

# Peptidomic discovery of short open reading frame-encoded peptides in human cells

Published in final edited form as:
*Nat Chem Biol*. 2013 January ; 9(1): 59–64. doi:10.1038/nchembio.1120.

# Peptidomic discovery of short open reading frame-encoded peptides in human cells

**Sarah A. Slavoff**[1], **Andrew J. Mitchell**[2,*], **Adam G. Schwaid**[1,*], **Moran N. Cabili**[3,4,5], **Jiao Ma**[1], **Joshua Z. Levin**[6], **Amir D. Karger**[7], **Bogdan A. Budnik**[8], **John L. Rinn**[3,5], and **Alan Saghatelian**[1,†]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA

[2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

[3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

[4]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA

[5]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA

[6]Genome Sequencing & Analysis Program, Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA

[7]Research Computing, Division of Science, Faculty of Arts and Sciences, Harvard University, 38 Oxford St, Room 211A, Cambridge, Massachusetts 02138, USA

[8]Center of Systems Biology, Mass Spectrometry and Proteomics Lab, Faculty of Arts and Sciences, Harvard University, 52 Oxford St, Northwest Labs, B243.20, Cambridge, Massachusetts 02138, USA

## Abstract

The amount of the transcriptome that is translated into polypeptides is of fundamental importance. We developed a peptidomic strategy to detect short ORF (sORF)-encoded polypeptides (SEPs) in human cells. We identified 90 SEPs, 86 of which are novel, the largest number of human SEPs ever reported. SEP abundances range from 10-1000 molecules per cell, identical to known proteins. SEPs arise from sORFs in non-coding RNAs as well as multi-cistronic mRNAs, and many SEPs initiate with non-AUG start codons, indicating that non-canonical translation may be more widespread in mammals than previously thought. In addition, coding sORFs are present in a small fraction (8/1866) of long intergenic non-coding RNAs (lincRNAs). Together, these results provide the strongest evidence to date that the human proteome is more complex than previously appreciated.

†Correspondence to: saghatelian@chemistry.harvard.edu. .
*These authors contributed equally to this work.

**Author Contributions**: A.J.M., S.A.S, A.G.S., M.N.C., J.M., J.Z.L., A.D.K., B.A.B., J.L.R., and A.S designed the experiments. A.J.M., S.A.S, A.G.S., M.N.C., A.D.K., B.A.B. performed the experiments. A.J.M., S.A.S., J.M. and A.G.S., and B.A.B. collected the peptidomics data and with A.D.K. searched this against the RefSeq database. J.Z.L. provided the RNA-seq data. M.N.C., A.J.M. and J.L.R. performed the lincRNA analysis. S.A.S. performed all cell imaging studies, cloning, and FRAT2 experiments. A.J.M., S.A.S, A.G.S., M.N.C., J.L.R., and A.S discussed the results and implications and wrote the manuscript together.

Competing Financial Interests: The authors claim no competing financial interests.

## Introduction

The complexity of the small proteome remains incompletely explored because genome annotation methods generally break down for small open reading frames (ORFs), generally with a length cutoff of 100 amino acids (this needs a citation - Frith MC et al. again). Computational[1] and ribosome profiling[2] studies have suggested that thousands of these non-annotated mammalian sORFs are translated. However, since these studies did not directly detect the presence of any sORF-encoded polypeptides (SEPs), it remains unknown whether sORFs produce polypeptides that persist in cells at biologically relevant concentrations, or are rapidly degraded. Indeed, biochemical analysis of the translation of two sORFs identified in the yeast GCN4 gene by ribosome profiling revealed that only one expressed detectable polypeptide product[3].

If SEPs do exist at physiologically relevant concentrations in cells, they may execute biological functions. Short open reading frames (sORFs) in the 5′-untranslated region (5′-UTR) of eukaryotic mRNAs (uORFs) are well studied[4-6] and some have been shown to produce detectable polypeptides[7, 8]. In addition to uORFs, other sORFs in bacteria[9], viruses[10], plants[11, 12], *Saccharomyces cerevisiae*[13], *Caenorhabditis elegans*[14], insects[15, 16], and humans[17] have recently been discovered to produce polypeptides. Notably, the peptides encoded by the polycistronic *tarsel-less* (*tal*) gene in *Drosophila*, which are as short as 11 amino acids, regulate fly morphogenesis[15, 16].

While no general method for discovering SEPs exists, attempts have been made to systematically identify these molecules. In *E. coli*, for example, experiments in which predicted sORFs were epitope-tagged revealed 18 SEPs[18] (which we define as polypeptides that are synthesized on the ribosome at a length of less than 150 amino acids). In another example, a combination of computational and experimental approaches identified 299 potentially coding sORFs in *S. cerevisiae*, four of which were confirmed to produce protein and 22 of which appeared to regulate growth[13]. In human cells, an unbiased proteomics approach identified a total of four SEPs in K562 and HEK293 cell lines with a length distribution of 88-148 amino acids[19]. The discordance between the small number of SEPs detected with previous methodologies in human cells[19] and the large number of coding sORFs described by ribosome profiling[2] and computational methods[1] leaves open the possibility that SEPs are not produced as predicted or are rapidly degraded and therefore not detectable.

To resolve this question we developed of a novel SEP discovery and validation strategy that combines peptidomics and massively parallel RNA sequencing (RNA-seq) (Fig. 1a). This strategy uncovered 90 SEPs, 86 of which are novel, demonstrating that SEPs are much more abundant that previously reported. In addition, characterization of the encoding sORFs revealed interesting non-canonical translation events that give rise to SEPs, including bicistronic expression and the use of non-AUG start codons. One SEP, derived from the DEDD2 gene, localizes to mitochondria, which suggests that SEPs could generally have specific cellular localizations and functions. Together, these results indicate that the human proteome is enriched in complexity through translation of sORFs.

## Results

### Discovering SEPs encoded by annotated transcripts

We developed a novel strategy that combines peptidomics and massively parallel RNA sequencing (RNA-seq) to discover human SEPs (Fig. 1a). Peptidomics augments the traditional liquid chromatography-tandem mass spectrometry (LC-MS/MS) proteomics workflow to preserve and enrich small polypeptides[20]. In this context, the use of

peptidomics increases the total number of SEPs detected, including a greater number of shorter SEPs. We isolated peptides from K562 cells, a human leukemia cell line, because we could use previously reported SEPs in this cell line as positive controls[19]. Endogenous K562 polypeptides were isolated using our standard peptidomics workflow[20] with great care being taken to reduce proteolysis. Proteolysis is detrimental because the processing of cellular proteins greatly increases the complexity of the peptidome, which deteriorates the signal-to-noise ratio during the subsequent analysis[21]. After isolation, the K562 polypeptides were digested with trypsin and analyzed by LC-MS/MS. Based on previous results from our lab[22] and others[23] that the optimal size for detection by LC-MS/MS is approximately 10-20 amino acids, trypsin digest is crucial for high-sensitivity SEP detection.

To identify SEPs it was necessary to use a modified protocol for LC-MS/MS data analysis. Standard proteomics and peptidomics approaches identify peptides by matching experimentally observed spectra to databases of predicted spectra based on annotated genes, which would not include SEPs. We therefore created a custom database containing all polypeptides that could possibly be translated from the human transcriptome (RefSeq) (Fig. 1a). Using Sequest, an analysis program used to identify peptides from MS/MS spectra[24, 25], we compared >200,000 MS/MS peptide spectra to this RefSeq-derived polypeptide database. This resulted in 6548 unique peptide identifications. We arrived at a tentative list of SEPs by keeping only those tryptic peptides that differed by at least two amino acids from every annotated protein to minimize the possibility of false positives arising from polymorphisms in annotated genes.

Due to the small size of SEPs, it is unlikely that an unbiased peptidomics experiment will detect more than one tryptic fragment of a given SEP, though for eleven SEPs we did observe two or more fragments (Supplementary Dataset 1). This contrasts with standard proteomics studies, which, on account of the numerous tryptic fragments generated from full size proteins, use the detection of more than one peptide to support the identification of a protein. Realizing that we would likely not be able to rely on the confidence contributed by the inherent redundancy of multiple-peptide protein identifications for SEP discovery, we submitted the candidate peptide spectrum matches (PSMs) to a rigorous evaluation procedure to ensure the highest confidence for each SEP.

First, we discarded any PSM with an Sf score of less than 0.75 (the threshold for a typical proteomics experiment is $Sf < 0.4$[26]). This eliminated over 95% of the candidate set. We then visually examined each remaining MS/MS spectrum to ensure that it met a stringent set of criteria (Supplementary Results, Supplementary Fig. 1). In particular, we required that there be a sequence tag of five consecutive b- or y-ions, a precursor mass error of <5 ppm, and sufficient sequence coverage to unambiguously differentiate each peptide from every annotated protein sequence. This step reduced the remaining peptide pool by approximately 75%, for a total of 39 putative SEPs. Our PSM evaluation procedure therefore selected the most confident ~1% of the peptide identifications in our original candidate set. As a check on the effectiveness of this procedure, we compared the experimentally collected MS/MS spectra of several identified peptides to that of identical synthetic peptides (Fig. 1b).

Lastly, to further reduce the probability of false positives, we comprehensively assembled and cataloged the K562 transcriptome using RNA-seq and crosschecked the assembled RNA-seq transcripts against our candidate sORF list. In this manner we confirmed that at least 37 of the 39 implicated sORFs are present in this cell line and that no other sequence in the assembled K562 RNA-seq transcripts could produce the detected peptides (Fig. 2 and Supplementary Dataset 1). This eliminated the possibility that the detected SEPs arose from point mutations in annotated genes, longer unannotated ORFs containing identical tryptic peptides, or post-transcriptional modification or editing of RNAs. We note that a similar

sample prepared without trypsin failed to identify any SEPs, demonstrating the importance of trypsin in generating an ideal sample for LC-MS/MS.

The 37 SEPs discovered through analysis of RefSeq transcripts fall into five major categories: (i) those located in the 5′-UTR, (ii) those located in the 3′-UTR, (iii) those located in a different reading frame inside an annotated protein coding sequence (CDS), (iv) those located on non-coding RNAs (ncRNAs), and (v) those located on antisense transcripts (Fig. 2a and 2b). The locations of these sORFs mirror the distribution obtained from ribosome profiling[2], indicating that our peptidomics coverage achieves the necessary breadth and depth to reveal global properties of sORFs (Fig. 2b). Many of these SEPs appear to be derived from polycistronic mRNAs, which is interesting because this phenomenon has historically been thought to be rare in eukaryotes. However, our findings here are again consistent with those of ribosome profiling studies[2].

## SEPs are derived from unannotated transcripts

Some SEPs may have been overlooked (false negatives) in our analysis of RefSeq transcripts due to the presence of RNAs in K562 cells that are not annotated in the RefSeq database. To account for such RNAs we also analyzed the LC-MS/MS peptidomics data using a second custom database derived from K562 RNA-seq data. Furthermore, recognizing that recent ribosome profiling studies identified a number of sORFs within the pool of long intergenic non-coding RNAs (lincRNAs) in mouse [2], we generated an extensive catalog of K562 lincRNAs by applying a previously described lincRNA-calling pipeline[27] to our RNA-seq data and searched the corresponding protein database against our data sets. We applied the same stringent criteria for scoring and assessing peptide-spectral matches, and eliminating peptides with fewer than two differences from annotated proteins; we also eliminated any peptides of fewer than 8 amino acids in order to further reduce false positives. These analyses yielded an additional 54 SEPs.

Combining the RefSeq and RNA-Seq results, we discovered 90 unannotated SEPs, four of which were previously reported and thus served as positive controls[19], and 86 of which are novel (Fig. 2c, Supplementary Dataset 1 and Supplementary Dataset 2). The average length of each tryptic peptide identified using this approach was 13-14 amino acids and 90% of the peptides were longer than 18 amino acids, which supports the use of trypsin to generate an ideal LC-MS/MS sample for SEP discovery (Supplementary Fig. 1). This is the largest number of SEPs ever reported in a single study and increases the total number of known human SEPs[17, 19] by ~18-fold, demonstrating the superior coverage afforded by our approach.,Analysis of the evolutionary conservation of the SEPs across 29 mammalian species suggested that SEPs are more conserved than introns, but not as conserved as known coding genes[28] (Supplementary Fig. 3).

## SEP translation is initiated at non-AUG codons

Because we performed mass spectrometry on trypsin-digested samples, we do not obtain full protein-level SEP sequence coverage, and in particular do not directly observe the N terminus. We therefore assigned the likely start codon for each SEP in order to determine their lengths. When present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence [29]. In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximal SEP length.

Using this approach, we determined the SEPs to be 18-149 amino acids long, with the majority (~ 80%) being <100 amino acids (Fig. 2c). If we take a more conservative approach

by using an AUG-to-stop or upstream-stop-to-stop, we obtain similar SEP length distribution and retain our smallest SEPs (Supplementary Fig. 4). As the shortest human SEP previously identified by mass spectrometry was 88 amino acids long[19], it is clear that our approach provides superior coverage of small SEPs. This is significant because many previously characterized, functional SEPs in other species are under 50 amino acids[9, 15-17].

Another interesting feature of our results is the preponderance of non-canonical translation start sites: 57% of the detected SEPs do not initiate at AUG codons (Fig. 2d). This finding is consistent with the results of ribosome profiling experiments in mouse, which indicate that, globally, most ORFs contain non-AUG start sites[2]. Below we obtain data demonstrating that these non-AUG sites are the actual initiation codons of the sORFs.

## Supporting SEP length assignments

We used two approaches to gain additional insight into the lengths of our SEPs. First, rather than relying solely on a molecular weight cutoff filter we used polyacrylamide gel electrophoresis (PAGE) to better separate the K562 lysate into different molecular weight fractions. PAGE can be used as a molecular weight fractionation method prior to proteomics and this approach has successfully been used to study proteolysis[30]. With SEPs, PAGE would provide a tighter molecular weight range, which would support the assigned lengths of the SEPs. Indeed, analysis of the ~10-15 kDa portion of the K562 found SEPs that we had identified as being 90-120 amino acids in length, supporting that these SEPs are intact in these cells which would lead them to migrate at ~10-15 kDa (Supplementary Dataset 1). Importantly, using this approach, we identified more than one tryptic peptide for several SEPs that previously presented only one, providing even greater confidence in the SEP assignments.

We still wanted to detect full-length SEPs directly in K562 lysates and therefore performed an isotope-dilution mass spectrometry (IDMS) experiment with chemically synthesized full-length SEPs. We prepared two SEPs, M***L***HSRKRELRQVLITNKNQVLITNKQVRLTLLLTLG and M***L***RCFFPKMCFSTTIGGMNQRGKRK, with a deuterated leucine (d10-Leu, amino acid that is bold, red and in italics). These two peptides were then added to K562 lysate and the sample was analyzed by LC-MS. These peptides co-eluted with peptides from the sample with the correct mass for the natural SEPs (Supplementary Fig. 5). Due to the high charge state of the peptides (+5 ions) the tandem MS (CID) was not informative, which led us to use additional methods for conformation including IDMS of trypsin fragments and cellular imaging experiments. Our current instrumentation configuration is not designed to easily measure full-length SEPs directly from lysates; however, other mass spectrometry methods including top-down proteomics[31] and high-resolution mass spectrometry approaches for peptide detection[32], should enable the discovery and/or validation of full-length SEPs in the future.

## Cellular concentrations of SEPs

We wished to explore the biological properties of SEPs. First, we examined the cellular concentrations (K562 cells) of three selected SEPs (ASNSD1-SEP, PHF19-SEP and H2AFx-SEP) using isotope dilution mass spectrometry[33] (Fig. 3a). (We refer to SEPs by appending “-SEP” to the name of the annotated CDS nearest the sORF; the sORF is given the same name but italicized.) These SEPs were found at concentrations between 10 and 2000 copies per cell (Supplementary Table 1). Thus, based on previous estimates of protein copy numbers, SEPs are found at concentrations well within the range of typical cellular proteins[34-36]. We further note that the MS/MS spectra from the synthetic standards used in

these experiments were nearly identical to those produced from the endogenous peptide and eluted at the same retention time as same, thus confirming these identifications (Fig. 3b).

## Heterologous expression of SEPs

We tested whether the implicated RNA transcripts and sORFs were competent to produce SEPs. Constructs were designed to produce full-length mRNAs, including 5′ and 3′ UTRs, that matched those in the RefSeq database[37]. We selected sORFs in the 5′-UTR, the 3′-UTR, or frameshifted within the CDS, and encoded a FLAG epitope tag at the 3′-end of each sORF (so that initiation is unperturbed). The uORFs *ASNSD1-SEP*, *PHF19-SEP, DNLZ-SEP, EIF5-SEP, FRAT2-SEP*, *YTHDF3-SEP, CCNA2-SEP, DRAP1-SEP, TRIP6-SEP*, and *C7ORF47-SEP* all produced cytoplasmically localized polypeptides, as detected by anti-FLAG immunofluorescence in transfected HEK293T cells (Fig. 4a and Supplementary Fig. 6). Most importantly, the fact that *FRAT2-SEP, YTHDF3-SEP*, *CCNA2-SEP, DRAP1-SEP, TRIP6-SEP, C7ORF47-SEP*, which do not have any upstream in frame AUG codons, produced SEPs verifies that sORFs with non-AUG start codons are translated (Fig. 4a).

By contrast, the *DEDD2-SEP* sORF was not translated from the full-length RefSeq construct. *DEDD2-SEP* is frameshifted deep within the main CDS of the DEDD2 transcript, so according to the scanning model of translation[38] it is not expected that this downstream sORF would be translated (Fig. 4a). One possible explanation for our observation of the DEDD2-SEP is that it is translated from a splice variant of the DEDD2 RNA that is present in K562 cells, but is not in RefSeq. In support of this hypothesis, we identified a truncated DEDD2 mRNA in the RNA-seq data wherein the first start codon is that of the *DEDD2-SEP* sORF (Supplementary Fig. 7). The 3′-UTR-embedded H2AFx-SEP was similarly not translated from the full-length mRNA construct; however, we were not able to clearly identify a truncated version of the H2AFx transcript in the K562 RNA-seq data. It is possible that a truncated H2AFx mRNA variant is present in K562 cells but is not detectable or not resolvable from the full-length H2AFx transcript.

## SEPs exhibit subcellular localization

We subcloned expression constructs for FLAG-tagged DEDD2-SEP and H2AFx-SEP to determine whether these SEPs are stable. The *H2AFx-SEP* sORF produced a cytoplasmic polypeptide in HEK293T cells (Supplementary Fig. 8). Interestingly, DEDD2-SEP localizes to mitochondria in HEK293T, mouse embryonic fibroblast (MEF), and COS7 cells, as demonstrated by co-localization with the mitochondrial marker MitoTracker Red (Fig. 4b and Supplementary Fig. 9). The N-terminus of DEDD2-SEP is predicted to contain a mitochondrial import signal[39]. Sequence-dependent trafficking and subcellular localization of SEPs could therefore be general phenomena related to their biological activities.

## Non-AUG start codons enable bicistronic expression

Since such a large proportion of SEPs putatively initiate at non-AUG sites, we wanted to rigorously identify the alternate start codon of one these sORFs. C-terminally FLAG-tagged FRAT2-SEP was expressed from the full-length mRNA construct in HEK293T cells and immunoprecipitated; mass spectrometry of the purified protein (Supplementary Fig. 10) was consistent with initiation at an ACG triplet embedded within a Kozak consensus sequence[29] (Supplementary Fig. 11). Mutating the ACG to an ATG resulted in increased FRAT2-SEP translation while deletion of this ACG abolished FRAT2-SEP production, as assessed by Western blotting, thus confirming our assignment (Fig. 5a). In addition, mutation of the Kozak consensus sequence to less favorable residues led to markedly lower FRAT2-SEP expression, which demonstrates the importance of the Kozak sequence at non-AUG initiation sites.

The scanning model of translation provided an explanation as to why the DEDD2 mRNA is not bi-cistronic; we hypothesized that upstream alternate start codons could provide a mechanism to promote polycistronic gene expression via leaky scanning. To test whether FRAT2 mRNA is bi-cistronic, we prepared a FRAT2 construct where the SEP and the downstream CDS were tagged with different epitopes (Fig. 5a), permitting their simultaneous detection by immunoblotting with two antibodies. We found that the FRAT2 RNA is bi-cistronic, as FRAT2 and FRAT2-SEP are both expressed (Fig. 5a). Remarkably, mutation of the ACG start codon of the *FRAT2-SEP* to an ATG increases FRAT2-SEP expression, but also completely eliminates the expression of FRAT2 protein, revealing that the translation of the downstream cistron absolutely requires leaky upstream initiation. Therefore, this experiment indicated that an upstream non-AUG initiation codon is necessary for efficient polycistronic gene expression.

While we attribute FRAT2-SEP translation and bi-cistronic expression to alternate start codon use, we note that another interesting mechanistic possibility for FRAT2-SEP translation is partial (or incomplete) RNA editing, which could modify the ACG to AUG post-transcriptionally. The role of RNA editing in generating sORF start codons at the RNA level could be studied in the future via genetic knockout of the enzymes responsible for this activity[40].

### A Small Subset of lincRNAs encode SEPs

Another intriguing feature of these experiments was the discovery of SEPs encoded by lincRNAs. lincRNAs have emerged as an important class of regulatory molecules with intrinsic biological functions (e.g., hotair, xist)[41, 42]. Ribosome profiling experiments in mouse cells indicate the presence of translated sORFs on nearly half of the lincRNAs analyzed[2], which is much higher than expected[41, 43, 44]. By contrast, our peptidomics analysis identified 8 SEP-encoding lincRNAs (Supplementary Dataset 1), which represents just 0.4% of the1866 lincRNAs detected in our RNA-seq analysis of K562.

This disparity may result from a number of factors, including false positive identifications by ribosome profiling techniques [3]. Additionally, ribosome profiling may identify rare translational events that do not generate enough protein to be detected by LC-MS/MS, since mass spectrometry is biased towards the detection of more abundant peptides[45]. It is also possible that some of the sORFs identified by ribosome profiling may produce polypeptides that are rapidly degraded and therefore would be undetectable using any analytical approach. Future work coupling ribosome profiling with mass spectrometry should help resolve these questions and provide a better understanding of the factors governing SEP expression.

## Discussion

In contrast to previous attempts to use mass spectrometry to discover unannotated human coding sequences, we successfully access the pool of SEPs that are under 50 amino acids in length. This is unprecedented for a global discovery technique and is a crucial step towards understanding the biology of these molecules, for many of the known SEPs[15-17] are below this size threshold. Moreover, the unbiased discovery of SEPs also provided insights into protein translation through the characterization of non-AUG codons and validation of mammalian polycistronic gene expression. Taken together, these findings provide the strongest evidence to date that coding sORFs constitute a significant human gene class. Moreover, due to the bias of mass spectrometry for more abundant species[45], which limits the scope of our technique to the most highly expressed SEPs, and our conservative identification criteria it is probable that there are many more as-yet-undiscovered human SEPs. Thus, we believe we have only begun to explore the breadth and diversity of this new family of polypeptides.

# Online Methods

## Cloning and mutagenesis

DNA constructs were prepared by standard ligation, Quikchange, or inverse PCR techniques. Human cDNA clones were obtained from Open Biosystems and subcloned into pcDNA3, which uses a CMV promoter. Gene synthesis was performed by DNA2.0. Plasmid sequences are publicly available upon request. We note that the YTHDF3-SEP construct consisted of the 5′-UTR putatively encoding the SEP only, obtained via gene synthesis because a full-length cDNA construct with an intact 5′-UTR was not commercially available.

## Cell culture

Cells were grown at 37°C under an atmosphere of 5% $CO_2$. HEK293T, HeLa, COS7 and MEF cells were grown in high-glucose DMEM supplemented with L-glutamine, 10% fetal bovine serum, penicillin and streptomycin. K562 cells were maintained at a density of 1-10 × $10^5$ cells/mL in RPMI1640 media with 10% fetal bovine serum, penicillin and streptomycin.

## Isolation and processing of polypeptides

Aliquots of 3 × $10^7$ growing K562 cells were placed in 1.5 ml Protein LoBind Tubes (Eppendorf), washed three times with PBS, pelleted and stored at −80 °C. Boiling water (500 μl) was added directly to the frozen cell pellets and the samples were then boiled for 20 minutes to eliminate proteolytic activity[20, 22]. After cooling to room temperature, samples were sonicated on ice for 20 bursts at output level 4 with a 40% duty cycle (Branson Sonifier 250; Ultrasonic Convertor). The cell lysate was then brought to 0.25% acetic acid by volume and centrifuged at 20,000 × g for 20 minutes at 4°C. The supernatant was sent through a 30 kD or 10 kD molecular weight cut-off (MWCO) filter (Modified PES Centrifugal Filter, VWR). The mix of small proteins and peptides in the flow-through was evaluated for protein content by BSA assay and then evaporated to dryness at low temperature in a SpeedVac. Pellets were re-suspended in 50 μl of 25mM TCEP in 50mM $NH_4HCO_3$ (pH=8) and incubated at 37 °C for 1 hour. The reaction was cooled to room temperature before 50 μl of a 50 mM iodoacetamide solution in 50 mM $NH_4HCO_3$. This solution was incubated in the dark for 1 hour. Samples were then dissolved in a 50 mM $NH_4HCO_3$ solution of 20 μg/μl trypsin (Promega) to a final protein to enzyme mass ratio of 50:1. This reaction was incubated at 37 °C for 16 hours, cooled to room temperature and then quenched by adding neat formic acid to 5% by volume. The digested peptide mix was then bound to a C18 Sep Pak cartridge (HLB, 1$cm^3$; 30mg, Oasis), washed thoroughly with water and eluted with 1:1 acetonitrile/water. The eluate was evaporated to dryness at low temperature on a SpeedVac.

## Offline electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) fractionation of polypeptide fraction

To simplify the sample and thereby improve detection sensitivity in the subsequent LC-MS/MS analysis, we separated the processed peptide mix by ERLIC[46,47]. ERLIC was performed using a PolyWax LP column (200 × 2.1 mm, 5μm, 300Å; PolyLC Inc.) connected to an Agilent Technologies 1200 Series HPLC equipped with a degasser and automatic fraction collector. All runs were performed at a flow rate of 0.3 ml/min and ultraviolet absorption was measured at a wavelength of 210 nm. Forty (30 kD sample) or 25 (10 kD sample) fractions were collected over a 70 minute gradient beginning with 0.1% acetic acid in 90% acetonitrile (aq.) and ending with 0.1% formic acid in 30% acetonitrile (aq.). The fractions

were then evaporated to dryness on a SpeedVac and dissolved in 15 μl 0.1% formic acid (aq.) in preparation for LC-MS/MS analysis.

LC-MS/MS analysis. Samples were injected onto a NanoAcquity HPLC system (Waters) equipped with a 5 cm x 100 μm capillary trapping column (New Objective) packed with 5 um C18 AQUA beads (Waters) and a PicoFrit SELF/P analytical column (15 μm tip, 25 cm length, New Objective) packed with 3 μm C18 AQUA beads (Waters) and separated over a 90 minute gradient at 200 nl/min. This HPLC system was online with an LTQ Orbitrap Velos (Thermo Scientific) instrument, which collected full MS (dynamic exclusion) and tandem MS (Top 20) data over 375-1600 m/z with 60,000 resolving power.

## Data processing

The acquired MS/MS spectra were analyzed with the SEQUEST algorithm using a database derived from 6-frame (forward and reverse) translation of RefSeq (National Center for Biotechnology Information) mRNA transcripts or 3-frame (forward only) translation of a transcriptome assembly generated by Cufflinks[48] using RNA-Seq data from the K562 cell line (data acquisition described below). The search was performed with the following parameters: variable modifications, oxidation (Met), N-acetylation; semitryptic requirement; maximum missed cleavages: 2; precursor mass tolerance: 20 ppm; and fragment mass tolerance: 0.7 Da. Search results were filtered such that the estimated false discovery rate of the remaining results was 1%. The Sf score is the final score for protein identification by the Proteomics Browser software based on a combination of the preliminary score, the cross-correlation and the differential between the scores for the highest scoring protein and second highest scoring protein[26].

Identified peptides were searched against the Uniprot human protein database using a string-searching algorithm. Peptides found to be identical to fragments of annotated proteins were eliminated from the SEP candidate pool. The remaining peptides were searched against non-redundant protein sequences using the Basic Local Alignment Search Tool (BLAST). Any peptides found to be less than two amino acids different from the nearest protein match (i.e., identical or different by one amino acid) were discarded.

The spectra of the remaining peptides were subjected to a rigorous manual validation procedure: spectra were rejected if they had a precursor mass error of >5 ppm, if they lacked a sequence tag of 5 consecutive b- or y-ions, if they had more than one missed cleavage, or if they lacked sufficient sequence coverage to differentiate from the nearest annotated protein. Finally, peptides under 8 amino acids in length were discarded in order to further minimize false positive identifications.

## RNA-Seq library preparation, alignment, and transcriptome assembly

Two types of cDNA libraries were generated from K-562 RNA (Ambion). In the first experiment, we used 50 nanograms of polyA$^+$ RNA to create standard, non-strand-specific cDNA libraries with paired-end adaptors as previously described [49] and sequenced it on one lane of an Illumina Genome Analyzer IIa machine. In the second experiment, we used eight different amounts of total RNA (30 ng, 100 ng, 250 ng, 500 ng, 1000ng, 3000 ng, and 10,000 ng) to create cDNA libraries with paired-end, indexed adaptors following the instructions for the Illumina TruSeq RNA sample prep kit, except that we used SuperScript III instead of SuperScript II and optimized PCR cycle number. These libraries were sequenced on a single lane of a HiSeq2000 machine. RNA-Seq reads were aligned to the human genome (Hg19 assembly) using TopHat [version V1.1.4;[50]] and transcriptome assembly was performed using Cufflinks [version V1.0.0;[48]]. lincRNAs were called based

on a lincRNA-calling pipeline as previously described[27]. The transcriptome data is deposited on GEO (GSE34740).

### Peptide synthesis, purification and concentration determination

Automated (PS3 Protein Technology, Inc.) solid-phase peptide synthesis was carried out using Fmoc amino acids. Crude peptides were HPLC (Shimadzu)-purified using a C18 column (150 mm × 20 mm, 10 μm particle size, Higgins Analytical). The mobile phase was adjusted for each peptide; buffer A was 99% $H_2O$, 1% acetonitrile, and 0.1% TFA; buffer B was 90% acetonitrile, 10% $H_2O$, and 0.07% TFA). Pure fractions were identified by MALDI-MS analysis, combined, and lyophilized. Peptide concentrations were determined by amino acid analysis (AlBio Tech).

### SEP analysis by PAGE

A total of 600μg of K562 protein was loaded on to 4 lanes (150 μg protein/lane) and run on a 16% Tricine gel 1.0 mm (Novex) at 100 V for 90 minutes. The gel was stained with Coomassie blue for 1 hour and destained. Dual Xtra Standards (Bio-Rad) was used as the molecular weight marker. The gel was then excised into three sections, 10-15 kDa and transferred into 1.5 ml Protein LoBind Tubes (Eppendorf). Each fraction of the gels was washed with 1 ml of 50% HPLC grade acetonitrile/water three times. The samples were stored at −80 °C before LC-MS/MS analysis. In gel trypsin digestion was performed and the sample was then analyzed using the standard LC-MS method.

### Confirmation of the existence of full-length SEPs

Automated (PS3 Protein Technology, Inc.) solid phase peptide synthesis was carried out using Fmoc amino acids and pyclock as an activation reagent. One leucine residue on each peptide was replaced with isotopically labeled d10 Leucine-Fmoc (Sigma). Successful peptide synthesis was confirmed via MALDI-TOF and LC-MS/MS. Peptides from $9\times10^7$ K562 cells were isolated as previously described except no tryptic digest was performed. Peptides were dissolved in 95% water and 5% acetonitrile. Synthetic peptides were added to the endogenous peptide aliquot to a concentration of 2.8nM. The sample was analyzed on a LC-MS/MS LTQ-Orbitrap Velos system as previously described except chromatography was conducted over a 360 minute gradient, and ions corresponding to the +5 charge state of the synthetic and endogenous full length peptides were targeted for fragmentation by CID.

### Absolute quantification of SEPs

Isotope dilution mass spectrometry (IDMS)[33] was used to determine the concentration of SEPs in K562 cells. All samples for this experiment were prepared by adding known amounts of heavy isotope-labeled peptides corresponding the detected fragment of the SEP of interest to a K562 cell pellet ($10^7$ cells) just before isolation of the polypeptides from these cells. The preparation of these samples was identical to that described above except that no ERLIC separation was done. The first step of the quantification procedure was to prepare a set of samples where each sample contained a different but known amount (1 fmol, 10 fmol, 50 fmol, 100 fmol, 500 fmol, 1 pmol or 10 pmol) of the heavy-labeled counterpart peptide. These samples were then analyzed by a selected ion monitoring (SIM) method on the previously described LC-MS/MS system and the resulting data was analyzed using Xcaliber 2.0 (Thermo Scientific). The areas of the peaks corresponding to the endogenous and isotope-labeled peptides were compared to determine the approximate concentration of the SEP and a standard curve was generated to verify that the quantity of the SEP fragment was within the linear range of the mass spectrometer. A second set of samples that each contained an amount of isotope-labeled peptide that was within the linear range of the instrument and within an order of magnitude of the amount of the

corresponding endogenous peptide in the cells was then prepared (N=4) and analyzed as described. The results of this experiment were used to determine the absolute cellular concentration of the selected SEPs.

### Imaging SEPs by immunofluorescence

HeLa, COS7, and MEF cells were grown to 80% confluency on glass coverslips in 48-well plates; HEK293T cells were grown to 50-75% confluency on fibronectin (Millipore)-coated glass coverslips in 48-well plates. Cells were transfected in Opti-MEM (Invitrogen) with 250 ng plasmid DNA using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. 24 hours after transfection, cells were fixed with 4% formalin in phosphate buffered saline (PBS) for 10 minutes at room temperature, and then permeabilized with methanol at −20°C for 10 minutes. Fixed cells were blocked with blocking buffer (3% BSA in PBS with 0.5% Tween-20), then incubated overnight at 4°C with anti-FLAG M2 antibody (Sigma) diluted 1:1000 in blocking buffer. After washing 3x with PBS, cells were then stained for one hour at room temperature with goat anti-mouse AlexaFluor 488 conjugate (Invitrogen) diluted 1:1000 in blocking buffer. Cells were washed 3x with PBS, post-fixed with 4% formalin for 10 minutes at room temperature, then counterstained with a final concentration of 270 ng/mL Hoescht 33258 (Invitrogen) in PBS for 15 minutes at room temperature. Cells were then imaged in PBS in glass-bottom imaging dishes (Matek Corp.). For mitochondrial co-localization analysis, transfected cells were treated with MitoTracker Red CMXRos (Invitrogen) at a final concentration of 100 nM in PBS at 37°C for 15 minutes, washed once with PBS, then fixed with formalin and methanol and immunostained as described above.

Images were acquired in the Harvard Center for Biological Imaging on a Zeiss LSM 510 inverted confocal microscope with the following lasers: 405 Diode, 488 (458,477,514) Argon, 543 HeNe and 633 HeNe. Image acquisition was with either AIM or Zen software. Images were acquired with a 60x oil immersion objective.

### Determination of the FRAT2-SEP start codon by immunoprecipitation and MALDI-MS

COS7 and HEK293T cells were grown in 10-cm dishes to 75% confluency, then transfected with 10 μg plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. 24 hours after transfection, cells were harvested by scraping and washed 3x with PBS. Cells were lysed in 400 μL Triton lysis buffer (1% Triton X-100 in Tris-buffered saline (TBS) with Roche Complete Mini Protease Inhibitor added) on ice for 15 minutes, then lysates were clarified by centrifugation at 16,100 × g for 20 minutes at 4°C. Clarified lysates were added to 50 μL of PBS-washed anti-FLAG M2 agarose resin (Sigma) and rotated at 4°C for 1 hour. Beads were washed 6x with TBS-T (Tris-buffered saline with 0.05% Tween-20). To elute bound proteins, 50 uL of 100 ug/mL 3x FLAG peptide (Sigma) in TBS-T was added to the resin and rotated at 4°C for 20 minutes. Eluates were stored at −80°C until further analysis.

For MALDI-MS analysis, the entire protein sample was desalted using a C18 Sep Pak cartridge (HLB, 1cm$^3$; 30mg, Oasis) and eluted in 50% acetonitrile. The sample was dried in a SpeedVac, and then dissolved in a final volume of 10 μL mass spectrometry-grade water (Burdick & Jackson). This solution (1 μL) was mixed with matrix (α-cyano-4-hydroxycinnamic acid in 50% acetonitrile, 1 μL) on a stainless steel MALDI plate and air-dried. Data were acquired on a Waters MALDI micro MX instrument operated in linear positive mode. Instrument control and spectral acquisition were with MassLynx software.

### Confirmation of the FRAT2-SEP initiation codon, Kozak sequence, and bicistronic expression by immunoblotting

HEK293T cells were grown to 75% confluency in 6-well plates, then transfected with 10 μg plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. Cells were harvested by vigorous pipetting and lysed in 100 μL Triton lysis buffer. Samples of clarified lysate (20 μL) were mixed with SDS-PAGE loading buffer, boiled, and electrophoresed on 4-20% Tris-HCl gels (Bio-Rad). Two replicate gels were run. Proteins were transferred to nitrocellulose (0.20 μm pore size, Thermo Scientific) and immunoblots were probed with anti-FLAG M2 antibody (Sigma) followed by goat anti-mouse IR dye 800 conjugate (LICOR). For bicistronic expression assays, immunoblots were probed with a mixture of rabbit anti-c-myc antibody (Sigma) and anti-FLAG M2, followed by a mixture of goat anti-mouse IR dye 800 and goat anti-rabbit IR dye 680 (LICOR). A replica immunoblot was probed with mouse anti-β-actin followed by goat anti-mouse IR dye 800. Antibodies were diluted 1:2000 in Rockland Immunochemicals fluorescent blocking buffer. Infrared imaging was performed on a LICOR Odyssey instrument.

### Annotation of SEPs in Supplementary Dataset 1

The full list of SEPs identified in this study including genome coordinates of the sORF, the actual LC-MS detected peptide(s), probable start codon and estimated length of SEP in amino acids are shown in Supplementary Dataset 1. There are a total of 90 SEPs that were identified by LC-MS approach and we validated several of these with a variety of approaches. First, for two of the peptides we confirmed the assignment of the tandem MS spectra by chemically synthesizing the peptide and visually comparing the MS2s. In Supplementary Dataset 1 these peptides are annotated by (synthesized to confirm). In addition, we also synthesized isotopically labeled tryptic peptides (deuterated petpides) for two of the SEPs. These peptides enabled us to quantify and simultaneously validate the assignment of these peptides. These peptides are noted on the table by the addition of (isotope-dilution mass spectrometry/trypsin (idms/tryp)). Next, we validated 10 of these SEPs by epitope tagging (FLAG) the putative sORF followed by heterologous expression and measurement of the epitope tag by cellular imaging. These peptides have (i) after their name in Supplementary Dataset 1.

To gain additional evidence for the lengths of these SEPs we utilized two methods. First, we separated the K562 lysate by polyacrylamide gel electrophoresis (PAGE), which provided much better resolution of the lysate, and enabled us to isolate a much smaller mass range for analysis. Specifically, the ~10-15-kilodalton region of this gel was isolated, trypsin digested, and then analyzed by proteomics. The SEPs we identified in this approach have predicted lengths of 83-136, providing additional support for the length assignments. In addition, additional peptides from many of these SEPs were also identified using this alternative fractionation method (peptides in red lettering) to bring the total number of SEPs with more than one peptide for assignment to 11.

Finally, to ensure the length assignment rigorously, we synthesized two full-length SEPs and introduced a deuterated leucine into these sequences (d10-Leu) to create an isotope labeled full-length SEP standard. This 'heavy'-labeled full-length SEP standard is added to K562 lysate and enables us to find the full-length endogenous (i.e. natural) SEP by co-elution. For the two sequences we prepared, we were able to validate the presence of the predicted full-length SEP in the K562 lysate. These two peptides are listed on this table with (isotope-dilution mass spectrometry (IDMS)/full-length SEP) after the name of the peptide.

In total, 16 SEPs were validated by one of these approaches and an additional 7 had multiple peptides from the same sORF to increase confidence in the SEP. Thus, we have multiple

data to support the identification of 23/90 (~26%) of these SEPs. Importantly, no SEPs were filtered out in these steps indicating that the stringent criteria used in the assignment of these peptides limited false positives. SEPs from non-coding RNAs are in the last column and the database used for their identification is also included.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank Xian Adiconis and Lin Fan for constructing the cDNA libraries used in this study. M.N.C. is supported by an HHMI International Student Research Fellowship and S.A.S. is supported by an NRSA postdoctoral fellowship (1F32GM099408-01). J.L.R. is supported by a Damon Runyon-Rachleff Innovator Award, a Searle Scholars Award and Richard and Susan Smith Family Foundation Fellowship. A.S. is supported by a Burroughs Wellcome Fund Career Award in Biomedical Sciences, a Searle Scholars Award, and an Alfred P. Sloan Fellowship. This work was also supported by the NIH training grant T32GM007598 (A.J.M.), the US National Human Genome Research Institute grant HG03067 (J.Z.L), Director's New Innovator Awards DP2OD00667 (J.L.R.) and DP2OD002374 (A.S.), and National Institute of General Medical Sciences grant R01GM102491 (A.S.) the and support from Harvard University (A.S.).

## References

1. Frith MC, et al. The abundance of short proteins in the mammalian proteome. PLoS Genet. 2006; 2:e52. [PubMed: 16683031]
2. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes. Cell. 2011; 147:789–802. [PubMed: 22056041]
3. Zhang F, Hinnebusch AG. An upstream ORF with non-AUG start codon is translated in vivo but dispensable for translational control of GCN4 mRNA. Nucleic Acids Res. 2011; 39:3128–3140. [PubMed: 21227927]
4. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A. 2009; 106:7507–7512. [PubMed: 19372376]
5. Abastado JP, Miller PF, Hinnebusch AG. A quantitative model for translational control of the GCN4 gene of Saccharomyces cerevisiae. New Biol. 1991; 3:511–524. [PubMed: 1883814]
6. Kozak M. Bifunctional messenger RNAs in eukaryotes. Cell. 1986; 47:481–483. [PubMed: 3779834]
7. Parola AL, Kobilka BK. The peptide product of a 5′ leader cistron in the beta 2 adrenergic receptor mRNA inhibits receptor synthesis. J Biol Chem. 1994; 269:4497–4505. [PubMed: 8308019]
8. Werner M, Feller A, Messenguy F. The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression. Cell. 1987
9. Wadler CS, Vanderpool CK. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. Proc Natl Acad Sci U S A. 2007; 104:20454–20459. [PubMed: 18042713]
10. Jay, G.; Nomura, S.; Anderson, CW.; Khoury, G. Identification of the SV40 agnogene product: a DNA binding protein. 1981.
11. Casson SA, et al. The POLARIS gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning. Plant Cell. 2002; 14:1705–1721. [PubMed: 12172017]
12. Rohrig H, Schmidt J, Miklashevichs E, Schell J, John M. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. Proc Natl Acad Sci U S A. 2002; 99:1915–1920. [PubMed: 11842184]
13. Kastenmayer JP, et al. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res. 2006; 16:365–373. [PubMed: 16510898]

14. Gleason CA, Liu QL, Williamson VM. Silencing a candidate nematode effector gene corresponding to the tomato resistance gene Mi-1 leads to acquisition of virulence. Mol Plant Microbe Interact. 2008; 21:576–585. [PubMed: 18393617]
15. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 2007; 5:e106. [PubMed: 17439302]
16. Kondo T, et al. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol. 2007; 9:660–665. [PubMed: 17486114]
17. Hashimoto Y, et al. A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. Proc Natl Acad Sci U S A. 2001; 98:6336–6341. [PubMed: 11371646]
18. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. Molecular Microbiology. 2008; 70:1487–1501. [PubMed: 19121005]
19. Oyama M, et al. Diversity of translation start sites may define increased complexity of the human short ORFeome. Mol Cell Proteomics. 2007; 6:1000–1006. [PubMed: 17317662]
20. Tinoco AD, Tagore DM, Saghatelian A. Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform. J Am Chem Soc. 2010; 132:3819–3830. [PubMed: 20178363]
21. Svensson M, Skold K, Svenningsson P, Andren PE. Peptidomics-based discovery of novel neuropeptides. J Proteome Res. 2003; 2:213–219. [PubMed: 12716136]
22. Tagore DM, et al. Peptidase substrates via global peptide profiling. Nat Chem Biol. 2009; 5:23–25. [PubMed: 19011639]
23. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res. 2010; 9:1323–1329. [PubMed: 20113005]
24. Eng JK, McCormack AL, Yates JR Iii. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry. 1994; 5:976–989.
25. Yates JR 3rd, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem. 1995; 67:1426–1436. [PubMed: 7741214]
26. Christofk HR, Vander Heiden MG, Wu N, Asara JM, Cantley LC. Pyruvate kinase M2 is a phosphotyrosine-binding protein. Nature. 2008; 452:181–186. [PubMed: 18337815]
27. Cabili MN, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]
28. Garber M, et al. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009; 25:i54–62. [PubMed: 19478016]
29. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 1986; 44:283–292. [PubMed: 3943125]
30. Dix MM, Simon GM, Cravatt BF. Global mapping of the topography and magnitude of proteolytic events in apoptosis. Cell. 2008; 134:679–691. [PubMed: 18724940]
31. Tran JC, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011; 480:254–258. [PubMed: 22037311]
32. Kersten RD, et al. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol. 2011; 7:794–802. [PubMed: 21983601]
33. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA. Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. Mol Cell Proteomics. 2007; 6:2212–2229. [PubMed: 17939991]
34. de Godoy LM, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature. 2008; 455:1251–1254. [PubMed: 18820680]
35. Schwanhausser B, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]

36. Beck M, et al. The quantitative proteome of a human cell line. Mol Syst Biol. 2011; 7:549. [PubMed: 22068332]
37. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35:D61–65. [PubMed: 17130148]
38. Hinnebusch AG. Molecular mechanism of scanning and start codon selection in eukaryotes. Microbiol Mol Biol Rev. 2011; 75:434–467. first page of table of contents. [PubMed: 21885680]
39. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 2004; 340:783–795. [PubMed: 15223320]
40. Wedekind JE, Dance GS, Sowden MP, Smith HC. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. Trends in genetics : TIG. 2003; 19:207–216. [PubMed: 12683974]
41. Guttman M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]
42. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009; 10:155–159. [PubMed: 19188922]
43. Guttman M, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010; 28:503–510. [PubMed: 20436462]
44. Khalil AM, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009; 106:11667–11672. [PubMed: 19571010]
45. Fonslow BR, et al. Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. J Proteome Res. 2011; 10:3690–3700. [PubMed: 21702434]
46. Alpert AJ. Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. Anal Chem. 2008; 80:62–76. [PubMed: 18027909]
47. Hao P, et al. Novel application of electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) in shotgun proteomics: comprehensive profiling of rat kidney proteome. J Proteome Res. 2010; 9:3520–3526. [PubMed: 20450224]
48. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
49. Levin JZ, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010; 7:709–715. [PubMed: 20711195]
50. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]

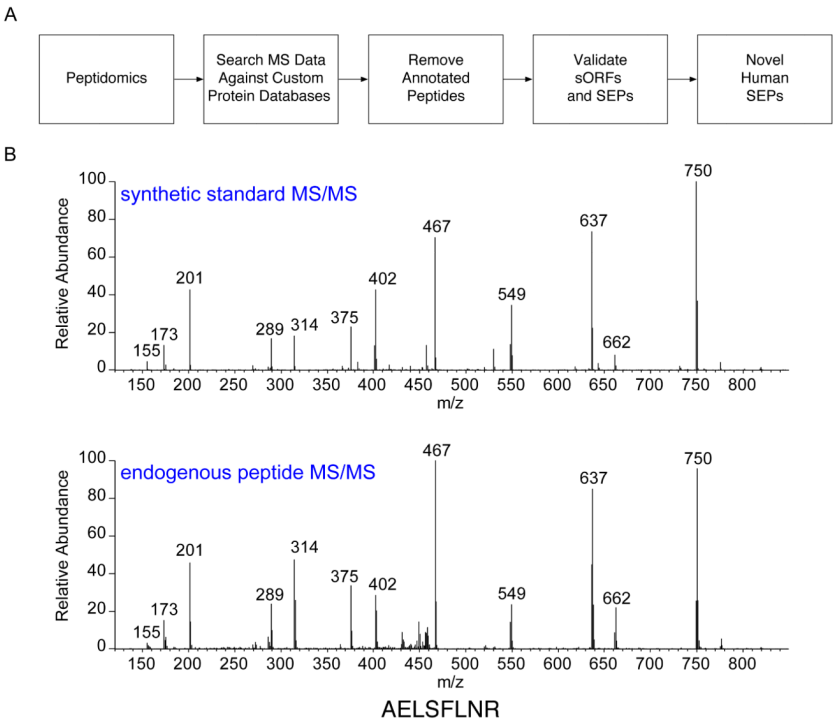


**Fig. 1. Discovering SEPs**

**(a)** An LC-MS/MS-based peptidomics platform was used to profile K562 cells. The MS/MS data were searched against a custom protein database (RefSeq or RNA-seq) to identify polypeptides in K562 cells. Peptides shorter than 8 amino acids were discarded. Tryptic peptides that were exact matches to a segment of an annotated protein were computationally filtered. In addition, tryptic peptides that differed from annotated proteins by a single amino acid were also removed to avoid the false identifications arising from point mutations in known proteins. The sequence assignment of these putative SEPs was validated by visual inspection of the tandem MS spectra. Lastly, K562 RNA-seq data to verify that that detected peptides were derived from a sORF rather than an unannotated ORF longer than 450 nucleotides or a mutated annotated ORF. Any tryptic peptide that fit these criteria was identified as arising from a novel human SEP. (**b**) We experimentally validated one of these assignments by chemically synthesizing the diagnostic peptide and comparing its tandem MS spectra of that of the endogenous peptide. This particular peptide is derived from a sORF found on a non-coding RNA (chr16:86563805-86589025).

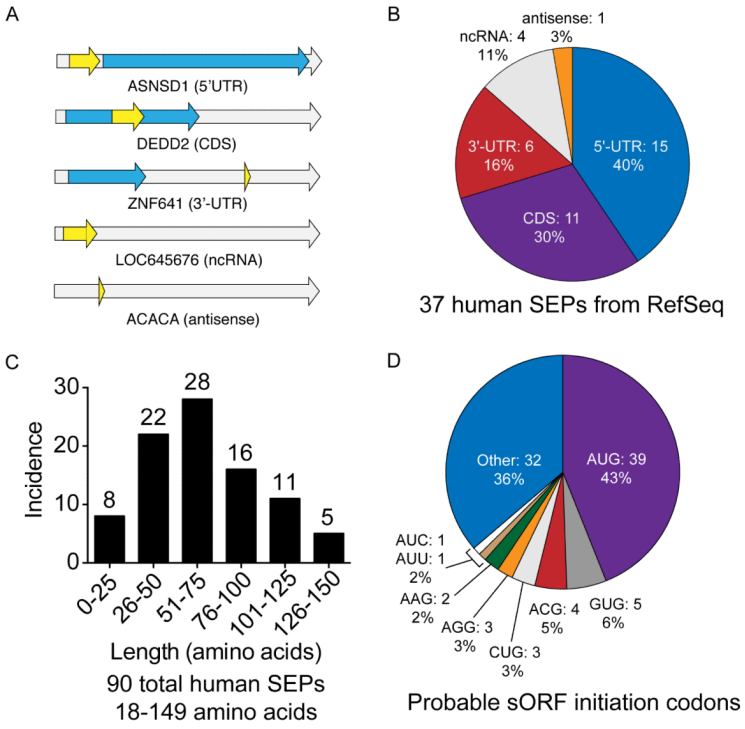


**Fig. 2. Overview of SEPs**

(**a**) RNA maps illustrating the categories of sORFs that are translated into SEPs, including 5′UTR, CDS, 3′UTR, non-coding RNAs and antisense RNAs. The gray arrow represents the RNA, the blue arrow represents annotated protein CDS (if present), and the yellow arrow represents the sORF. (**b**) Incidence of SEPs in each category within RefSeq mRNAs. (**c**) Using protein databases derived from K562 RNA-seq data revealed an additional 54 SEPs for a total of 90 human SEPs, 86 of which are novel. SEP length was estimated by defining sORFs as follows: when present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence [29]. In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximal SEP length. (**d**) Probable sORF initiation codon usage. (Note: RNA maps are not to scale. See Supplementary Fig. 12 for lengths of the RNAs and sORFs.)

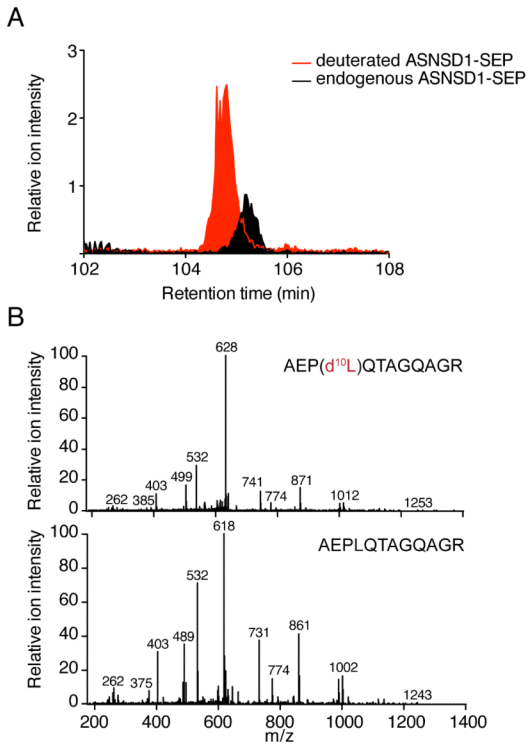


**Fig. 3. SEP quantitation**

(**a**) SEPs were quantified by isotope dilution mass spectrometry (IDMS). We synthesized a deuterated (heavy-labeled) variant of the diagnostic SEP peptide we detected. Upon isolation of K562 cells this peptide was added and the entire mixture was prepared using our standard approach to isolate SEPs. SEPs are then quantified by comparing the peak areas for the deuterated peptide to the endogenous peptide by LC-MS. Since the concentration of the deuterated SEP is known this enables the absolute amount of the endogenous SEP to be determined. Overlap between the endogenous SEP and the deuterated SEP in the LC-MS chromatogram. (**b**) Matching MS/MS spectra (note: 10 Da shift for heavy peptide for some fragments) confirm the peptide sequence assignment in addition to quantifying the peptide.

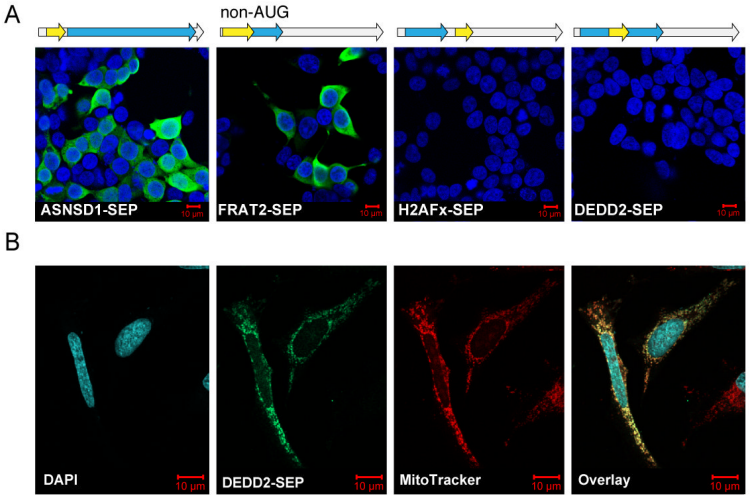


**Fig. 4. Expression of SEPs**

(**a**) Transient transfection of HEK293T cells with constructs containing a cDNA sequence corresponding to the full-length RefSeq mRNA (i.e., including the 5′- and 3′-UTRs). We appended a C-terminal FLAG-tag on the SEP coding sequence that could be detected by immunofluorescence. In these images the nuclei are stained with DAPI (blue) and the SEPs are detected with anti-FLAG antibody (green). ASNSD1-SEP and FRAT2-SEP sORFs in the 5′-UTR (uORFs) but FRAT2-SEP starts with a non-AUG codon. DEDD2-SEP (CDS) and H2AFx-SEP (3′-UTR) were not translated from the RefSeq RNAs, which is consistent with a scanning model of eukaryotic translation. (**b**) DEDD2-SEP was subcloned and expressed in HeLa cells to examine is expression and localization by immunofluorescence. Co-staining with MitoTracker (red) indicated that the DEDD2-SEP localizes to the mitochondria (overlay). (Note: RNA maps are not to scale. See Supplementary Fig. 12 for lengths of the RNAs and sORFs.)

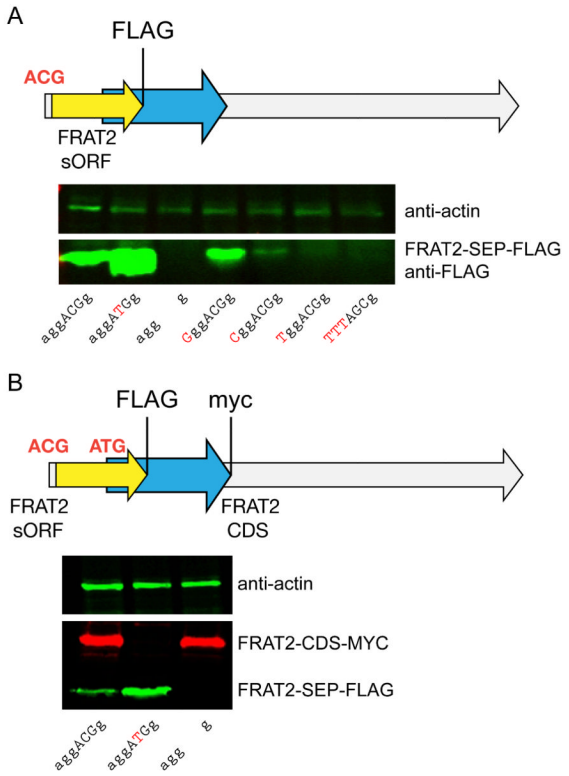


**Fig. 5. Characterization of the non-AUG initiation codon of the *FRAT2-SEP* sORF**

(**a**) An ACG was confirmed as the *FRAT2-SEP* initiation codon by site-directed mutagenesis followed by western blots of FRAT2-SEP-FLAG using an anti-FLAG antibody. Conversion of the ACG to an ATG resulted in higher expression (lane 2), while ablation of this codon removed all expression (lane 3). In addition, perturbation of the Kozak sequence (lanes 4-7) revealed the importance of context when using non-AUG codons, as substitution of less favorable residues [29] at the most important positions in the Kozak sequence resulted in lower FRAT2-SEP-FLAG expression. (**b**) Epitope tagging of the sORF and CDS of the FRAT2 mRNA demonstrates that the FRAT2 mRNA is bi-cistronic. Specifically, the FRAT2 CDS was *c*-myc tagged and the FRAT2-SEP was FLAG tagged. Conversion of the FRAT2-SEP initiation codon from ACG to ATG ablates the expression of the downstream FRAT2-CDS, indicating the importance of alternate start codons for polycistronic expression. (Note: RNA maps are not to scale. See Supplementary Fig. 12 for lengths of the RNAs and sORFs.)