MOLECULAR BIOLOGY

Pervasive functional translation of noncanonical human open reading frames

Jin Chen[1,2], Andreas-David Brunner[3], J. Zachery Cogan[1,2], James K. Nuñez[1,2], Alexander P. Fields[1,2]*, Britt Adamson[1,2]†, Daniel N. Itzhak[4], Jason Y. Li[4], Matthias Mann[3,5], Manuel D. Leonetti[4], Jonathan S. Weissman[1,2]‡

Ribosome profiling has revealed pervasive but largely uncharacterized translation outside of canonical coding sequences (CDSs). In this work, we exploit a systematic CRISPR-based screening strategy to identify hundreds of noncanonical CDSs that are essential for cellular growth and whose disruption elicits specific, robust transcriptomic and phenotypic changes in human cells. Functional characterization of the encoded microproteins reveals distinct cellular localizations, specific protein binding partners, and hundreds of microproteins that are presented by the human leukocyte antigen system. We find multiple microproteins encoded in upstream open reading frames, which form stable complexes with the main, canonical protein encoded on the same messenger RNA, thereby revealing the use of functional bicistronic operons in mammals. Together, our results point to a family of functional human microproteins that play critical and diverse cellular roles.

Efforts to bioinformatically discover and annotate protein-coding open reading frames (ORFs) in genomes, termed coding sequences (CDSs), have traditionally relied on rules such as amino acid conservation and homology, translation initiation from an AUG start codon, and minimum length (i.e., 100 amino acids) (*1*). These rules have been widely adopted on the basis of the assumption that short peptides are unlikely to fold into stable structures to perform functions. However, the generality of these rules has been challenged. For example, the ribosomal protein RPL41 is a 25–amino acid (aa) peptide and both sarcolipin (SLN, 31 aa) and phospholamban (PLN, 52 aa) bind to and regulate the sarcoplasmic Ca^{2+} transporter SERCA (*2*, *3*). Additionally, MYC can be translated from a noncanonical start codon CUG (*4*), which demonstrates that non-AUG initiation can produce functional proteins. Recent studies have added a handful of examples of short proteins, or microproteins (also called micropeptides or just peptides), performing diverse functions (*5*–*18*), some encoded on transcripts annotated as long noncoding RNAs (lncRNAs). Finally, upstream ORFs (uORFs), located in the 5′ untranslated regions of mRNAs, have long been implicated in cis-acting translational control of the main, canonical CDS (*19*–*21*), though it has remained unclear whether they can generate stable, functional peptides.

Systematic identification of functional short CDSs remains challenging. Recent ribosome profiling (deep sequencing of ribosome-protected fragments) and mass spectrometry (MS) studies have identified thousands of previously unannotated CDSs (*22*–*25*) across bacteria, yeasts, viruses, and mammalian cells. However, for most cases, the cellular functions of these identified CDSs or their peptide products remain unexplored. We reasoned that the advent of CRISPR and its ability to precisely disrupt protein-coding regions (*26*), when combined with ribosome profiling, provides an opportunity to define and empirically characterize the functional protein-coding capacity of a given genome. In this work, we applied various types of approaches—including ribosome profiling, MS, and multiple CRISPR-based techniques—to systematically discover noncanonical CDSs encoded in the human genome and validate their critical roles in diverse cellular pathways.

To annotate potential CDSs comprehensively and accurately, we first investigated genome-wide translation by ribosome profiling across multiple cell types and conditions, including human induced pluripotent stem cells (iPSCs), iPSC-derived cardiomyocytes, human foreskin fibroblasts (HFFs), and HFFs infected with cytomegalovirus (*27*, *28*) (fig. S1A). We leveraged the ORF-RATER algorithm to annotate ORFs (*27*), incorporating multiple lines of evidence to identify ORFs undergoing active translation. This included consideration of the accumulation of ribosome densities at the start and stop codons, three-nucleotide periodicity, and additional experimental results, such as data from harringtonine-treated cells in which ribosomes are stalled at initiation sites (*27*). In iPSCs and cardiomyocytes, in addition to 9490 annotated CDSs (62% of the identified CDSs), we identified 3455 distinct, noncanonical CDSs (22%, i.e., with no in-frame overlap with previously annotated CDSs) and 2466 variant CDSs of annotated proteins (16%) in our high–statistical confidence set (Fig. 1A and materials and methods) (*27*). Among the distinct CDSs, 818 were CDSs on transcripts lacking prior protein-coding annotations ("new", i.e., lncRNAs), 2342 were upstream CDSs (i.e., uORFs or start overlaps: CDSs that overlap annotated start codons in a different reading frame), and only 13 were downstream CDSs. Similar numbers of CDSs were present in HFFs (fig. S1B), with 75% of the CDSs shared between the two cell types. Of the distinct CDSs, 96% are less than 100 aa in length, and 36% of the CDSs use non-AUG start codons (Fig. 1, B and C; see also fig. S2 for further characterizations).

Multiple lines of evidence suggest that the noncanonical CDSs are actively translated. The average ribosome density (metagene) of the lncRNA CDSs and of the translated uORFs closely mirrors footprints from that of annotated coding regions with strong three-nucleotide periodicities, a hallmark of active translation, as exemplified by traces from the lncRNA *LINC00998* transcript and a uORF of *ARL5A* (Fig. 1, D and E, and fig. S3). Our analysis also successfully recapitulated well-characterized short ORFs, such as the uORF on *ATF4* (*29*) and the recently discovered lncRNA-encoded microproteins MOXI/mitoregulin (*11*, *12*) and NoBody (*10*). Bona fide lncRNAs, such as *XIST*, *HOTAIR*, and *NEAT1*, were not identified to be protein coding (fig. S3E). Moreover, many of the CDSs were differentially translated during iPSC differentiation or viral infection (fig. S3F), providing evidence for translational control in different cell states.

MS-based proteomics in iPSCs and major human leukocyte antigen class I (HLA-I) peptidomics confirmed the stable expression of hundreds of noncanonical CDS peptides (Fig. 1F and figs. S4 and S5). HLA-I peptidomics identified 240 noncanonical peptides, which suggests that these peptides enter the HLA-I presentation pathway and contribute to the antigen repertoire and possible immunogenicity (Fig. 1F) (*30*). HLA-I prediction analysis cross-validated strong binding ($K_d \leq 50$ mM, where K_d is the dissociation constant) of noncanonical CDS HLA-I peptides to their respective allotypes (fig. S6) (*30*). MS-based proteomics using tryptic digestion identified far fewer noncanonical peptides, which may be due to challenges in detecting the trypsin-digested products from short, noncanonical CDSs or possibly to more rapid turnover of these noncanonical peptides (fig. S7).

To test whether translation of the noncanonical CDSs is important for cell growth and potentially yields functional peptides, we measured the growth phenotypes resulting from CRISPR-mediated ORF knockout in

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94158, USA. [2]Howard Hughes Medical Institute, University of California, San Francisco, CA 94158, USA. [3]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany. [4]Cell Atlas Initiative, Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [5]Clinical Proteomics Group, Proteomics Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen 2200, Denmark.
*Present address: GRAIL, Inc., Menlo Park, CA 94025, USA.
†Present address: Department of Molecular Biology and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA.
‡Corresponding author. Email: jonathan.weissman@ucsf.edu

pooled screens (*26*). We designed a Cas9 ORF single guide RNA (sgRNA) library to specifically knock out thousands of the noncanonical CDSs identified by ribosome profiling (Fig. 2A and materials and methods) (*31*, *32*), targeting 1098 uORFs, 613 lncRNA CDSs, 352 extensions of annotated coding regions, 283 start overlaps, and 7 downstream CDSs. We performed pooled Cas9 knockout screens in iPSC and K562 chronic myeloid leukemia cells expressing Cas9 and the sgRNA library, akin to conventional pooled screens for essential proteins (*26*, *31*). We measured sgRNA abundance in the cell populations shortly after library transduction and after 10 additional population doublings by deep sequencing to quantify the fitness defect conferred by each sgRNA. We then calculated a phenotype score (γ) and

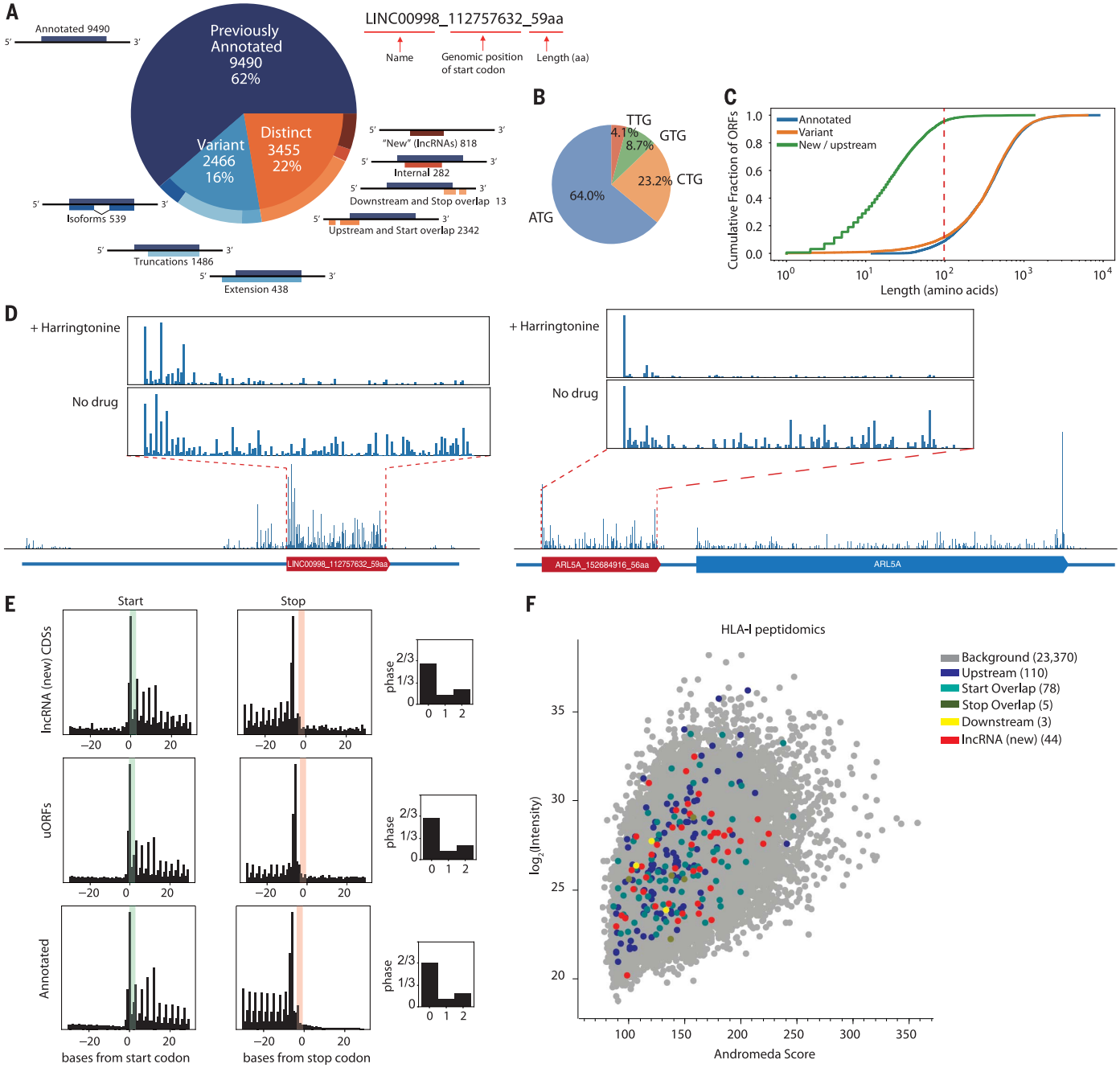
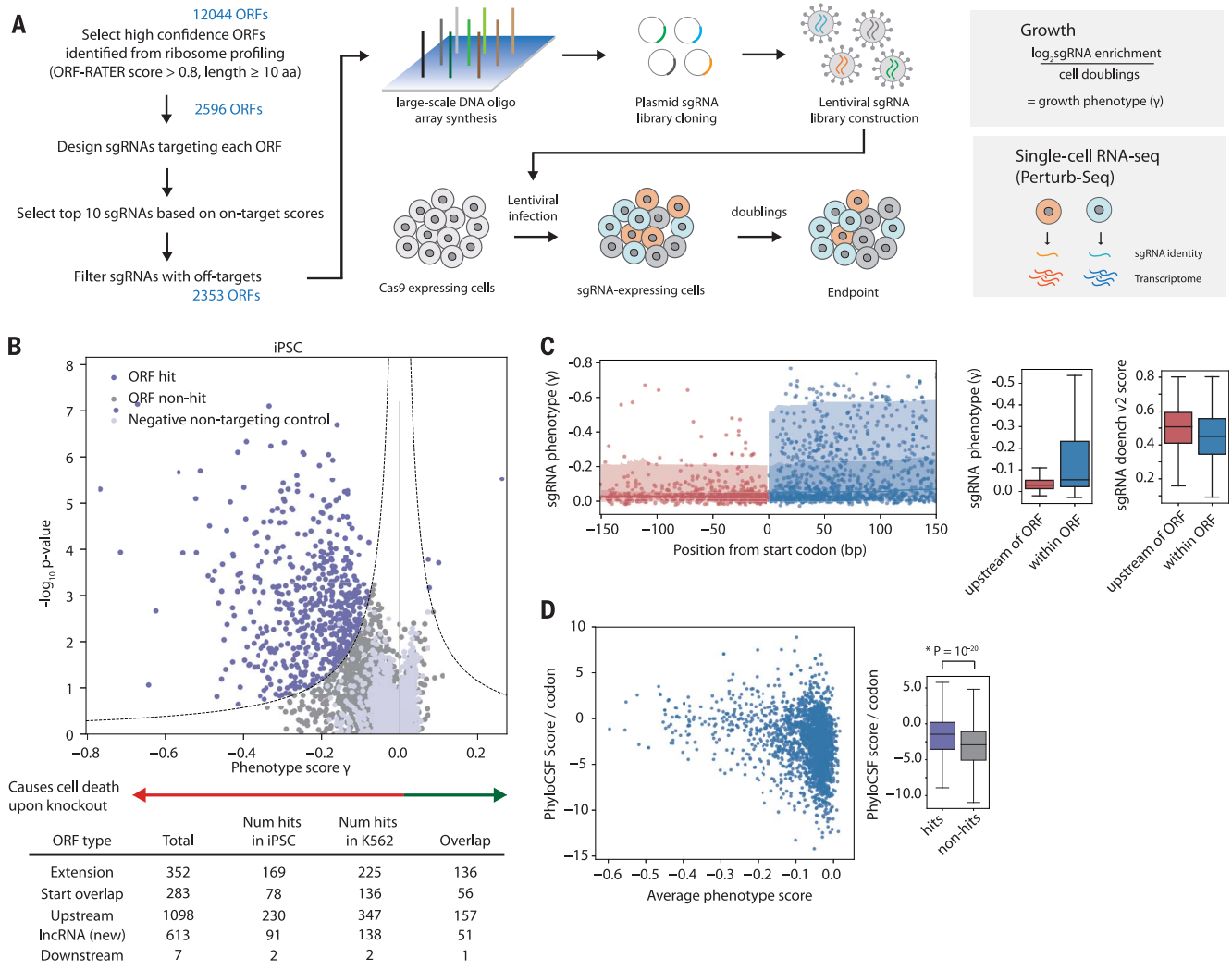


Fig. 1. Ribosome profiling and MS reveal translation of unannotated CDSs. (**A**) ORF-RATER analysis of ribosome profiling data: 62% are previously annotated coding sequences, whereas 16% are variants of canonical coding sequences that share portions of the coding sequence and 22% are distinct from annotated coding sequences. The naming convention of the identified ORFs is shown on the right. (**B**) Start-codon usage of the identified CDSs. (**C**) Cumulative distribution of CDS length. For distinct CDSs, 96% are smaller than 100 amino acids. (**D**) Example ribosome profiling traces of a lncRNA peptide from *LINC00998* and a uORF peptide from *ARL5A* displaying the hallmarks of translation, including peaks of density around the start codon following harringtonine treatment and three-nucleotide periodicities along the coding region. (**E**) Metagene analysis shows that the signatures of translation, including three-nucleotide periodicity in the expected reading frame, for uORFs and lncRNA CDSs are similar to those for annotated coding regions. (**F**) Identification of >200 noncanonical CDS peptides from HLA-I peptidomics, cross-validating their existence across the whole abundance range, with a mean Andromeda score of 141 compared with a total mean Andromeda score of 144. See materials and methods for further details.



ORF type	Total	Num hits in iPSC	Num hits in K562	Overlap
Extension	352	169	225	136
Start overlap	283	78	136	56
Upstream	1098	230	347	157
lncRNA (new)	613	91	138	51
Downstream	7	2	2	1

Fig. 2. Genome-scale CRISPR screens to identify functional, non-canonical CDSs. (**A**) Schematic of CRISPR library design and screening strategies, either by growth screens or Perturb-seq. For growth screens, frequencies of cells expressing a given sgRNA are determined by next-generation sequencing, and phenotype scores are quantified with the formula shown. For Perturb-seq, single-cell transcriptomes and sgRNA identities were obtained by single-cell RNA-seq. (**B**) Volcano plot summarizing knockout phenotypes and statistical significance (determined by Mann-Whitney *U* test) for ORFs targeted in the pooled screen in iPSCs. Each dot represents a targeted ORF, and ORF hits are labeled in purple, with a more negative phenotype score indicating a stronger growth defect. See materials and methods for further details. (**C**) Plot of the sgRNA phenotypes and distance from the start codon across all ORF hits. sgRNAs targeting the genome immediately upstream of the ORF (shown in red) have significantly lower phenotype scores than sgRNAs targeting within the ORF (shown in blue). Note the axis is increasingly negative (stronger) phenotype. The sgRNA phenotypes are quantified by the boxplot to the right. The difference is not because of differences in sgRNA on-target efficiencies, as quantified by the Doench v2 score. (**D**) The PhyloCSF score per codon (higher scores are more conserved across the Euarchontoglires) is generally higher for ORF hits (*$P = 10^{-20}$, Kolmogorov-Smirnov test) and ORFs with a stronger phenotype. Note that lack of a growth phenotype does not necessarily imply a low PhyloCSF score.

confidence (*P* value) for each ORF from the relative enrichment or depletion of sgRNAs targeting a particular ORF (Fig. 2B and materials and methods). In iPSCs, our screen identified >500 ORF knockout hits that resulted in statistically significant phenotypes. The hits include 169 genes that are variants of annotated proteins, 78 start overlap hits, 230 uORF hits, 91 lncRNA CDS hits, and 2 downstream CDS hits. iPSC and K562 cells had 401 shared hits, suggesting housekeeping or general cellular roles as well as CDSs that may play cell-specific functions (fig. S8).

A fraction of the uORF hits do not have main, canonical CDSs with fitness defects upon knockout. This suggests an independent function of the uORFs or that disruption of the uORFs leads to increases in main CDS expression, which results in the growth phenotype (fig. S8E). Thus, unannotated CDSs with important functions across multiple cell types are an abundant feature of the genome.

Several lines of evidence further suggested that our screen reported specifically on the phenotypes of the selected ORFs. First, the phenotypes of control sgRNAs targeted directly upstream of each ORF in the genome (Fig. 2C) are significantly weaker than those of sgRNAs targeted within the ORF ($P = 10^{-26}$, Mann-Whitney test). Second, sgRNA phenotypes are independent of distance to other annotated proteins, splice sites, or transcriptional start sites (fig. S9A). Functionally, ORF hits are, on average, more phylogenetically conserved with a higher conservation score than non-hits (PhyloCSF score per codon, $P = 10^{-20}$, Kolmogorov-Smirnov test; Fig. 2D) (*33*), and they have other distinguishing sequence features (e.g., enrichment for Kozak consensus

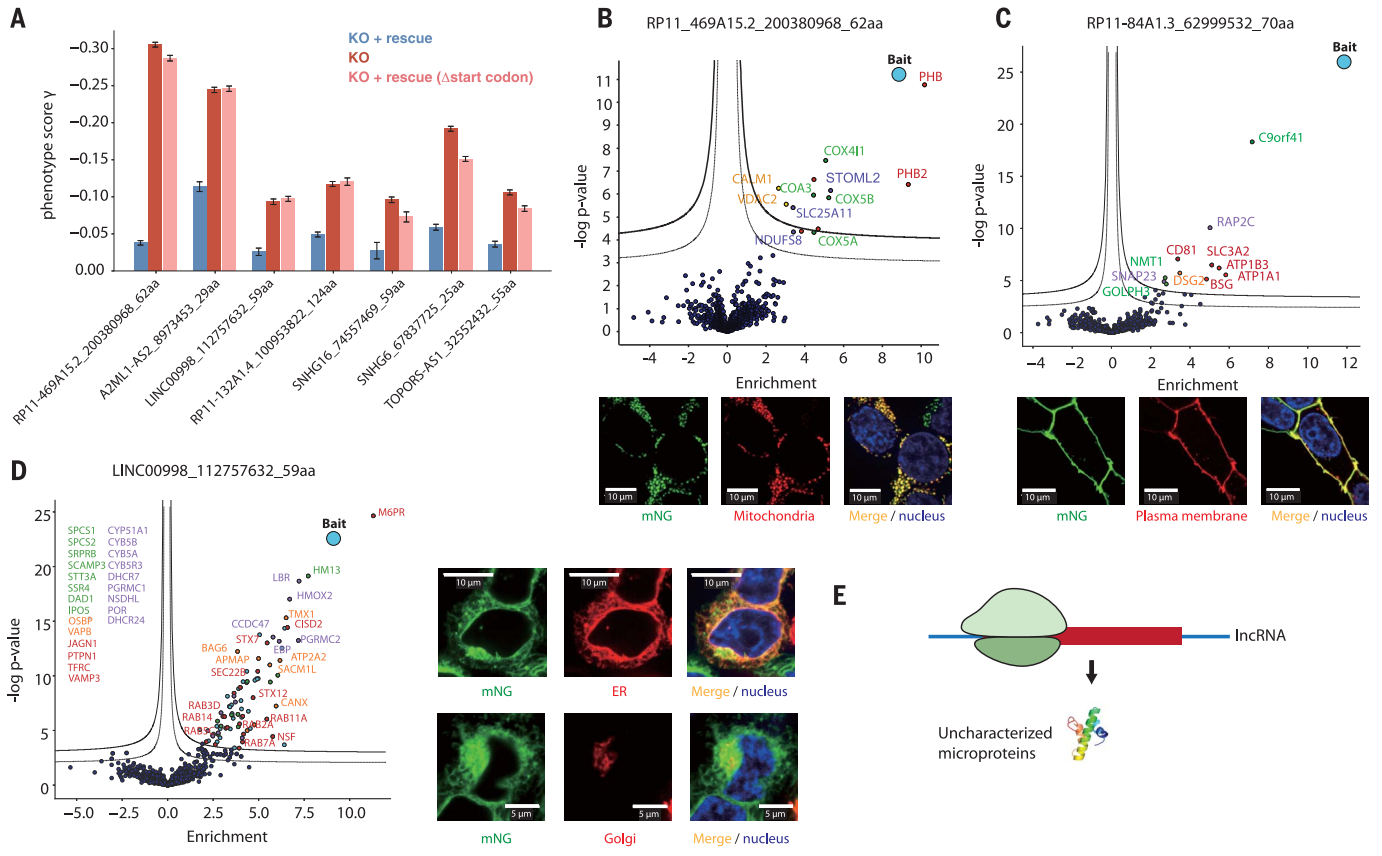
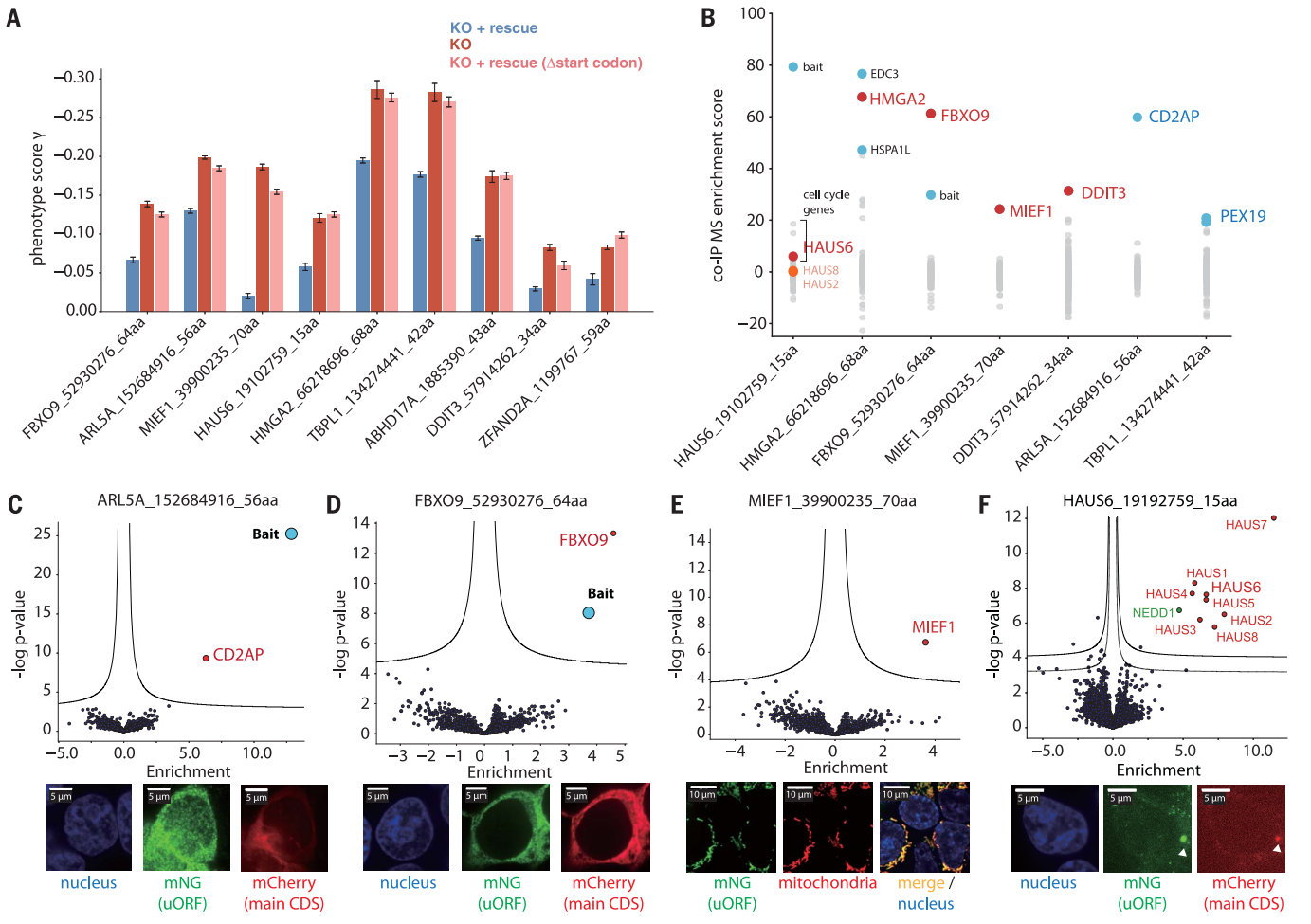


Fig. 3. Short lncRNA CDSs encode functional microproteins. (**A**) Rescue of lncRNA CDS knockout growth phenotypes by ectopic expression of the transcript encoding the peptide, as well as controls in which the initiating start codon is removed (Δstart codon). Error bars represent standard deviation of triplicates. $P < 0.05$ for all comparisons between knockout (KO) and KO + rescue. (**B** to **D**) Microscopy images and volcano plots of the co-IP MS of three example lncRNA-encoded microproteins tagged with mNG11, expressed ectopically (in the native transcript context) in a HEK293T cell line expressing mNG1-10. Green is mNG, red is the indicated organelle localization, and blue is Hoechst 33342, which stains for the nucleus. Scale bar dimensions are labeled. Statistically significant interactors are shown in the top, right corner of the volcano plots. Thick threshold line is 1% FDR (false discovery rate), and the thin threshold line is 5% FDR. The bait (the tagged peptide) is labeled in blue. The interactors are colored according to their functional groups. (**E**) lncRNA-encoded microproteins are uncharacterized proteins that may play important regulatory roles in cells.

sequence) (fig. S10). However, the noncanonical CDSs, on average, have lower PhyloCSF scores compared with canonical proteins (fig. S2B). Finally, sgRNAs targeting ORF hits versus non-hits have indistinguishable off-target and on-target scores (fig. S9B) (*32*). We then performed validation follow-ups with individual sgRNAs, which recapitulated the growth phenotypes from our genome-scale screen (fig. S8D). Sequencing of the targeted genomic regions revealed insertions and deletions (indels) of <50 base pairs (bp) (fig. S9, C and D). Together, these analyses independently support the conclusion that our screen phenotypes result specifically from the disruption of the target ORFs.

To survey function of the noncanonical CDSs at scale, we combined CRISPR screening with single-cell RNA sequencing (Perturb-seq) (*34*, *35*). Disruptions of the various noncanonical CDSs resulted in broad and diverse changes in RNA-sequencing profiles across a variety of critical pathways, suggesting that the candidate CDSs play diverse cellular roles (fig. S11). As an example, disruption of the CDS on *LINC00998* resulted in differentially expressed genes related to glycosylation ($P < 10^{-10}$), suggesting a function at the Golgi or endoplasmic reticulum (ER). The transcriptional phenotype also allowed us to functionally profile CDSs that are not essential for robust growth (fig. S11C). Furthermore, we found that CRISPR-targeted transcripts did not show detectable changes in abundance that might result from processes such as nonsense-mediated decay, which indicates that the phenotypes we observed were not caused by decreasing the abundance of the entire transcript (fig. S11D). Thus, similar to screens for essential protein-coding genes (*26*, *31*), our screen for noncanonical CDSs required for robust cell growth underestimated the true number of functional CDSs in the genome. This finding further underscores the pervasiveness of functional, unannotated CDSs in the genome that affect a wide range of cellular activities.

We next explored the functional role of the peptides encoded by the noncanonical CDSs identified from our screen, focusing first on lncRNA CDSs. For seven lncRNAs, we ectopically expressed the transcript encoding for the peptide and found, in all cases, that knockout-induced growth defect was partially or completely rescued. This rescue was abrogated by the removal of the initiating start codon (Δstart codon) (Fig. 3A), which suggests an essential role of the peptide itself in cell growth. To further investigate the specific functions of the noncanonical microproteins, we adopted a split-fluorescent protein approach using mNeonGreen (mNG), in which we fused each peptide with a minimally disruptive 16-aa tag (mNG11). Coexpression of the tagged peptide with the remainder of the mNG protein (mNG1-10) results in a fluorescence signal upon complementation (*36*, *37*). This creates



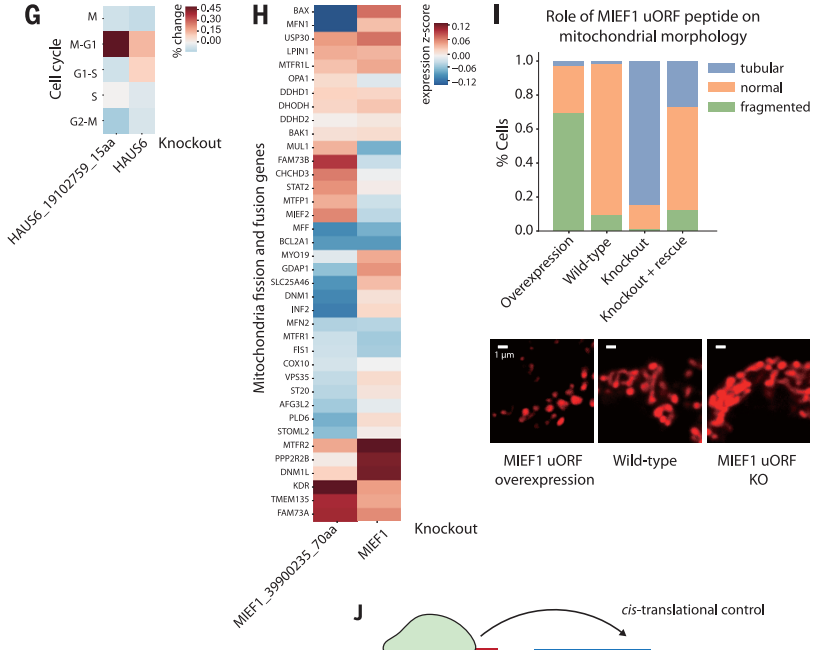


Fig. 4. Bicistronic mRNAs can encode uORF peptides that function in trans. (**A**) Rescue of uORF knockout growth phenotypes by ectopic expression of a transcript encoding the uORF peptide alone, as well as controls in which the initiating start codon is removed (Δstart codon). Error bars represent standard deviation of triplicates. $P < 0.05$ for all comparisons between KO and KO + rescue. (**B**) Summary of co-IP MS interactions of uORF peptides tagged with mNG11, showing five uORF peptides that interact with their downstream-encoded protein (shown in red). The tagged peptides are expressed alone ectopically (in the native transcript context) in an HEK293T cell line expressing mNG1-10. Other statistically significant interactors are shown in blue. (**C** to **E**) Example volcano plots of co-IP MS in (B) reveal statistically significant interactors with uORF peptides. Threshold line is 1% FDR. The bait (the tagged peptide) is labeled in blue. For microscopy in (C) and (D), the main, canonical protein tagged with mCherry (red) is coexpressed. For (E), the mNG11-tagged MIEF1 uORF peptide (green) localizes to the mitochondria (red). (**F**) Volcano plot of co-IP MS from endogenously mNG11-tagged HAUS6 uORF. For microscopy, the mNG11-tagged uORF is expressed alone ectopically (green), and the canonical HAUS6 tagged with mCherry (red) is coexpressed. (**G**) Percent change for each cell cycle state for HAUS6 knockout (KO) and HAUS6 uORF KO, compared with control cells. (**H**) Transcriptome response of the MIEF1 uORF KO compared with the main CDS KO from Perturb-seq. (**I**) Quantification of mitochondria morphology upon MIEF1 uORF peptide overexpression and knockout, as well as rescue of knockout phenotype. Representative microscopy images of the different mitochondria morphologies are shown to the right. (**J**) Possible model of uORF peptide functions and regulatory roles in cells.

both a fluorescent reporter to detect stable expression and cellular localization and a handle for coimmunoprecipitation (co-IP) and MS to define interaction partners (*36*) (fig. S12A). We probed the functions of six essential lncRNA CDSs and found that five of them formed specific complexes that were consistent with their subcellular localization. For example, the 62-aa peptide encoded by lncRNA *RP11_469A15.2* specifically localized to the mitochondria. The peptide has a predicted transmembrane domain and coimmunoprecipitates with the cytochrome c oxidase (COX) complex and the mitochondrial Prohibitin complex (Fig. 3B). Moreover, the 70-aa peptide encoded by *RP11-84A1.3* localizes to the plasma membrane and interacts with various cell surface proteins (Fig. 3C). Third, the 59-aa peptide encoded by lncRNA *LINC00998*, which contains two predicted transmembrane domains, localizes specifically to both the ER and Golgi and coimmunoprecipitates with lysosomal and vesicular transport proteins (Fig. 3D). Finally, the 55-aa peptide encoded on *TOPORS-AS1* and the 124-aa peptide on *RP11-132A1.4* also form functional complexes consistent with their cellular localization (fig. S12, C and D, and fig. S13). Consistent with prior studies (*5*–*18*), these examples demonstrate that lncRNAs can encode uncharacterized proteins, and they highlight the need to fully extend the annotation of lncRNAs and the proteome.

We next explored the functional effects of uORF translation, which is complicated by the fact that phenotypes can, in principle, be mediated by the peptide product (*24*, *38*–*41*), the effect of uORF translation on expression of the main, canonical CDS (*20*), or both. To distinguish between these possibilities, we first separately tagged the uORF and the main CDS and used Western blot to confirm the independent expression of uORF peptides from the canonical protein (fig. S14). Furthermore, we established that ectopic expression of a transcript encoding only the uORF peptide could at least partially rescue the growth phenotype caused by disruption of the endogenous uORF. In all cases this rescue is dependent on the initiating start codon in the ectopically expressed message, which demonstrates that the rescue is the result of production of the expressed peptide (Fig. 4A). Consistent with this, in all cases we tested, deleting the start codon for the uORFs only minimally increased (~20% to 60%) the expression of the main CDS. This suggests that the growth defect observed is mediated by the peptide and not by increased expression of the canonical protein (fig. S14, E and F). Taken together, these findings establish that uORFs could function through the peptide they produce independently of any cis-regulatory effects.

To explore the functions of uORF-encoded microproteins, we examined their localization and protein binding partners by tagging the uORF peptides with mNG11. Out of the 10 uORF peptides further tested by co-IP MS, we failed to detect statistically significant interaction partners for three of the tagged peptides. Two peptides, encoded by the uORFs of *TBPL1* and *ARL5A*, localize generally to the cytoplasm, whereas the main CDS proteins exhibit different cellular localization patterns. Consistent with our observed cellular localizations, these two uORF peptides specifically immunoprecipitate proteins with functions that are independent of the main CDS protein (Fig. 4, B and C, and fig. S12). Thus, these uORF peptides and their main CDS protein have independent functions.

We found that 5 of the 10 uORF peptides colocalized and formed a stable physical complex with the downstream-encoded, canonical protein on their shared mRNA. These include *MIEF1*, *DDIT3*, *FBXO9*, *HMGA2*, and *HAUS6* (Fig. 4, B, D, and E, and fig. S12). In all cases, we expressed the tagged peptides in their native transcript context but without the downstream CDS, thereby eliminating the possibility of stop codon read-through. We further confirmed this interaction by co-IP of the canonical protein and immunoblotting for the uORF peptide (fig. S12F), as well as with endogenously tagged clonal lines (fig. S15 and Fig. 4F). This physical interaction between the proteins encoded by the uORF and the canonical CDS on the same transcript is notable (*39*, *42*, *43*) because it implies an additional layer of regulation beyond the propensity of uORFs to modulate translation of downstream CDSs.

Next, we further investigated the function of uORF-expressed microproteins in *HAUS6* and *MIEF1*. In both cases, disrupting the uORF led to minimal increase in the expression of the main CDS protein, and the ectopic expression of a peptide-encoding transcript rescued the knockout-induced growth phenotype (Fig. 4A and fig. S14). mNG11-tagged HAUS6 uORF expressed from its endogenous locus efficiently pulled down key components of the HAUS6 complex, localized to the centrosome, and knockout of the uORF caused cells to arrest in the G1 stage, consistent with the role of HAUS6 microtubule attachment to the kinetochore and central spindle formation (Fig. 4, F and G, fig. S12, and fig. S15). Similarly, the MIEF1 uORF peptide localized to the mitochondria, consistent with the localization of the MIEF1 protein (Fig. 4E), which regulates mitochondrial fission and fusion (*44*). The MIEF1 uORF peptide knockout induced differential expression of mitochondrial fusion and fission genes, with a transcriptional signature that was distinct from that seen in the knockout of the MIEF1 protein (Fig. 4H). We observed that overexpression of the MIEF1 uORF peptide alone induced a fragmented mitochondrial phenotype (increased fission), whereas a clonal knockout of the MIEF1 uORF (with the sequence disrupted but nonetheless preserving an upstream ORF; see fig. S15) resulted in a tubular and more elongated mitochondrial phenotype (increased fusion). Notably, this knockout morphology could be rescued by the exogenous expression of the MIEF1 uORF peptide (Fig. 4I). Together, our results indicated a possible role of the uORF-encoded peptide in regulating the downstream-encoded protein, thereby challenging the monocistronic assumption about mammalian genomes. We speculate that this type of genomic architecture may be general, opening the doors to investigation of the cooperative and regulatory nature of bicistronic human mRNAs. Indeed, a number of stress-regulated alternate translation initiation factors can modulate translation initiation site choice and uORF usage, which suggests that regulation of bicistronic expression could play roles in both normal biology and diseases states (*21*, *45*).

We described a strategy that combines ribosome profiling, MS-based proteomics, microscopy, and CRISPR-based genetic screens to discover and characterize widespread translation of functional microproteins and define the protein-coding potential of complex genomes. We identified a subset of lncRNAs that can encode stable, functional proteins, which suggests that they may be misannotated RNAs or potentially have dual roles at the RNA and protein levels. Furthermore, we provided examples of uORFs encoding functional peptides, highlighting the diverse cellular roles that uORFs may play beyond translational control. We also identified uORF-encoded peptides binding to the downstream-encoded protein on the same mRNA. Thus, our data highlight a previously unappreciated complexity of the functional mammalian proteome, as well as the full spectrum of antigens presented by the HLA system.

REFERENCES AND NOTES

1. M. A. Basrai, P. Hieter, J. D. Boeke, *Genome Res.* **7**, 768–771 (1997).
2. A. Odermatt *et al.*, *Genomics* **45**, 541–553 (1997).
3. D. H. MacLennan, E. G. Kranias, *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, *Cell* **52**, 185–195 (1988).
5. R. Jackson *et al.*, *Nature* **564**, 434–438 (2018).
6. T. Kondo *et al.*, *Science* **329**, 336–339 (2010).
7. B. R. Nelson *et al.*, *Science* **351**, 271–275 (2016).
8. D. M. Anderson *et al.*, *Cell* **160**, 595–606 (2015).
9. E. G. Magny *et al.*, *Science* **341**, 1116–1120 (2013).
10. N. G. D'Lima *et al.*, *Nat. Chem. Biol.* **13**, 174–180 (2017).
11. C. S. Stein *et al.*, *Cell Rep.* **23**, 3710–3720e8 (2018).
12. C. A. Makarewich *et al.*, *Cell Rep.* **23**, 3701–3709 (2018).
13. A. Matsumoto *et al.*, *Nature* **541**, 228–232 (2017).
14. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, *J. Biol. Chem.* **289**, 10950–10957 (2014).
15. P. Bi *et al.*, *Science* **356**, 323–327 (2017).
16. J.-Z. Huang *et al.*, *Mol. Cell* **68**, 171–184.e6 (2017).
17. Q. Zhang *et al.*, *Nat. Commun.* **8**, 15664 (2017).
18. A. Pauli *et al.*, *Science* **343**, 1248636 (2014).
19. T. G. Johnstone, A. A. Bazzini, A. J. Giraldez, *EMBO J.* **35**, 706–723 (2016).
20. G. L. Chew, A. Pauli, A. F. Schier, *Nat. Commun.* **7**, 11663 (2016).

21. S. R. Starck *et al*., *Science* **351**, aad3867 (2016).
22. N. T. Ingolia *et al*., *Cell Rep.* **8**, 1365–1379 (2014).
23. N. T. Ingolia, L. F. Lareau, J. S. Weissman, *Cell* **147**, 789–802 (2011).
24. S. A. Slavoff *et al*., *Nat. Chem. Biol.* **9**, 59–64 (2013).
25. A. A. Bazzini *et al*., *EMBO J.* **33**, 981–993 (2014).
26. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, *Science* **343**, 80–84 (2014).
27. A. P. Fields *et al*., *Mol. Cell* **60**, 816–827 (2015).
28. N. Stern-Ginossar *et al*., *Science* **338**, 1088–1093 (2012).
29. K. M. Vattem, R. C. Wek, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11269–11274 (2004).
30. M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, M. Mann, *Mol. Cell. Proteomics* **14**, 658–673 (2015).
31. L. A. Gilbert *et al*., *Cell* **159**, 647–661 (2014).
32. A. R. Perez *et al*., *Nat. Biotechnol.* **35**, 347–349 (2017).
33. M. F. Lin, I. Jungreis, M. Kellis, *Bioinformatics* **27**, i275–i282 (2011).
34. B. Adamson *et al*., *Cell* **167**, 1867–1882.e21 (2016).
35. P. Datlinger *et al*., *Nat. Methods* **14**, 297–301 (2017).
36. M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang, *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3501–E3508 (2016).
37. S. Feng *et al*., *Nat. Commun.* **8**, 370 (2017).
38. Z. Ji, R. Song, A. Regev, K. Struhl, *eLife* **4**, e08890 (2015).
39. S. Samandi *et al*., *eLife* **6**, e27860 (2017).
40. A. Rathore *et al*., *Biochemistry* **57**, 5564–5575 (2018).
41. V. Delcourt *et al*., *Mol. Cell. Proteomics* **17**, 2402–2411 (2018).
42. D. Bergeron *et al*., *J. Biol. Chem.* **288**, 21824–21835 (2013).
43. C. F. Lee, H. L. Lai, Y. C. Lee, C. L. Chien, Y. Chern, *J. Biol. Chem.* **289**, 1257–1270 (2014).
44. R. Yu *et al*., *Sci. Rep.* **7**, 880 (2017).
45. A. Sendoel *et al*., *Nature* **541**, 494–499 (2017).

ACKNOWLEDGMENTS

We thank the Weissman laboratory members, particularly T. M. Norman, M. Jost, M. J. Shurtleff, M. A. Horlbeck, L. A. Gilbert, M. Y. Hein, S. E. Torres, C. R. Liem, D. A. Santos, J. M. Replogle, and A. Xu. We also thank B. R. Conklin and M. P. Olvera (Gladstone Institute) for iPSC culturing; N. Cho (Chan-Zuckerberg Biohub) for endogenous tagging; and S. E. O'Leary and I W. Lin for discussions. **Funding:** This work was funded by NIH (RM1 HG009490), DARPA (HR0011-17-2-0043), and the Chan-Zuckerberg Initiative. J.S.W. is a HHMI Investigator. A.-D.B. and M.M. are supported by the Max Planck Society. J.C. is funded by the Jane Coffin Childs Memorial Fund for Medical Research and the NIH K99/R00 Pathway to Independence Award (K99 GM134154). J.K.N. is a fellow of the Hanna H. Gray Fellows Program. Oligonucleotide pools were courtesy of the Innovative Genomics Institute. **Author contributions:** J.C. designed and performed all experiments and analyzed and interpreted all the data with guidance from J.S.W. A.-D.B. and M.M. performed the proteomic MS and analysis. J.Z.C. and J.K.N. assisted with sample preparation, experiments, and key discussions. A.P.F. performed preliminary experiments and provided key analytical pipelines. B.A. contributed to Perturb-seq experiments. M.D.L. and J.Y.L. performed the pull-downs. M.D.L. designed the endogenous tagging methods. D.N.I. performed MS for the pull-downs. J.C. and J.S.W. conceived the study and wrote the manuscript with input from all authors. **Competing interests:** J.S.W. consults for and holds equity in KSQ Therapeutics and Maze Therapeutics and consults for 5AM Ventures. B.A. is an advisory board member and has restricted stock in Celsius Therapeutics, Inc. **Data and materials availability:** Raw sequencing data are deposited on the National Center for Biotechnology Information Gene Expression Omnibus database with accession number GSE131650. MS-based proteomics data are deposited to the ProteomeXchange Consortium via the Proteomics Identifications Database (PRIDE) partner repository with the dataset identifier PXD014031. Processed data are included as supplementary tables.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6482/1140/suppl/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S8
References (*46*–*70*)

View/request a protocol for this paper from *Bio-protocol*.

26 May 2019; resubmitted 22 November 2019
Accepted 13 January 2020
10.1126/science.aay0262

Science

Pervasive functional translation of noncanonical human 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 and Jonathan S. Weissman

Science **367** (6482), 1140-1146.
DOI: 10.1126/science.aay0262

Expanding the human proteome

Using mass spectrometry, ribosome profiling, and several CRISPR-based screens, Chen *et al.* identified hundreds of previously uncharacterized functional micropeptides in the human genome (see the Perspective by Wei and Guo). Protein translation outside of annotated open reading frames (ORFs) in messenger RNAs and within ORFs in long noncoding RNAs is pervasive. A functional screen using CRISPR-Cas9 with single-cell transcriptomics suggested critical roles for hundreds of micropeptides. Micropeptides encoded by multiple short, upstream ORFs form stable protein complexes with the downstream canonical proteins encoded on the same messenger RNAs.

Science, this issue p. 1140; see also p. 1074

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.