nature chemical biology

ARTICLE

PUBLISHED ONLINE: 18 NOVEMBER 2012 | DOI: 10.1038/NCHEMBIO.1120

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

Sarah A Slavoff[1], Andrew J Mitchell[2,9], Adam G Schwaid[1,9], Moran N Cabili[3–5], Jiao Ma[1], Joshua Z Levin[6], Amir D Karger[7], Bogdan A Budnik[8], John L Rinn[3,5] & Alan Saghatelian[1]*

The complete extent to which the human genome is translated into polypeptides is of fundamental importance. We report a peptidomic strategy to detect short open reading frame (sORF)-encoded polypeptides (SEPs) in human cells. We identify 90 SEPs, 86 of which are previously uncharacterized, which is the largest number of human SEPs ever reported. SEP abundances range from 10–1,000 molecules per cell, identical to abundances of known proteins. SEPs arise from sORFs in noncoding RNAs as well as multicistronic mRNAs, and many SEPs initiate with non-AUG start codons, indicating that noncanonical translation may be more widespread in mammals than previously thought. In addition, coding sORFs are present in a small fraction (8 out of 1,866) of long intergenic noncoding RNAs. Together, these results provide strong evidence that the human proteome is more complex than previously appreciated.

The complexity of the small proteome remains incompletely explored because genome annotation methods break down for sORFs, generally with a length cutoff of 100 amino acids[1]. Computational[1] and ribosome profiling[2] studies have suggested that thousands of these nonannotated mammalian sORFs are translated. However, as these studies did not directly detect the presence of any 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. 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 (tal-1A, tal-2A, tal-3A and tal-AA) encoded by the polycistronic *tarsel-less* (*tal*) gene in *Drosophila*, which are as short as 11 amino acids, regulate fly morphogenesis[15,16].

Although no general method for discovering SEPs exists, attempts have been made to systematically identify these molecules. In *Escherichia 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 seemed to regulate growth[13]. In human cells, an unbiased proteomics approach identified a total of four SEPs in K562 (human leukemia) 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 a 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 previously uncharacterized, demonstrating that SEPs are much more abundant that previously reported. In addition, characterization of the encoding sORFs revealed noncanonical 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 new strategy that combines peptidomics and massively parallel RNA-seq to discover human SEPs (**Fig. 1a**). Peptidomics augments the traditional 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 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,

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, USA. [2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [6]Genome Sequencing and Analysis Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [7]Research Computing, Division of Science, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, USA. [8]Center of Systems Biology, Mass Spectrometry and Proteomics Laboratory, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, USA. [9]These authors contributed equally to this work. *e-mail: saghatelian@chemistry.harvard.edu

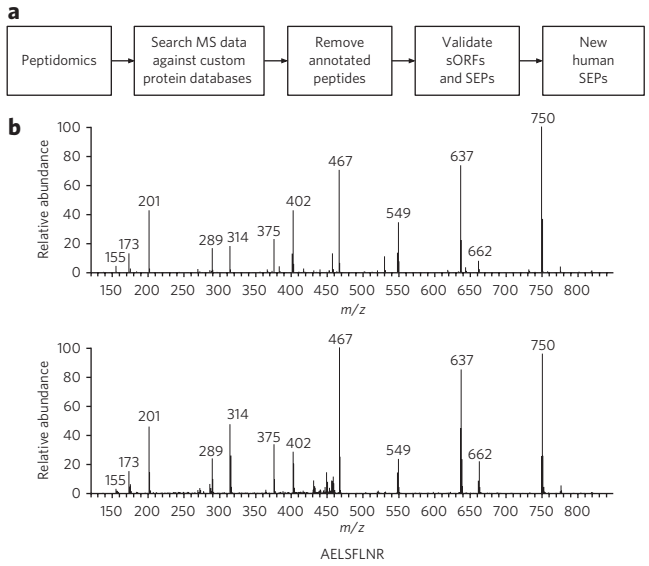

Figure 1 | Discovering SEPs. (**a**) An LC/MS/MS-based peptidomics platform was used to profile K562 cells. The MS/MS data were searched against a custom protein database (RefSeq or RNA-seq) to identify polypeptides in K562 cells. Peptides shorter than eight amino acids were discarded. Tryptic peptides 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, we used K562 RNA-seq data to verify 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 new human SEP. (**b**) We experimentally validated one of these assignments by chemically synthesizing the diagnostic peptide (top) and comparing its tandem MS spectra of that of the endogenous peptide (bottom). This particular peptide is derived from a sORF found on a noncoding RNA (chr16:86563805–86589025).

the K562 polypeptides were digested with trypsin and analyzed by LC/MS/MS. Previous results from our lab[22] and others[23], showing that the optimal size for detection by LC/MS/MS is approximately 10–20 amino acids, suggest that 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 on the basis of annotated genes, which would not include SEPs. We therefore created a custom database containing all polypeptides that could possibly be translated from the annotated human transcriptome (National Center for Biotechnology Information), which we called 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 6,548 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.

Because of 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 Data Set 1**). In contrast, standard proteomic studies, 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 most 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 high-confidence identifications.

First, we discarded any PSM with an Sf score (defined in Online Methods) of less than 0.75 (the threshold for a typical proteomics experiment is Sf < 0.4 (ref. 26)), which 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 p.p.m. and sufficient sequence coverage to unambiguously differentiate each peptide from annotated protein sequences. This step reduced the remaining peptide pool by approximately 75%, giving 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 catalogued the K562 transcriptome using RNA-seq and cross-checked 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 K562 cells and that no other sequence in the assembled K562 RNA-seq transcripts could produce the detected peptides (**Fig. 2** and **Supplementary Data Set 1**). We conclude that observed SEPs must arise from the assigned sORFs and cannot arise from point mutations in annotated genes, longer unannotated ORFs containing identical tryptic peptides, post-transcriptional modifications 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 noncoding RNAs (ncRNAs) and (v) those located on antisense transcripts (**Fig. 2a**,**b**). 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 seem to be derived from polycistronic mRNAs, which is notable 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 owing 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 noncoding RNAs (lincRNAs) in mice[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 for eliminating peptides with fewer than two differences from

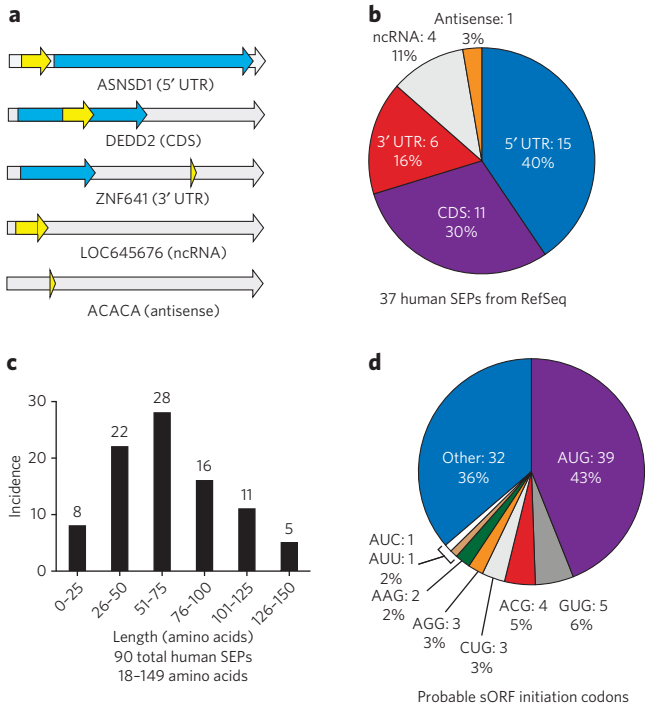

Figure 2 | Overview of SEPs. (**a**) RNA maps illustrating the categories of sORFs that are translated into SEPs, including 5′ UTR, CDS, 3′ UTR, noncoding RNAs (ncRNAs) and antisense RNAs. The gray arrows represents the RNA, the blue arrows represents annotated protein CDS (if present), and the yellow arrows 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 previously uncharacterized. 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. RNA maps are not to scale. Lengths of the RNAs and sORFs are in **Supplementary Figure 12**.

annotated proteins; we also eliminated any peptides of fewer than eight amino acids to further reduce false positives. These analyses yielded an additional 53 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 previously uncharacterized (**Fig. 2c** and **Supplementary Data Set 1**). The average length of each tryptic peptide identified using this approach was 13 or 14 amino acids, and 90% of the peptides were longer than 18 amino acids, which explains why the lack of trypsin yields no SEPs (**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 are not as conserved as known coding genes[28] (**Supplementary Fig. 3**).

SEP translation is initiated at non-AUG codons

Because we performed MS on trypsin-digested samples, we did not obtain full protein-level SEP sequence coverage, and, in particular, we usually did not directly observe the N terminus. We therefore assigned the likely start codon for each SEP to determine its length. 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 MS was 88 amino acids long[19], it is clear that our approach provides superior coverage of small SEPs. This is notable because many previously characterized, functional SEPs in other species are under 50 amino acids in length[9,15–17].

Another feature of our results is the preponderance of noncanonical 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 mice, which indicate that, globally, most ORFs contain non-AUG start sites[2]. Below, we show data demonstrating that these non-AUG sites are the actual initiation codons of the sORFs.

Supporting SEP length assignments

We used two approaches to confirm our SEP length assignments. First, rather than relying on a molecular weight cutoff filter, we used PAGE to better separate the K562 lysate into fractions of different molecular weight. PAGE can be used as a molecular weight fractionation method before proteomics, and this approach has successfully been used to study proteolysis[30]. Indeed, analysis of the gel band corresponding to the ~10- to 15-kDa portion of the K562 proteome found SEPs that we had identified as being 90–120 amino acids in length, as expected (**Supplementary Data Set 1**). 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.

To detect full-length SEPs directly in K562 lysates, we performed an isotope-dilution MS (IDMS) experiment with chemically synthesized full-length SEPs. We prepared two SEPs, MLHSRKRELRQVLITNKNQVLITNKQVRLTLLLTLG and MLRC FFPKMCFSTTIGGMNQRGKRK, with a deuterated leucine (d10-Leu, shown as underlined L). These two peptides were then added to K562 lysate, and the sample was analyzed by LC/MS. These standard peptides co-eluted with endogenous peptides from the sample with the expected mass for the full-length SEP (**Supplementary Fig. 5**). Owing to the high charge state of the peptides (+5 ions), the tandem MS (collision-induced dissociation) was not informative. Our current instrumentation configuration is not designed to easily measure full-length SEPs directly from lysates; however, other MS methods including top-down proteomics[31] and high-resolution MS approaches for peptide detection[32] should enable the discovery and/or validation of full-length SEPs in the future. We therefore pursued additional methods to confirm the existence of SEPs in cells, including IDMS of trypsin fragments and cellular imaging.

Cellular concentrations of SEPs

We examined the cellular concentrations (K562 cells) of three selected SEPs (ASNSD1-SEP, PHF19-SEP and H2AFx-SEP) using isotope dilution MS[33] (**Fig. 3a**) (we refer to SEPs by appending '-SEP' to the gene name producing the encoding RNA). These SEPs

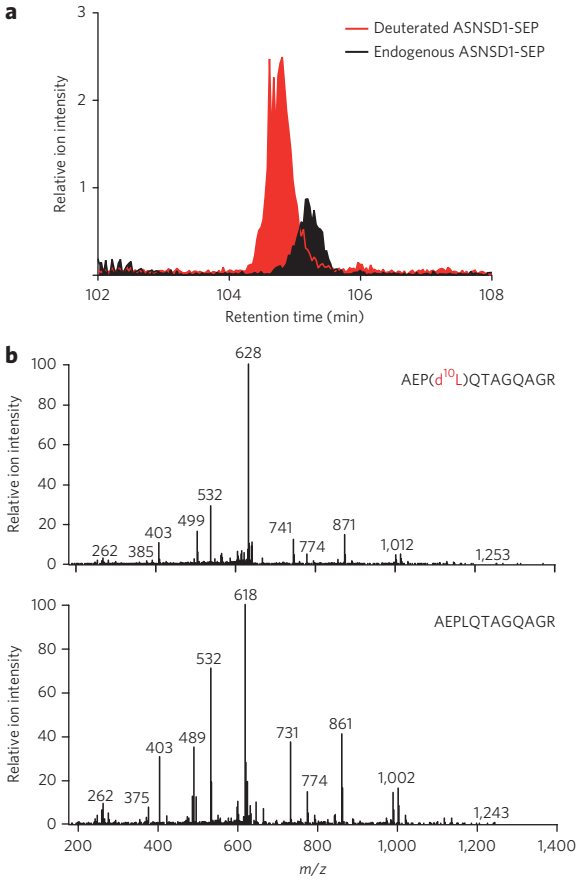

Figure 3 | SEP quantification. (**a**) SEPs were quantified by IDMS. We synthesized deuterated (heavy-labeled) variants of several SEP tryptic peptides. Upon preparation of the K562 peptidome, the deuterated peptides were added exogenously, and the entire mixture was subjected to LC/MS. SEPs were then quantified by comparing the peak areas for the deuterated peptide to the endogenous peptide by LC/MS. As the concentration of the deuterated peptide is known, this enables the absolute amount of the endogenous SEP to be determined. Co-elution of the endogenous and deuterated peptides in the LC/MS chromatogram confirms the identity of the endogenous SEP peptide. (**b**) Matching MS/MS spectra (note the 10-Da shift for heavy peptide for some fragments) confirm the sequence assignment.

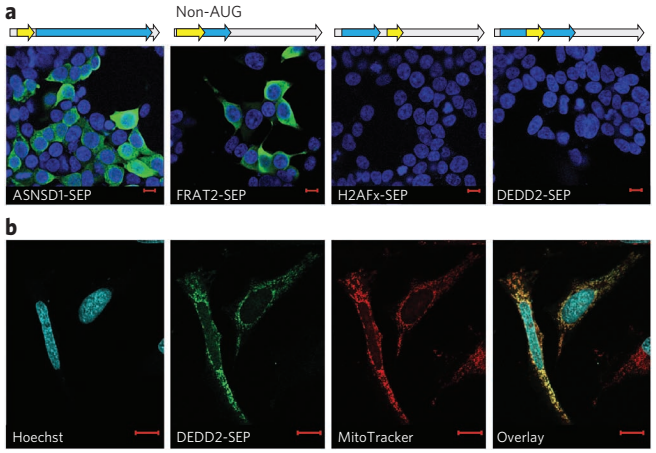

Figure 4 | Expression of SEPs. (**a**) Transient transfection of HEK293T cells with constructs containing a cDNA sequence corresponding to the full-length RefSeq mRNA. We appended a C-terminal Flag tag on the SEP coding sequence that could be detected by immunofluorescence. In these images, the nuclei are stained with DAPI (blue), and the SEPs are detected with Flag-specific antibody (green). ASNSD1-SEP and FRAT2-SEP sORFs are both found in the 5′ UTR of their transcripts. FRAT2-SEP initiates 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**) The DEDD2-SEP sORF was subcloned and expressed in HeLa cells to examine its expression and localization. Costaining with MitoTracker (red) indicated that DEDD2-SEP localizes to the mitochondria (overlay). RNA maps are not to scale. Lengths of the RNAs and sORFs are in **Supplementary Figure 12**. Scale bars, 10 μm.

were found at concentrations between 10 and 2,000 copies per cell (**Supplementary Table 1**). Thus, on the basis of previous estimates of protein copy numbers, SEPs are found at concentrations well within the range of typical cellular proteins[34–36]. Additionally, the MS/MS spectra from the synthetic standards used in these experiments were nearly identical to and eluted at the same retention time as those produced from the endogenous peptide, thus confirming these identifications (**Fig. 3b**).

Heterologous expression of SEPs

We tested whether the implicated RNA transcripts were competent to produce SEPs. Mammalian expression constructs were designed to produce full-length mRNAs, including 5′ and 3′ UTRs, that matched those in the RefSeq database[37]. We selected sORFs that were in the 5′ UTR or 3′ UTR or were frameshifted within the CDS and appended an epitope tag at the 3′ end of each sORF (so that initiation is unperturbed). The uORFs from the genes *ASNSD1*, *PHF19*, *DNLZ*, *EIF5*, *FRAT2*, *YTHDF3*, *CCNA2*, *DRAP1*, *TRIP6* and *C7ORF47* all produced cytoplasmically localized polypeptides, as detected by immunofluorescence, in transfected HEK293T cells (**Fig. 4a** and **Supplementary Fig. 6**). The fact that uORFs from the *FRAT2*, *YTHDF3*, *CCNA2*, *DRAP1*, *TRIP6* and *C7ORF47* genes, 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* sORF was not translated from the full-length RefSeq construct. The *DEDD2* sORF 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 DEDD2-SEP production is that it is translated from a splice variant of the DEDD2 RNA that is present in K562 cells, but it 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* 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 epitope-tagged DEDD2-SEP and H2AFx-SEP to determine whether these SEPs are stable. The *H2AFx* sORF produced a cytoplasmic polypeptide in HEK293T cells (**Supplementary Fig. 8**). DEDD2-SEP localizes to mitochondria in HEK293T, mouse embryonic fibroblast (MEF) and COS7 cells, as demonstrated by colocalization 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.

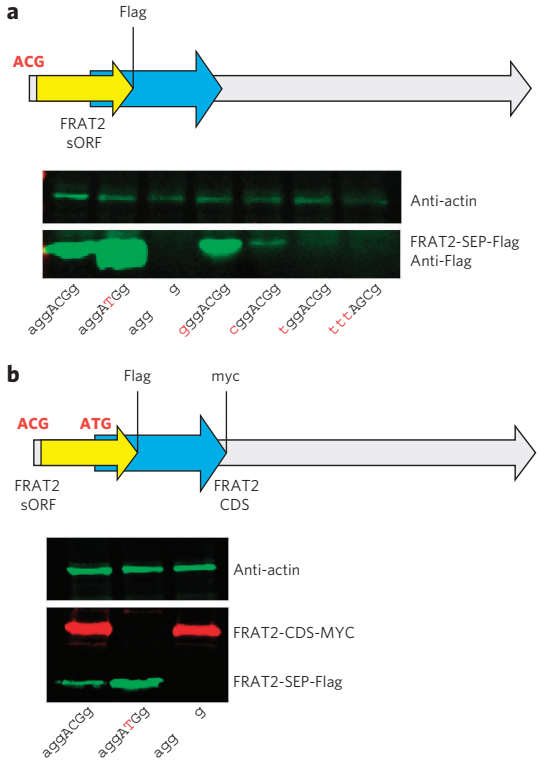

Figure 5 | Characterization of the non-AUG initiation codon of the *FRAT2* sORF. (**a**) The FRAT2 cDNA expression construct, with a Flag epitope tag appended at the C terminus of the FRAT2-SEP sORF, was subjected to site-directed mutagenesis to probe the identity of the sORF start codon by expression in HEK 293T cells followed by western blotting. Below the immunoblot, the sORF Kozak and start codon sequences of the expressed construct are shown, with the start codon shown in uppercase letters and sites of mutations highlighted in red. Conversion of the putative ACG start codon to an ATG resulted in higher expression (lane 2), whereas ablation of this codon eliminated 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] resulted in lower FRAT2-SEP expression. Equal loading was demonstrated with actin-specific (anti-actin) immunoblotting. (**b**) Epitope tagging of the sORF and CDS of the FRAT2 mRNA demonstrates that the FRAT2 mRNA is bicistronic. 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 expression of the downstream FRAT2 CDS, indicating the importance of alternate start codons for polycistronic expression. RNA maps are not to scale. Lengths of the RNAs and sORFs are in **Supplementary Figure 12**.

Non-AUG start codons enable bicistronic expression

As such a large proportion of SEPs putatively initiate at non-AUG codons, we wanted to rigorously identify the alternate start codon of one these sORFs. C-terminally epitope-tagged FRAT2-SEP was expressed from the full-length mRNA construct in HEK293T cells and immunoprecipitated; MS 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 ~three-fold, whereas 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 (**Fig. 5a**).

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 bicistronic, we prepared a FRAT2 construct where the SEP and the downstream CDS were tagged with different epitopes (**Fig. 5b**), permitting their simultaneous detection by immunoblotting with two antibodies. We found that the FRAT2 RNA is bicistronic, as FRAT2 protein and FRAT2-SEP are both expressed (**Fig. 5b**). Remarkably, mutation of the ACG start codon of the SEP to an ATG increases FRAT2-SEP expression, but it also completely eliminates the expression of the downstream 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. We note that another mechanistic possibility for FRAT2-SEP translation is partial 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

lincRNAs have emerged as a class of regulatory molecules with intrinsic biological functions (for example, hotair and 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]. We therefore applied a previously described lincRNA discovery pipeline to our K562 RNA-seq data and determined what percentage of these K562 lincRNAs are translated to produce SEPs. Our peptidomics analysis identified eight SEP-encoding lincRNAs (**Supplementary Data Set 1**), which represents just 0.4% of the 1,866 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, as MS is biased toward the detection of 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 MS should help resolve these questions and provide a better understanding of the factors governing SEP expression.

DISCUSSION

In contrast to previous attempts to use MS to discover unannotated human coding sequences, we have globally accessed the pool of SEPs that are under 50 amino acids in length. This is a crucial step toward understanding the biology of these molecules, for many known SEPs[15–17] are below this size threshold. Moreover, the unbiased discovery of SEPs provided insights into protein translation through the characterization of non-AUG initiation codons and mammalian polycistronic gene expression. Taken together, these findings provide the strongest evidence to date that coding sORFs constitute a human gene class. Moreover, owing to the bias of MS 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 diversity of this new family of genetically encoded polypeptides.

Received 22 March 2012; accepted 16 October 2012;
published online 18 November 2012

METHODS

Methods and any associated references are available in the online version of the paper.

References

1. Frith, M.C. *et al.* The abundance of short proteins in the mammalian proteome. *PLoS Genet.* **2**, e52 (2006).
2. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
3. Zhang, F. & Hinnebusch, A.G. An upstream ORF with non-AUG start codon is translated *in vivo* but dispensable for translational control of GCN4 mRNA. *Nucleic Acids Res.* **39**, 3128–3140 (2011).
4. Calvo, S.E., Pagliarini, D.J. & Mootha, V.K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. USA* **106**, 7507–7512 (2009).
5. Abastado, J.P., Miller, P.F. & Hinnebusch, A.G. A quantitative model for translational control of the *GCN4* gene of *Saccharomyces cerevisiae*. *New Biol.* **3**, 511–524 (1991).
6. Kozak, M. Bifunctional messenger RNAs in eukaryotes. *Cell* **47**, 481–483 (1986).
7. Parola, A.L. & Kobilka, B.K. The peptide product of a 5′ leader cistron in the β2 adrenergic receptor mRNA inhibits receptor synthesis. *J. Biol. Chem.* **269**, 4497–4505 (1994).
8. Werner, M., Feller, A. & Messenguy, F. The leader peptide of yeast gene *CPA1* is essential for the translational repression of its expression. *Cell* **49**, 805–813 (1987).
9. Wadler, C.S. & Vanderpool, C.K. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc. Natl. Acad. Sci. USA* **104**, 20454–20459 (2007).
10. Jay, G., Nomura, S., Anderson, C.W. & Khoury, G. Identification of the SV40 agnogene product: a DNA binding protein. *Nature* **8**, 346–349 (1981).
11. Casson, S.A. *et al.* The *POLARIS* gene of *Arabidopsis* encodes a predicted peptide required for correct root growth and leaf vascular patterning. *Plant Cell* **14**, 1705–1721 (2002).
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. USA* **99**, 1915–1920 (2002).
13. Kastenmayer, J.P. *et al.* Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res.* **16**, 365–373 (2006).
14. Gleason, C.A., Liu, Q.L. & Williamson, V.M. Silencing a candidate nematode effector gene corresponding to the tomato resistance gene *Mi-1* leads to acquisition of virulence. *Mol. Plant Microbe Interact.* **21**, 576–585 (2008).
15. Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A. & Couso, J.P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* **5**, e106 (2007).
16. Kondo, T. *et al.* Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol.* **9**, 660–665 (2007).
17. Hashimoto, Y. *et al.* A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Aβ. *Proc. Natl. Acad. Sci. USA* **98**, 6336–6341 (2001).
18. Hemm, M.R., Paul, B.J., Schneider, T.D., Storz, G. & Rudd, K.E. Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol. Microbiol.* **70**, 1487–1501 (2008).
19. Oyama, M. *et al.* Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol. Cell. Proteomics* **6**, 1000–1006 (2007).
20. Tinoco, A.D., Tagore, D.M. & Saghatelian, A. Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform. *J. Am. Chem. Soc.* **132**, 3819–3830 (2010).
21. Svensson, M., Skold, K., Svenningsson, P. & Andren, P.E. Peptidomics-based discovery of novel neuropeptides. *J. Proteome Res.* **2**, 213–219 (2003).
22. Tagore, D.M. *et al.* Peptidase substrates via global peptide profiling. *Nat. Chem. Biol.* **5**, 23–25 (2009).
23. Swaney, D.L., Wenger, C.D. & Coon, J.J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **9**, 1323–1329 (2010).
24. Eng, J.K., McCormack, A.L. & Yates Iii, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
25. Yates, J.R. III, Eng, J.K., McCormack, A.L. & Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67**, 1426–1436 (1995).
26. Christofk, H.R., Vander Heiden, M.G., Wu, N., Asara, J.M. & Cantley, L.C. Pyruvate kinase M2 is a phosphotyrosine-binding protein. *Nature* **452**, 181–186 (2008).
27. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
28. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).
29. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292 (1986).
30. Dix, M.M., Simon, G.M. & Cravatt, B.F. Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell* **134**, 679–691 (2008).
31. Tran, J.C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).
32. Kersten, R.D. *et al.* A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794–802 (2011).
33. Keshishian, H., Addona, T., Burgess, M., Kuhn, E. & Carr, S.A. Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell. Proteomics* **6**, 2212–2229 (2007).
34. de Godoy, L.M. *et al.* Comprehensive mass-spectrometry–based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
35. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
36. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7**, 549 (2011).
37. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
38. Hinnebusch, A.G. Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol. Mol. Biol. Rev.* **75**, 434–467 (2011).
39. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
40. Wedekind, J.E., Dance, G.S., Sowden, M.P. & Smith, H.C. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet.* **19**, 207–216 (2003).
41. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
42. Mercer, T.R., Dinger, M.E. & Mattick, J.S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).
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.* **28**, 503–510 (2010).
44. Khalil, A.M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
45. Fonslow, B.R. *et al.* Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *J. Proteome Res.* **10**, 3690–3700 (2011).

Acknowledgments

We thank X. Adiconis and L. Fan for constructing the cDNA libraries used in this study. M.N.C. is supported by a Howard Hughes Medical Institute International Student Research Fellowship, and S.A.S. is supported by an National Research Service Award postdoctoral fellowship (1F32GM099408-01). J.L.R. is supported by a Damon Runyon-Rachleff Innovator Award, a Searle Scholars Award and a 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 US National Institutes of Health training grant T32GM007598 (A.J.M.), the US National Human Genome Research Institute grant 3U54HG003067 (J.Z.L.), Director's New Innovator Awards DP2OD00667 (J.L.R.) and DP2OD002374 (A.S.), National Institute of General Medical Sciences grant R01GM102491 (A.S.) and support from Harvard University (A.S.).

Author contributions

A.J.M. and A.G.S. contributed equally to this work. 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. and B.A.B. performed the experiments. A.J.M., S.A.S., J.M., 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 declare no competing financial interests.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to A.S.

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 available upon request. We note that the YTHDF3-SEP and CCNA2-SEP constructs consisted of the 5′ UTR putatively encoding the SEP only and were 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% FBS, penicillin and streptomycin. K562 cells were maintained at a density of 1–10 × 10^5 cells/mL in RPMI 1640 medium with 10% FBS, 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 and 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 15 min to eliminate proteolytic activity[20,22]. After cooling to room temperature, samples were sonicated on ice for 20-s 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 min at 4 °C. The supernatant was sent through a 30-kDa or 10-kDa 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 BCA assay and then evaporated to dryness at low temperature in a SpeedVac. Pellets were resuspended in 50 µl of 25 mM TCEP in 50 mM NH_4HCO_3 (pH 8) and incubated at 37 °C for 1 h. The reaction was cooled to room temperature before addition of 50 µl of a 50-mM iodoacetamide solution in 50 mM NH_4HCO_3. This solution was incubated in the dark for 1 h. Samples were then dissolved in a 50-mM NH_4HCO_3 solution of 20 µg/µl trypsin (Promega) to a final protein to enzyme-to-mass ratio of 50:1. This reaction was incubated at 37 °C for 16 h, 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; 30 mg, 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 mm × 2.1 mm, 5 µm, 300 Å; 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-kDa sample) or twenty-five (10-kDa sample) fractions were collected over a 70-min 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 × 100 µm capillary trapping column (New Objective) packed with 5-µm 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-min 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 375–1,600 *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 six-frame (forward and reverse) translation of RefSeq (National Center for Biotechnology Information) mRNA transcripts or three-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, two maximum missed cleavages; precursor mass tolerance of 20 p.p.m. and fragment mass tolerance of 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 nonredundant human 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 (that is, 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 p.p.m., if they lacked a sequence tag of five 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 eight amino acids in length were discarded to further minimize false positive identifications.

RNA-Seq library preparation, alignment and transcriptome assembly. Two types of cDNA libraries were generated from K562 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 them 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, 1,000 ng, 3,000 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 (ref. 50)), and transcriptome assembly was performed using Cufflinks (version V1.0.0 (ref. 48)). lincRNAs were called on the basis of 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 600 µg of K562 protein was loaded on to four lanes (150 µg protein per lane) and run on a 16% Tricine gel 1.0 mm (Novex) at 100 V for 90 min. The gel was stained with Coomassie blue for 1 h and destained. Dual Xtra Standards (Bio-Rad) was used as the molecular weight marker. The 10- to 15-kDa band was excised and transferred into 1.5-ml Protein LoBind Tubes (Eppendorf). Each gel slice was washed with 1 ml of 50% HPLC grade acetonitrile in 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 × 10^7 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 2.8 nM. The sample was analyzed on an LC/MS/MS LTQ-Orbitrap Velos system as previously described, except chromatography was conducted over a 360-min 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. 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 cover slips in 48-well plates; HEK293T cells were grown to 50–75% confluency on fibronectin (Millipore)-coated glass cover slips 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. Twenty-four hours after transfection, cells were fixed with 4% formalin in PBS for 10 min at room temperature and then permeabilized with methanol at −20 °C for 10 min. Fixed cells were blocked with blocking buffer (3% BSA in PBS with 0.5% Tween-20), then incubated overnight at 4 °C with mouse monoclonal Flag-specific antibody (Sigma, clone M2) diluted 1:1,000 in blocking buffer. After washing 3× with PBS, cells were then stained for 1 h at room temperature with goat mouse-specific AlexaFluor 488 conjugate (Invitrogen) diluted 1:1,000 in blocking buffer. Cells were washed 3× with PBS, post-fixed with 4% formalin for 10 min at room temperature, then counterstained with a final concentration of 270 ng/mL Hoechst 33258 (Invitrogen) in PBS for 15 min at room temperature. Cells were then imaged in PBS in glass-bottom imaging dishes (Matek Corp.). For mitochondrial colocalization analysis, transfected cells were treated with MitoTracker Red CMXRos (Invitrogen) at a final concentration of 100 nM in PBS at 37 °C for 15 min, 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 60× 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. Twenty-four hours after transfection, cells were harvested by scraping and washed 3× 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 min, then lysates were clarified by centrifugation at 16,100*g* for 20 min at 4 °C. Clarified lysates were added to 50 µL of PBS-washed Flag-specific M2 agarose resin (Sigma) and rotated at 4 °C for 1 h. Beads were washed 6× with TBS-T (Tris-buffered saline with 0.05% Tween-20). To elute bound proteins, 50 µL of 100 µg/mL 3× Flag peptide (Sigma) in TBS-T was added to the resin and rotated at 4 °C for 20 min. 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, 1 cm^3; 30 mg, Oasis) and eluted in 50% acetonitrile. The sample was dried in a SpeedVac and then dissolved in a final volume of 10-µL MS-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 performed 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 six-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 mouse monoclonal Flag-specific antibody (Sigma, clone M2) followed by goat mouse-specific IR dye 800 conjugate (LICOR). For bicistronic expression assays, immunoblots were probed with a mixture of rabbit c-myc–specific antibody (Sigma) and mouse monoclonal M2 Flag-specific antibody, followed by a mixture of goat mouse-specific IR dye 800 and goat rabbit-specific IR dye 680 (LICOR). A replica immunoblot was probed with mouse β-actin–specific antibody followed by goat mouse-specific IR dye 800. Antibodies were diluted 1:2,000 in Rockland Immunochemicals fluorescent blocking buffer. Infrared imaging was performed on a LICOR Odyssey instrument.

46. Alpert, A.J. Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. *Anal. Chem.* **80**, 62–76 (2008).
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.* **9**, 3520–3526 (2010).
48. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
49. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
50. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

doi:10.1038/nchembio.1120