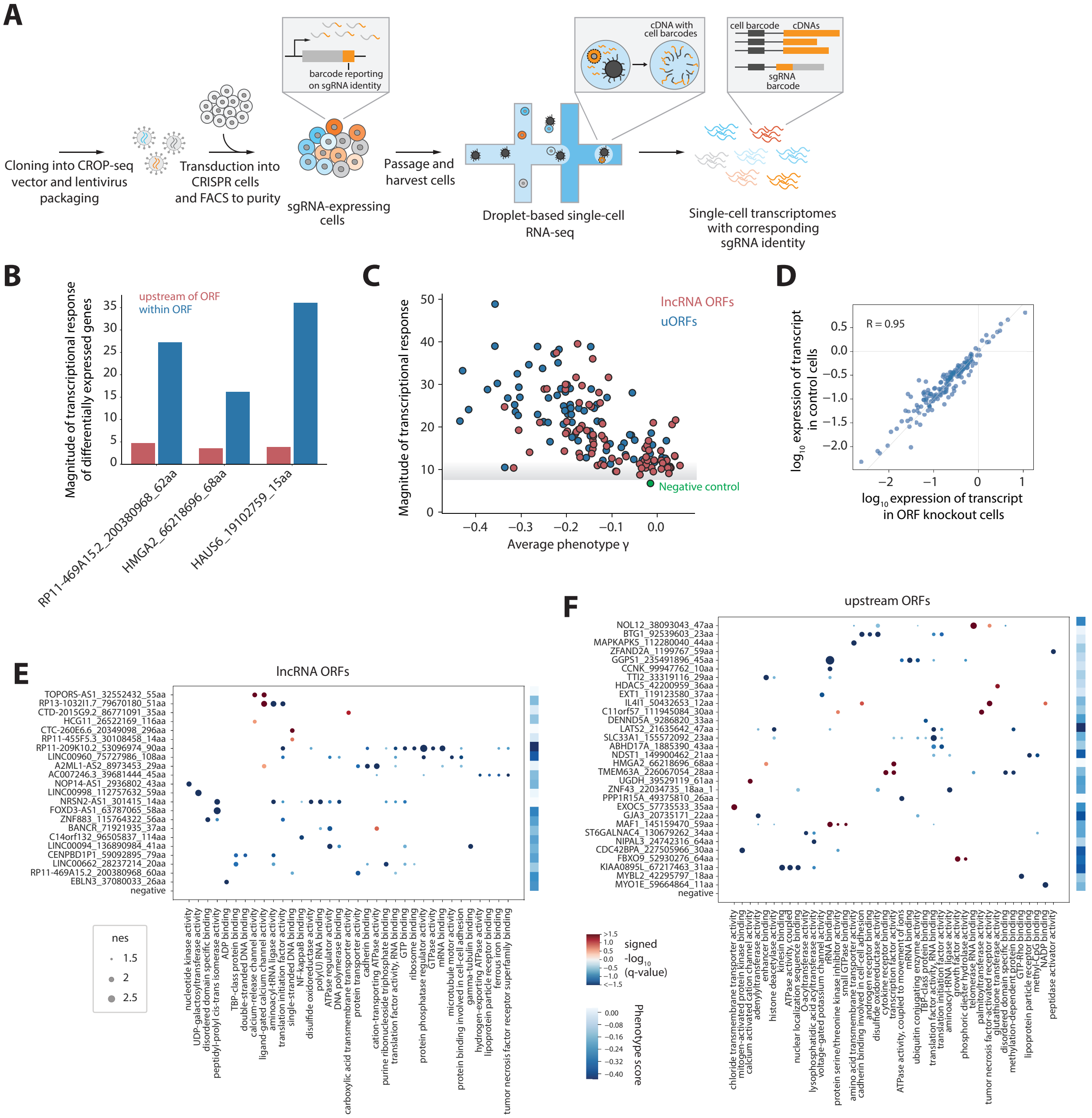


Fig. S11. Perturb-Seq reveals non-canonical CDSs play diverse cellular roles.

(**A**) Schematic of Perturb-seq strategy to capture single-cell transcriptomes with matched sgRNA identities. (**B**) As an additional control, we targeted sgRNAs immediately upstream in the genome of the ORF, similar to Fig. 2C. Targeting upstream of the ORF elicits a much weaker transcriptional response compared to targeting within the ORF. The magnitude of transcriptional response is defined by the absolute sum of z-scores from differentially expressed genes identified by a random-forest classifier. (**C**) The magnitude of transcriptional response upon ORF knockout from the Perturb-Seq screen in iPSCs versus the ORF knockout growth phenotype score from the CRISPR screen. Perturb-Seq allows for the identification of functional CDSs where the knockout gave no growth phenotype (**D**) Comparison of the transcript levels in control cells versus ORF-disrupted cells using Perturb-Seq. There is no significant difference, suggesting that, at least in these cases, knockout does not induce a decrease of transcript abundance through mechanisms such as non-sense mediated decay. (**E, F**) Dot plots of signed q-values and normalized enrichment scores (nes) of Gene Ontologies from Gene Set Enrichment Analysis (GSEA) of the transcriptome response for knockouts of lncRNA ORFs and uORFs from Perturb-Seq, showing activities in diverse biological processes. Only the q-values that are significant (less than 0.25) are shown. Heatmap of the corresponding ORF knockout growth phenotype score is shown to the right.

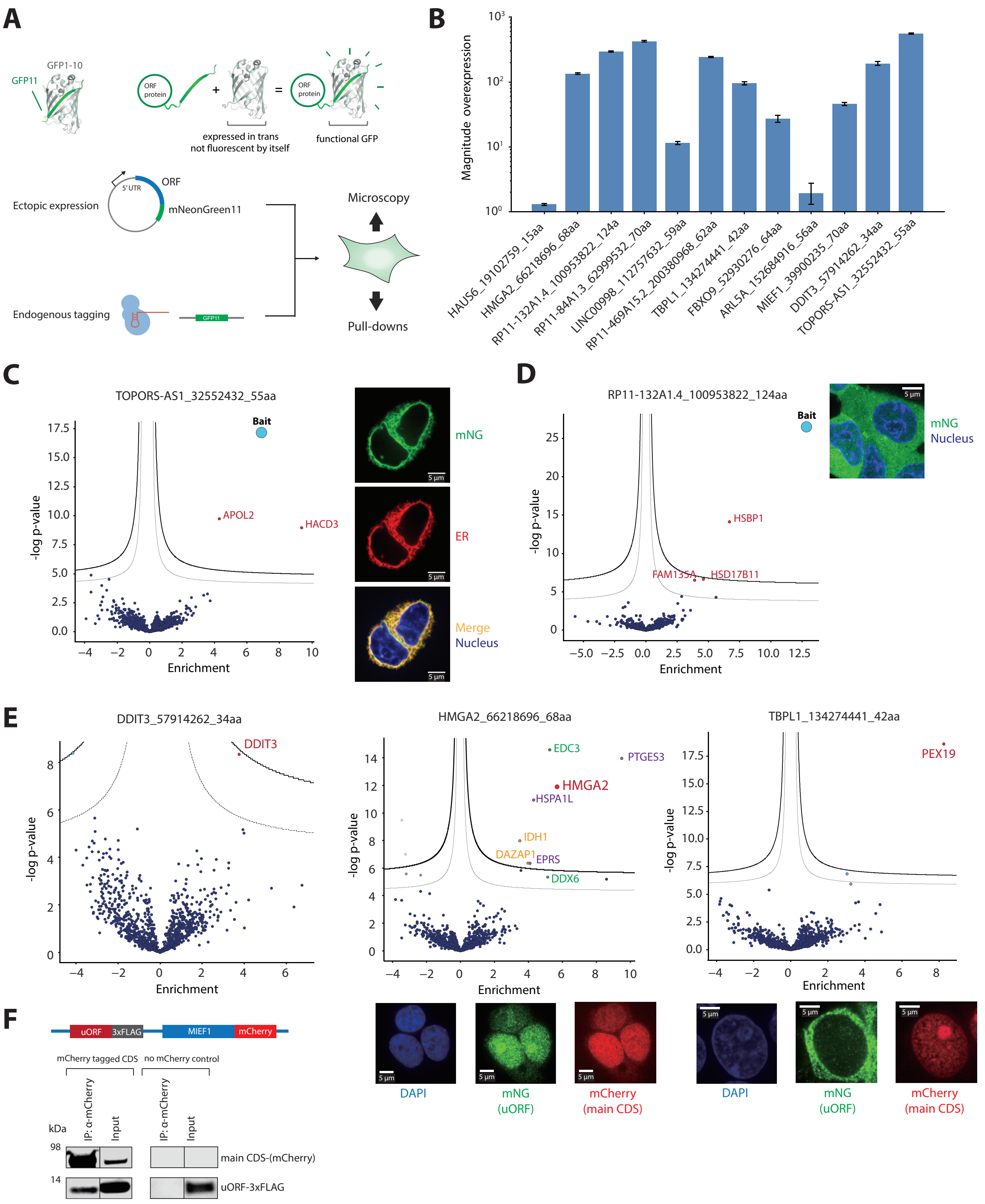


Fig. S12. Additional co-immunoprecipitation and mass spectrometry (co-IP MS) of tagged microproteins.

(**A**) Schematic of the split-GFP system. The split-mNeonGreen (mNG) system works similarly. (**B**) Estimates of the magnitude of overexpression for the ectopically-expressed constructs used for co-IP MS, compared to endogenous transcript levels, as quantified by qPCR. (**C**) Volcano plot of the co-IP MS of the 55 amino acid peptide encoded on lncRNA *TOPORS-AS1.* Thick threshold line is 1% FDR, and the thin threshold line is 5% FDR. The bait (the tagged peptide) is labeled in blue. The peptide interacts with APOL2, a cytosolic protein that may affect the movements of lipids or allow the binding of lipids to organelles, and HACD3, which plays a part in the catalysis of the long-chain fatty acids elongation cycle. The peptide localizes to the endoplasmic reticulum. (**D**) *RP11-132A1.4* encodes a 124 amino acid peptide that interacts with HSBP1, a heat shock binding protein, HSD17B11, a dehydrogenase, and FAM135A, a protein with predicted hydrolase and acyltransferase activity. The pepide localizes generally to the cytoplasm. (**E**) Co-IP MS of three uORF microproteins. The DDIT3 uORF peptide interacts exclusively with DDIT3, the downstream-encoded protein. The uORF peptide of HMGA2 co-immunoprecipitates with HMGA2, as well as other proteins such as the mRNA decapping protein EDC3 and the heat-shock related proteins HSPA1L and PTGES3. The uORF peptide for TBPL1 interacts with PEX19, a cytosolic chaperone and import receptor for peroxisomal membrane proteins. In some cases, here and in Fig. 4, the bait was not identified. This may be because the amino acid composition of the peptide sequence or the peptide length after Trypsin digestion is not amenable to being detected by the mass spectrometer. Microscopy images of uORF peptides tagged with mNG11 (green), expressed ectopically (in the native transcript context) in a HEK293T cell line expressing mNG1-10. The main CDS protein tagged with mCherry (red) is co-expressed. The uORF peptides are expressed and may localize to the same or distinct parts of the cell relative to the main CDS protein. (**F**) The uORF peptide interaction is confirmed by pulling down on the main MIEF1 protein and immunoblotting against the uORF peptide.

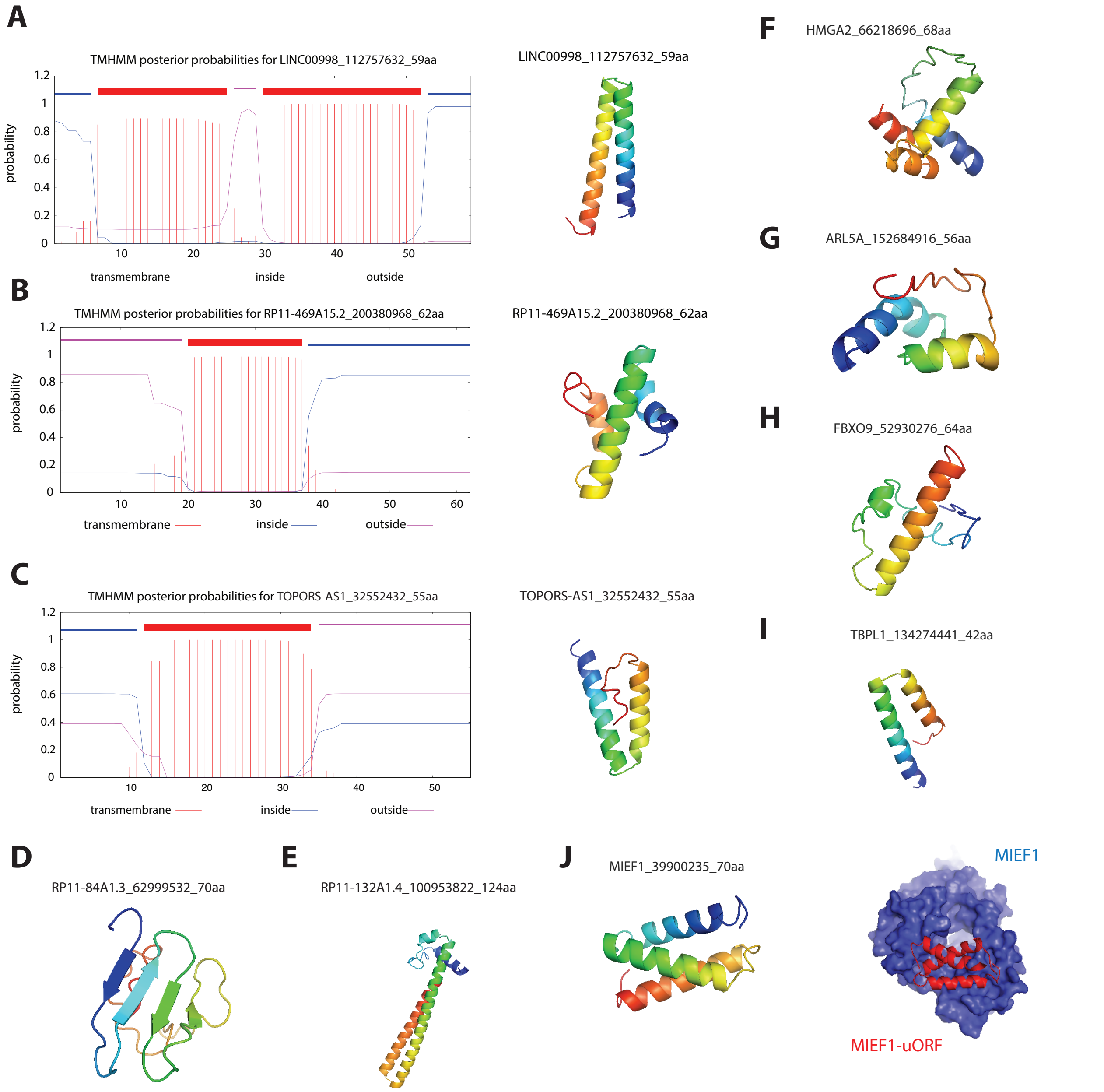


Fig. S13. Predicted structures for lncRNA and uORF microproteins.

(**A-C**) Predicted transmembrane domains from TMHMM, as well as predicted structures using Quark(*69*), for three lncRNA peptides. (**D-E**) Predicted structure of two lncRNA peptides that did not have a predicted transmembrane domain. (**F-I**) Predicted structure of four uORF peptides. (**J**). Predicted structure of the 70 amino acid peptide encoded by the uORF of MIEF1. The uORF peptide potentially interacts with the canonical MIEF1 protein, as predicted by ClusPro(*70*), a protein-protein docking software. Further biochemical and structural studies will be needed to understand the nature of such an interaction.

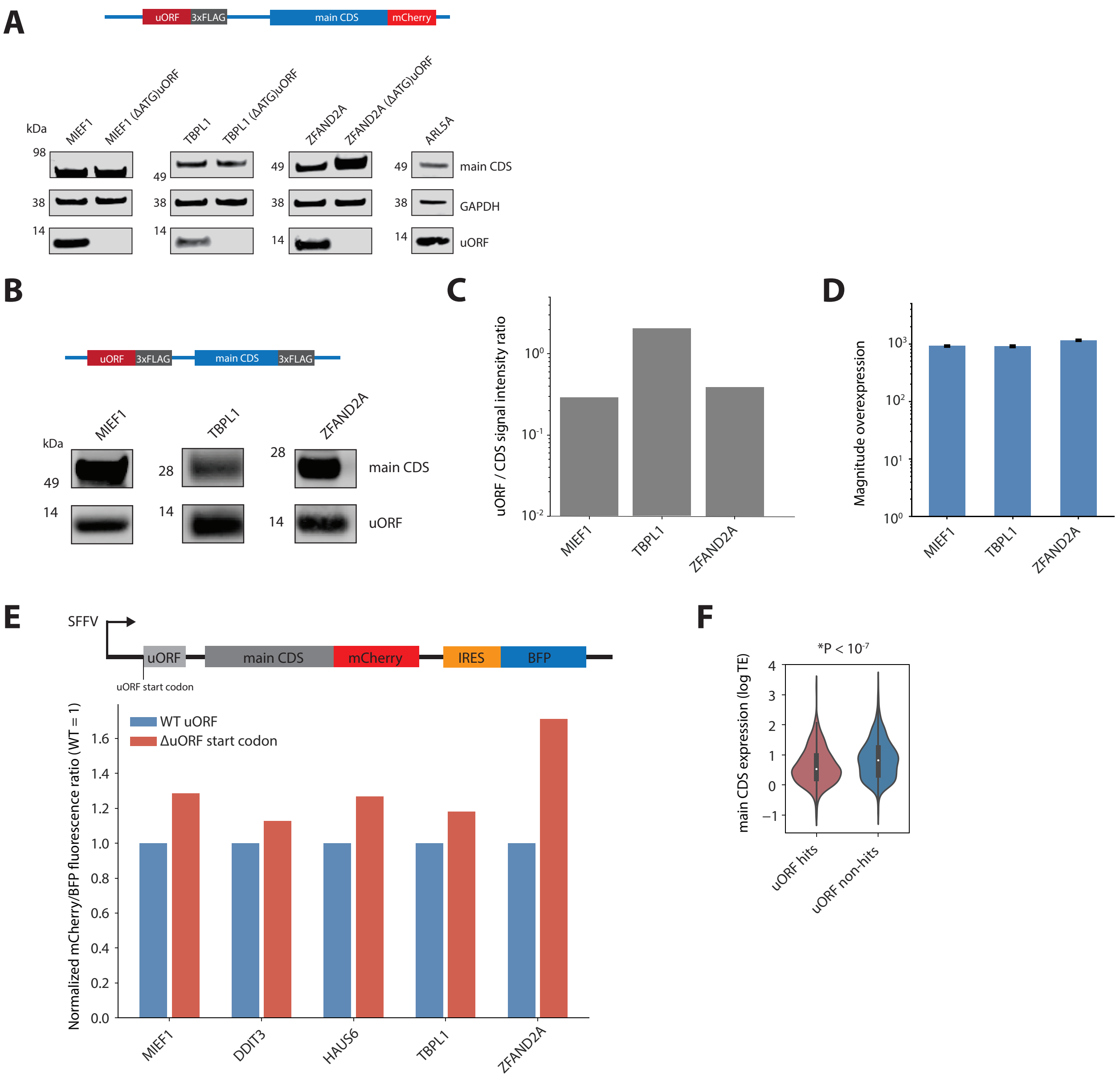


Fig. S14. Further characterizations of uORFs.

(**A**) Western blot of tagged uORF and the main CDS in their native transcript context, indicating two separate protein products are translated. (**B**) Western blots similar to (A), except both the uORF and main CDS are tagged with 3xFLAG. (**C**) Quantification of the Western blot in (B) to determine the relative stoichiometry of the uORF peptide and main CDS protein. (**D**) Magnitude of overexpression for the transfected constructs used in (B), compared to endogenous transcript levels, as quantified by qPCR. (**E**) Top, the reporter construct design. To enable normalization for transfection and transcription efficiency, a BFP sequence whose translation was driven by an internal ribosome entry site (IRES) was included. Bottom, fluorescence measurements following transient transfection of reporter constructs into HEK 293T cells. Removal of the initiating uORF start codon increased mCherry fluorescence of the main CDS, suggesting that translation of the uORFs inhibits downstream translation, though this effect is not strong. (**F**) The expression (translational efficiency, TE) of the main, annotated protein downstream of uORF hits are, in general, lower in expression compared with non-hits (Mann-Whitney P-value < 10^{-7}).

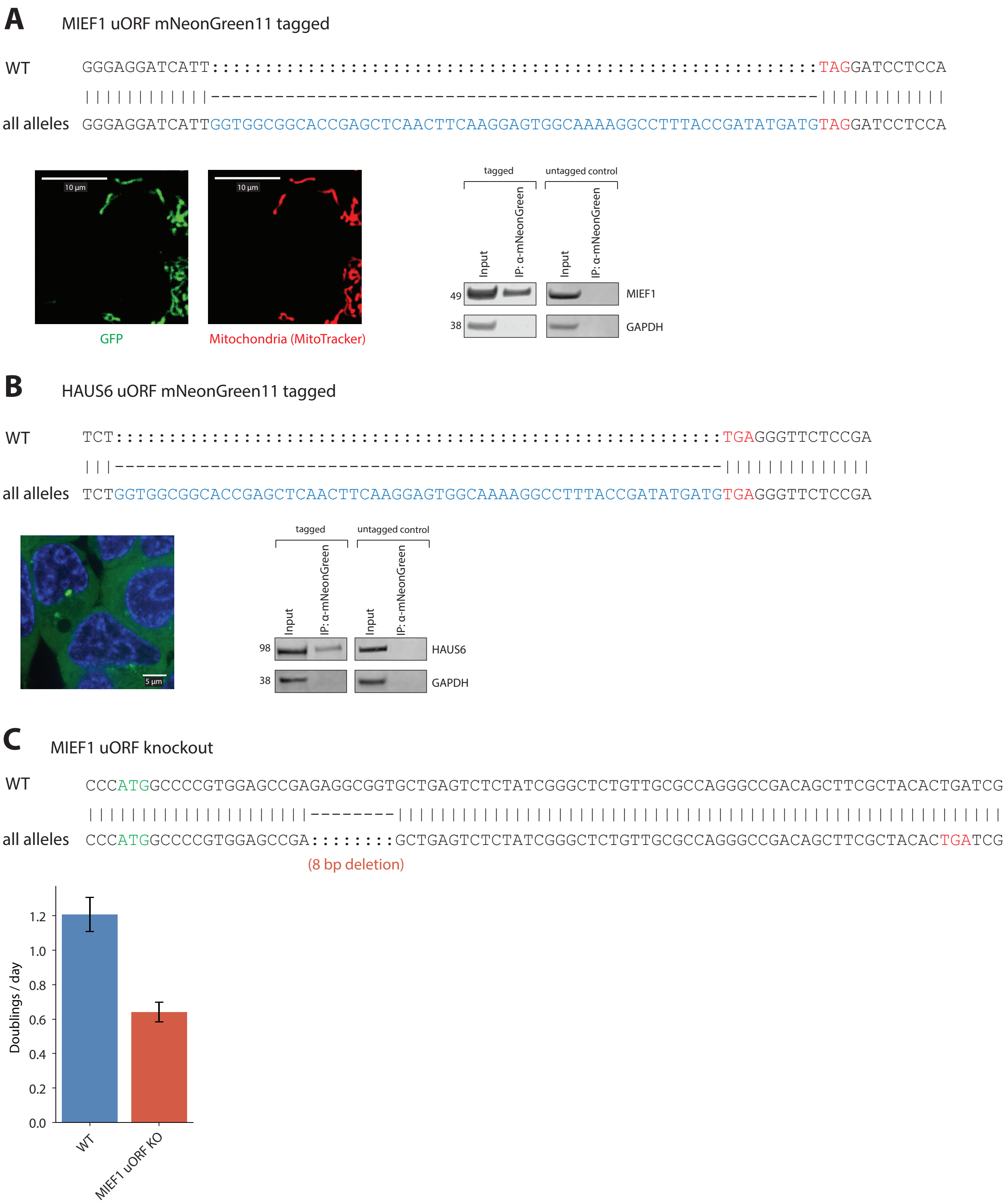


Fig. S15. Validation with endogenously tagged and knockout clonal cell lines.

(**A**) Sequence alignments around the knock-in sites of wild-type (WT) and MIEF1 uORF tagged (with mNeonGreen11, mNG11) clonal cell lines. Red is the stop codon. The data were obtained from PCR cloning followed by Illumina sequencing. Bottom: microscopy images showing the MIEF1 uORF peptide localizes to the mitochondria, and co-IP Western blots showing endogenous interaction of the MIEF1 uORF peptide with the MIEF1 protein. (**B**) Sequence alignments around the knock-in sites of wild-type (WT) and HAUS6 uORF tagged clonal cell lines, similar to (A). Bottom: microscopy image showing the HAUS6 uORF peptide localizes to the centrosome and cytoplasm, and co-IP Western blots showing endogenous interaction of the HAUS6 uORF peptide with the HAUS6 protein. (**C**) Sequence alignments around the edited sgRNA targeting sites of WT and knockout clonal cell line, with an 8 bp deletion. The 8 bp deletion creates a new in-frame stop codon, indicated in red. Green is the start codon. The data were obtained from PCR cloning followed by Illumina sequencing. Bottom: the knockout line has a slower doubling rate, recapitulating the growth defect phenotype from the pooled CRISPR knockout.

Supplementary Tables

Table S1. High-confidence translated ORFs identified by ORF-RATER in iPSCs and HFFs.

Table S2. Amino acid sequence of identified ORFs from ribosome profiling (all scores).

Table S3. Library composition of the CRISPR ORF sgRNA library.

Table S4. sgRNA read counts and growth phenotypes for K562 and iPSC CRISPR screens.

Table S5. ORF growth phenotypes and p-values for K562 and iPSC CRISPR screens.

Table S6. Comparison of ORF growth phenotypes in this study with published CRISPRi and CRISPRn datasets.

Table S7. sgRNA library composition of Perturb-Seq screen and Perturb-Seq results.

Table S8. sgRNA and construct sequences used for validation experiments.

References

1. M. A. Basrai, P. Hieter, J. D. Boeke, Small open reading frames: beautiful needles in the haystack. *Genome research* **7**, 768-771 (1997).
2. A. Odermatt, P. E. Taschner, S. W. Scherer, B. Beatty, V. K. Khanna, D. R. Cornblath, V. Chaudhry, W. C. Yee, B. Schrank, G. Karpati, M. H. Breuning, N. Knoers, D. H. MacLennan, Characterization of the gene encoding human sarcolipin (SLN), a proteolipid associated with SERCA1: absence of structural mutations in five patients with Brody disease. *Genomics* **45**, 541-553 (1997).
3. D. H. MacLennan, E. G. Kranias, Phospholamban: a crucial regulator of cardiac contractility. *Nat Rev Mol Cell Biol* **4**, 566-577 (2003).
4. S. R. Hann, M. W. King, D. L. Bentley, C. W. Anderson, R. N. Eisenman, A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. *Cell* **52**, 185-195 (1988).
5. R. Jackson, L. Kroehling, A. Khitun, W. Bailis, A. Jarret, A. G. York, O. M. Khan, J. R. Brewer, M. H. Skadow, C. Duizer, C. C. D. Harman, L. Chang, P. Bielecki, A. G. Solis, H. R. Steach, S. Slavoff, R. A. Flavell, The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434-438 (2018).
6. T. Kondo, S. Plaza, J. Zanet, E. Benrabah, P. Valenti, Y. Hashimoto, S. Kobayashi, F. Payre, Y. Kageyama, Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science* **329**, 336-339 (2010).
7. B. R. Nelson, C. A. Makarewich, D. M. Anderson, B. R. Winders, C. D. Troupes, F. Wu, A. L. Reese, J. R. McAnally, X. Chen, E. T. Kavalali, S. C. Cannon, S. R. Houser, R. Bassel-Duby, E. N. Olson, A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271-275 (2016).
8. D. M. Anderson, K. M. Anderson, C. L. Chang, C. A. Makarewich, B. R. Nelson, J. R. McAnally, P. Kasaragod, J. M. Shelton, J. Liou, R. Bassel-Duby, E. N. Olson, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160**, 595-606 (2015).
9. E. G. Magny, J. I. Pueyo, F. M. Pearl, M. A. Cespedes, J. E. Niven, S. A. Bishop, J. P. Couso, Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341**, 1116-1120 (2013).
10. N. G. D'Lima, J. Ma, L. Winkler, Q. Chu, K. H. Loh, E. O. Corpuz, B. A. Budnik, J. Lykke-Andersen, A. Saghatelian, S. A. Slavoff, A human microprotein that interacts with the mRNA decapping complex. *Nature chemical biology* **13**, 174-180 (2017).
11. C. S. Stein, P. Jadiya, X. Zhang, J. M. McLendon, G. M. Abouassaly, N. H. Witmer, E. J. Anderson, J. W. Elrod, R. L. Boudreau, Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell reports* **23**, 3710-3720 e3718 (2018).
12. C. A. Makarewich, K. K. Baskin, A. Z. Munir, S. Bezprozvannaya, G. Sharma, C. Khemtong, A. M. Shah, J. R. McAnally, C. R. Malloy, L. I. Szweda, R. Bassel-Duby, E. N. Olson, MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid beta-Oxidation. *Cell reports* **23**, 3701-3709 (2018).
13. A. Matsumoto, A. Pasut, M. Matsumoto, R. Yamashita, J. Fung, E. Monteleone, A. Saghatelian, K. I. Nakayama, J. G. Clohessy, P. P. Pandolfi, mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541**, 228-232 (2017).

14. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *The Journal of biological chemistry* **289**, 10950-10957 (2014).
15. P. Bi, A. Ramirez-Martinez, H. Li, J. Cannavino, J. R. McAnally, J. M. Shelton, E. Sanchez-Ortiz, R. Bassel-Duby, E. N. Olson, Control of muscle formation by the fusogenic micropeptide myomixer. *Science* **356**, 323-327 (2017).
16. J. Z. Huang, M. Chen, Chen, X. C. Gao, S. Zhu, H. Huang, M. Hu, H. Zhu, G. R. Yan, A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Molecular cell* **68**, 171-184 e176 (2017).
17. Q. Zhang, A. A. Vashisht, J. O'Rourke, S. Y. Corbel, R. Moran, A. Romero, L. Miraglia, J. Zhang, E. Durrant, C. Schmedt, S. C. Sampath, The microprotein Minion controls cell fusion and muscle formation. *Nature communications* **8**, 15664 (2017).
18. A. Pauli, M. L. Norris, E. Valen, G. L. Chew, J. A. Gagnon, S. Zimmerman, A. Mitchell, J. Ma, J. Dubrulle, D. Reyon, S. Q. Tsai, J. K. Joung, A. Saghatelian, A. F. Schier, Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343**, 1248636 (2014).
19. T. G. Johnstone, A. A. Bazzini, A. J. Giraldez, Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO journal* **35**, 706-723 (2016).
20. G. L. Chew, A. Pauli, A. F. Schier, Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nature communications* **7**, 11663 (2016).
21. S. R. Starck, J. C. Tsai, K. Chen, M. Shodiya, L. Wang, K. Yahiro, M. Martins-Green, N. Shastri, P. Walter, Translation from the 5' untranslated region shapes the integrated stress response. *Science* **351**, aad3867 (2016).
22. N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. Talhouarne, S. E. Jackson, M. R. Wills, J. S. Weissman, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports* **8**, 1365-1379 (2014).
23. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
24. S. A. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, A. D. Karger, B. A. Budnik, J. L. Rinn, A. Saghatelian, Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature chemical biology* **9**, 59-64 (2013).
25. A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, A. J. Giraldez, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981-993 (2014).
26. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).
27. A. P. Fields, E. H. Rodriguez, M. Jovanovic, N. Stern-Ginossar, B. J. Haas, P. Mertins, R. Raychowdhury, N. Hacohen, S. A. Carr, N. T. Ingolia, A. Regev, J. S. Weissman, A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Molecular cell* **60**, 816-827 (2015).
28. N. Stern-Ginossar, B. Weisburd, A. Michalski, V. T. Le, M. Y. Hein, S. X. Huang, M. Ma, B. Shen, S. B. Qian, H. Hengel, M. Mann, N. T. Ingolia, J. S. Weissman, Decoding human cytomegalovirus. *Science* **338**, 1088-1093 (2012).

29. K. M. Vattem, R. C. Wek, Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11269-11274 (2004).
30. M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, M. Mann, Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics* **14**, 658-673 (2015).
31. L. A. Gilbert, M. A. Horlbeck, B. Adamson, J. E. Villalta, Y. Chen, E. H. Whitehead, C. Guimaraes, B. Panning, H. L. Ploegh, M. C. Bassik, L. S. Qi, M. Kampmann, J. S. Weissman, Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647-661 (2014).
32. A. R. Perez, Y. Pritykin, J. A. Vidigal, S. Chhangawala, L. Zamparo, C. S. Leslie, A. Ventura, GuideScan software for improved single and paired CRISPR guide RNA design. *Nature biotechnology* **35**, 347-349 (2017).
33. M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-282 (2011).
34. B. Adamson, T. M. Norman, M. Jost, M. Y. Cho, J. K. Nunez, Y. Chen, J. E. Villalta, L. A. Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A. Regev, J. S. Weissman, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882 e1821 (2016).
35. P. Datlinger, A. F. Rendeiro, C. Schmidl, T. Krausgruber, P. Traxler, J. Klughammer, L. C. Schuster, A. Kuchler, D. Alpar, C. Bock, Pooled CRISPR screening with single-cell transcriptome readout. *Nature methods* **14**, 297-301 (2017).
36. M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang, A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3501-3508 (2016).
37. S. Feng, S. Sekine, V. Pessino, H. Li, M. D. Leonetti, B. Huang, Improved split fluorescent proteins for endogenous protein labeling. *Nature communications* **8**, 370 (2017).
38. Z. Ji, R. Song, A. Regev, K. Struhl, Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).
39. S. Samandi, A. V. Roy, V. Delcourt, J. F. Lucier, J. Gagnon, M. C. Beaudoin, B. Vanderperre, M. A. Breton, J. Motard, J. F. Jacques, M. Brunelle, I. Gagnon-Arsenault, I. Fournier, A. Ouangraoua, D. J. Hunting, A. A. Cohen, C. R. Landry, M. S. Scott, X. Roucou, Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* **6**, (2017).
40. A. Rathore, Q. Chu, D. Tan, T. F. Martinez, C. J. Donaldson, J. K. Diedrich, J. R. Yates, 3rd, A. Saghatelian, MIEF1 Microprotein Regulates Mitochondrial Translation. *Biochemistry* **57**, 5564-5575 (2018).
41. V. Delcourt, M. Brunelle, A. V. Roy, J. F. Jacques, M. Salzet, I. Fournier, X. Roucou, 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 Proteomics* **17**, 2402-2411 (2018).
42. D. Bergeron, C. Lapointe, C. Bissonnette, G. Tremblay, J. Motard, X. Roucou, An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel

ataxin-1 interacting protein. *The Journal of biological chemistry* **288**, 21824-21835 (2013).

43. C. F. Lee, H. L. Lai, Y. C. Lee, C. L. Chien, Y. Chern, The A2A adenosine receptor is a dual coding gene: a novel mechanism of gene usage and signal transduction. *The Journal of biological chemistry* **289**, 1257-1270 (2014).
44. R. Yu, T. Liu, S. B. Jin, C. Ning, U. Lendahl, M. Nister, J. Zhao, MIEF1/2 function as adaptors to recruit Drp1 to mitochondria and regulate the association of Drp1 with Mff. *Scientific reports* **7**, 880 (2017).
45. A. Sendoel, J. G. Dunn, E. H. Rodriguez, S. Naik, N. C. Gomez, B. Hurwitz, J. Levorse, B. D. Dill, D. Schramek, H. Molina, J. S. Weissman, E. Fuchs, Translation from unconventional 5' start sites drives tumour initiation. *Nature* **541**, 494-499 (2017).
46. M. A. Mandegar, N. Huebsch, E. B. Frolov, E. Shin, A. Truong, M. P. Olvera, A. H. Chan, Y. Miyaoka, K. Holmes, C. I. Spencer, L. M. Judge, D. E. Gordon, T. V. Eskildsen, J. E. Villalta, M. A. Horlbeck, L. A. Gilbert, N. J. Krogan, S. P. Sheikh, J. S. Weissman, L. S. Qi, P. L. So, B. R. Conklin, CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. *Cell stem cell* **18**, 541-553 (2016).
47. X. Lian, C. Hsiao, G. Wilson, K. Zhu, L. B. Hazeltine, S. M. Azarin, K. K. Raval, J. Zhang, T. J. Kamp, S. P. Palecek, Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E1848-1857 (2012).
48. N. J. McGlincy, N. T. Ingolia, Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112-129 (2017).
49. L. F. Lareau, D. H. Hite, G. J. Hogan, P. O. Brown, Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* **3**, e01257 (2014).
50. M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, J. L. Rinn, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25**, 1915-1927 (2011).
51. M. K. Iyer, Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette, J. R. Prensner, J. R. Evans, S. Zhao, A. Poliakov, X. Cao, S. M. Dhanasekaran, Y. M. Wu, D. R. Robinson, D. G. Beer, F. Y. Feng, H. K. Iyer, A. M. Chinnaiyan, The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics* **47**, 199-208 (2015).
52. J. G. Dunn, J. S. Weissman, Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC genomics* **17**, 958 (2016).
53. N. A. Kulak, G. Pichler, I. Paron, N. Nagaraj, M. Mann, Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature methods* **11**, 319-324 (2014).
54. N. A. Kulak, P. E. Geyer, M. Mann, Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. *Mol Cell Proteomics* **16**, 694-705 (2017).
55. M. Andreatta, B. Alvarez, M. Nielsen, GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic acids research* **45**, W458-W463 (2017).
56. M. Andreatta, M. Nielsen, Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511-517 (2016).

57. H. Xu, T. Xiao, C. H. Chen, W. Li, C. A. Meyer, Q. Wu, D. Wu, L. Cong, F. Zhang, J. S. Liu, M. Brown, X. S. Liu, Sequence determinants of improved CRISPR sgRNA design. *Genome research* **25**, 1147-1157 (2015).

58. J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten, D. E. Root, Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature biotechnology* **34**, 184-191 (2016).

59. M. A. Horlbeck, L. A. Gilbert, J. E. Villalta, B. Adamson, R. A. Pak, Y. Chen, A. P. Fields, C. Y. Park, J. E. Corn, M. Kampmann, J. S. Weissman, Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* **5**, (2016).

60. S. J. Liu, M. A. Horlbeck, S. W. Cho, H. S. Birk, M. Malatesta, D. He, F. J. Attenello, J. E. Villalta, M. Y. Cho, Y. Chen, M. A. Mandegar, M. P. Olvera, L. A. Gilbert, B. R. Conklin, H. Y. Chang, J. S. Weissman, D. A. Lim, CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, (2017).

61. B. Linder, A. V. Grozhik, A. O. Olarerin-George, C. Meydan, C. E. Mason, S. R. Jaffrey, Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature methods* **12**, 767-772 (2015).

62. D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, Q. Morris, The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research* **38**, W214-220 (2010).

63. M. M. Chan, Z. D. Smith, S. Grosswendt, H. Kretzmer, T. M. Norman, B. Adamson, M. Jost, J. J. Quinn, D. Yang, M. G. Jones, A. Khodaverdian, N. Yosef, A. Meissner, J. S. Weissman, Molecular recording of mammalian embryogenesis. *Nature*, (2019).

64. M. Jost, Y. Chen, L. A. Gilbert, M. A. Horlbeck, L. Krenning, G. Menchon, A. Rai, M. Y. Cho, J. J. Stern, A. E. Prota, M. Kampmann, A. Akhmanova, M. O. Steinmetz, M. E. Tanenbaum, J. S. Weissman, Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent. *Molecular cell* **68**, 210-223 e216 (2017).

65. C. S. Palmer, L. D. Osellame, D. Laine, O. S. Koutsopoulos, A. E. Frazier, M. T. Ryan, MiD49 and MiD51, new components of the mitochondrial fission machinery. *EMBO reports* **12**, 565-573 (2011).

66. M. C. Bassik, M. Kampmann, R. J. Lebbink, S. Wang, M. Y. Hein, I. Poser, J. Weibezahn, M. A. Horlbeck, S. Chen, M. Mann, A. A. Hyman, E. M. Leproust, M. T. McManus, J. S. Weissman, A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell* **152**, 909-922 (2013).

67. D. Mas-Ponte, J. Carlevaro-Fita, E. Palumbo, T. Hermoso Pulido, R. Guigo, R. Johnson, LncATLAS database for subcellular localization of long noncoding RNAs. *RNA* **23**, 1080-1087 (2017).

68. S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301-2319 (2016).

69. D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-1735 (2012).

70. D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov, S. Vajda, The ClusPro web server for protein-protein docking. *Nat Protoc* **12**, 255-278 (2017).