Supplementary Information

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

*These authors contributed equally to this work.

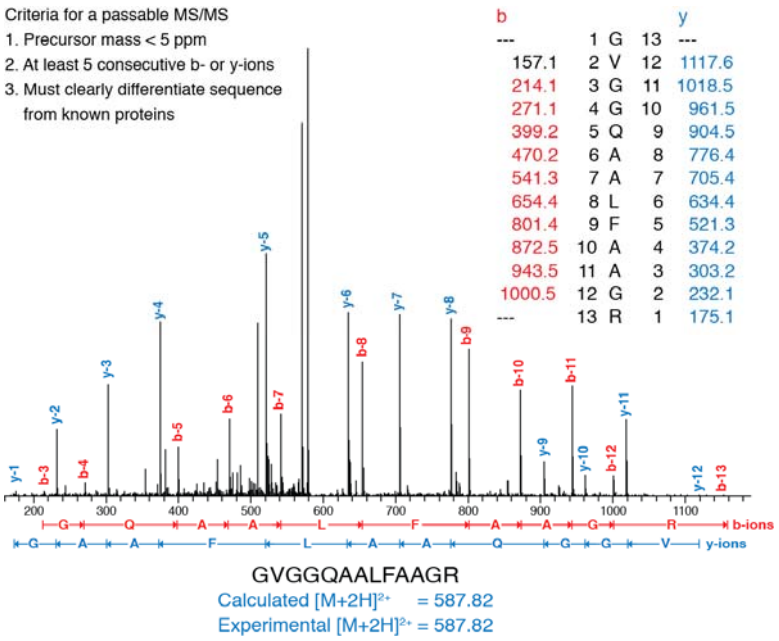
[†]Correspondence to: saghatelian@chemistry.harvard.edu.

Supplementary Results

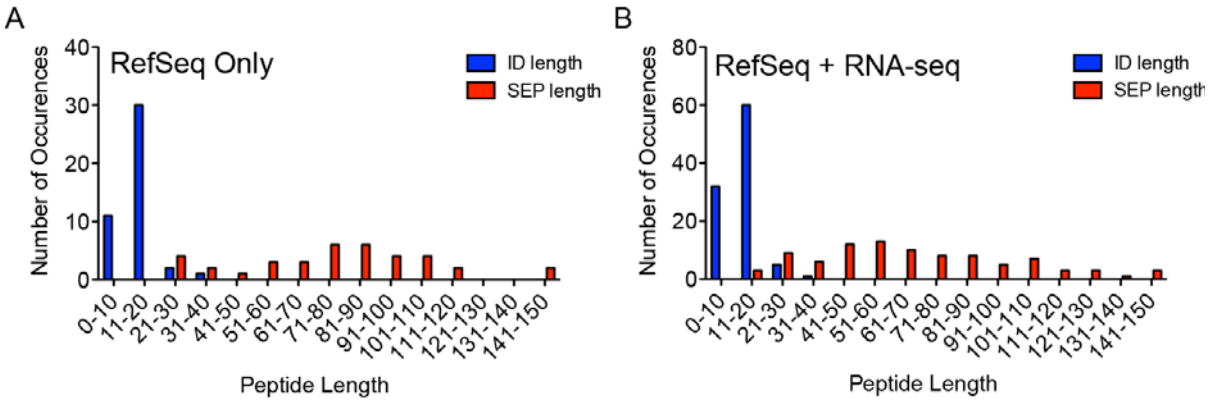
Supplementary Table 1. Quantification of SEP trypsin peptides by IDMS.

SEP	Peptide	Copies/Cell
H2AFX	AEPLQTAGQAGR	1728
ASNSD1	EYQEIENLDK	386
PHF19	LQVGPADTQPR	6

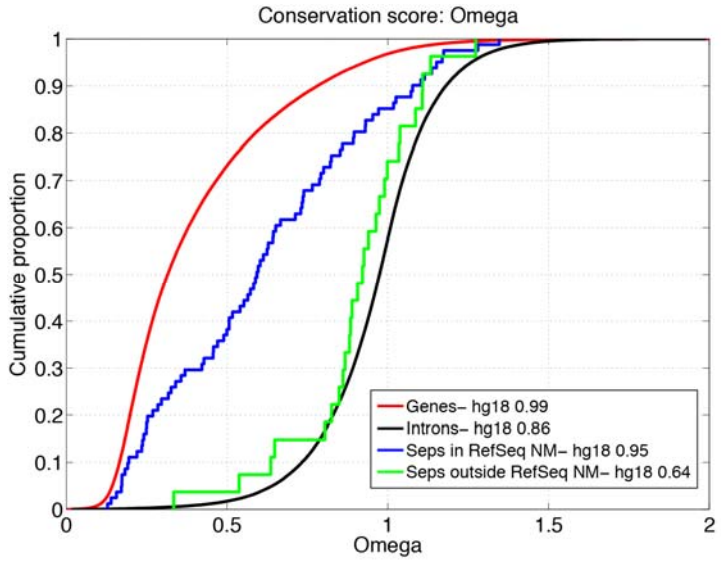
Supplementary Dataset 1. Full-list of identified SEPS. SEPs validated through alternative methods or having more than one peptide ID per SEP. The following were used to annotate the different methods: imaging (i), isotope dilution-MS of tryptic fragments (idmd/tryp), comparison of tandem MS (MS/MS) spectra of natural peptides to synthetic peptides (synthesized to confirm), co-elution IDMS of full-length synthetic heavy-labeled peptides with endogenous SEPs (idms/full-length SEP), and SEP peptides identified by PAGE followed by trypsin and LC-MS analysis of the 10-15 kDa region of the gel are showin in red. SEPs from non-coding RNAs are in the last column and the database used for their identification is also included.



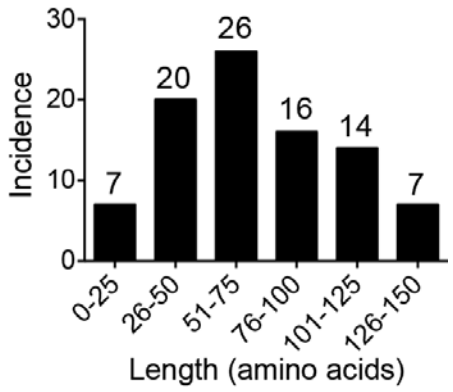
Supplementary Fig. 1. MS/MS spectra for a SEP tryptic peptide. The MS/MS spectra for GVGGQAALFAAGR was visually inspected to ensure sequence coverage that this peptide is unique to the sORF that encodes this SEP.



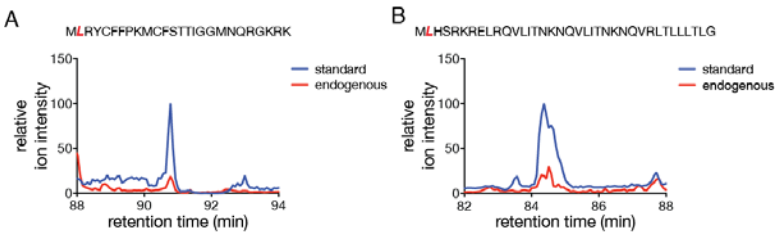
Supplementary Fig. 2. Length of SEPs and peptides identified by LC-MS/MS shows overlap between these data sets. (A) Length distribution for the actual peptides identified by LC-MS/MS (blue) and SEP length (red) using the RefSeq database. (B) Length distribution for the actual peptides identified by LC-MS/MS (blue) and SEP length (red) using the Ref-Seq and the RNA-seq database. The data shows that trypsin is necessary to generate peptides between 10-20 amino acids even for shorter SEPs.



Supplementary Fig. 3. SEP-encoding sequences are under stronger evolutionary selection than the introns of known coding genes. The curves show the cumulative distribution of sequence conservation levels calculated by SiPhy[4] across 29 mammalian species (Omega) in the exons of protein coding genes (red), the RefSeq sORFs producing SEPs (blue), sORF producing SEPs not in RefSeq (green) and introns of coding genes (black). Lower Omega scores reflect higher conservation. Only transcripts with a sufficient cross-species alignment support (branch length > 0.5) are included in the plot. 85% of SEP exons met this threshold, compared to 99% for known gene exons and 86% for known gene introns. The intron set was created by uniformly sampling a size matched intronic fragment from the intron neighboring each coding exon.



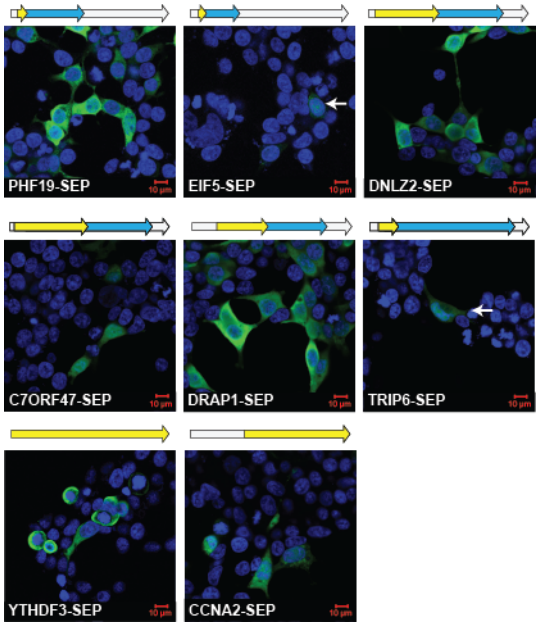
Supplementary Fig. 4. Length distribution for SEPs determined by defining sORF initiation sites the codon immediately 3’ of the upstream stop codon unless an AUG was present, in which case the upstream-most AUG was defined as the start.



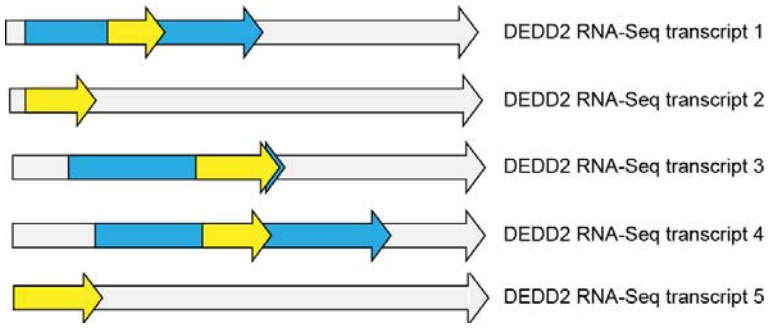
C

Sequence	Predicted Ion Mass	Observed Ion Mass
MLHSRKRELRQVLITNKNQVRLTLLLTLG	689.61525^{5+}	689.62^{5+}
MLHSRKRELRQVLITNKNQVRLTLLLTLG	691.62645^{5+}	691.62^{5+}
MLRYC*FFPKMC*FSTTIGGMNQRGKRK	643.72013^{5+}	643.72^{5+}
MLRYC*FFPKMC*FSTTIGGMNQRGKRK	645.73133^{5+}	645.73^{5+}

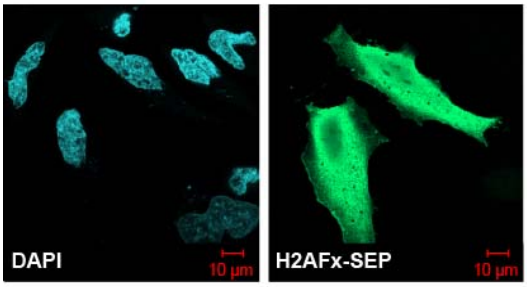
Supplementary Fig. 5. Confirmation of the presence of full-length SEPs in the K562 lysates by isotope-dilution mass spectrometry (IDMS). Full-length SEPs were synthesized by solid phase peptide synthesis and a deuterated leucine (d10-Leucine) was included to create a ‘heavy’-labeled SEP (red amino acid in sequences). Addition of these synthetic SEPs (blue lines) to K562 lysates enabled the identification of endogenous full-length SEPs (red lines) MLRYCFFPKMCFSTTIGGMNQRGKRK (A) and MLHSRKRELRQVLITNKNQVLITNKNQVRLTLLLTLG (B). The CID for these spectra was uninterruptable but these co-elution studies support the predicted full-length SEPs in the K562 lysates. (C) Predicted and observed masses for the ‘heavy’ standards and ‘light’ endogenous SEPs for the charge state detected.



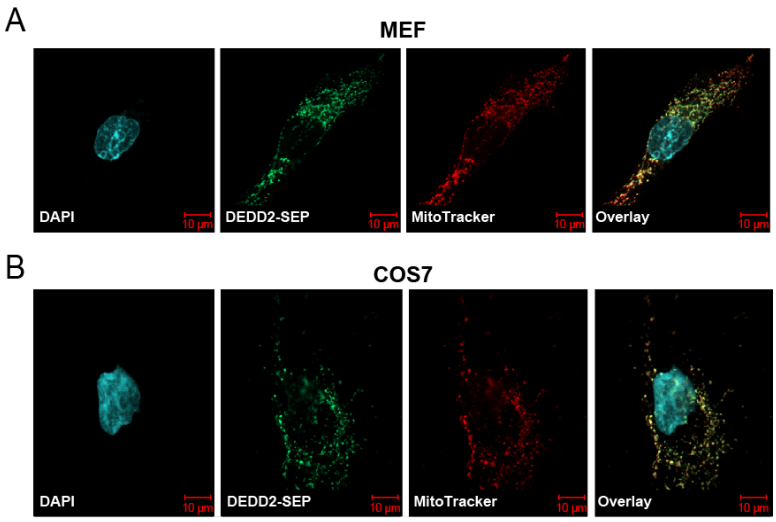
Supplementary Fig. 6. Validation of sORFs by SEP expression. Transient transfection of HEK293T cells with constructs containing a cDNA sequence corresponding to the full-length RefSeq mRNA (i.e., including the 5'- and 3'-UTRs). We appended a C-terminal FLAG-tag on the SEP coding sequence that could be detected by immunofluorescence. In these images the nuclei are stained with DAPI (blue) and the SEPs are detected with anti-FLAG antibody (green). PHF19-SEP, EIF5-SEP and DNLZ-SEP are all derived from sORFs in the 5'-UTR; cells expressing EIF5 are indicated with a white arrow. Additional 5'-UTR sORFs, C7ORF47-SEP, DRAP1-SEP, TRIP6-SEP, YTHD3-SEP and CCNA2-SEP produce SEPs but do not have an upstream in-frame AUG codon (i.e. these SEPs use non-AUG codons). (Note: RNA maps are not to scale. See **Supplementary Fig. 12** for lengths of the RNAs and sORFs.)



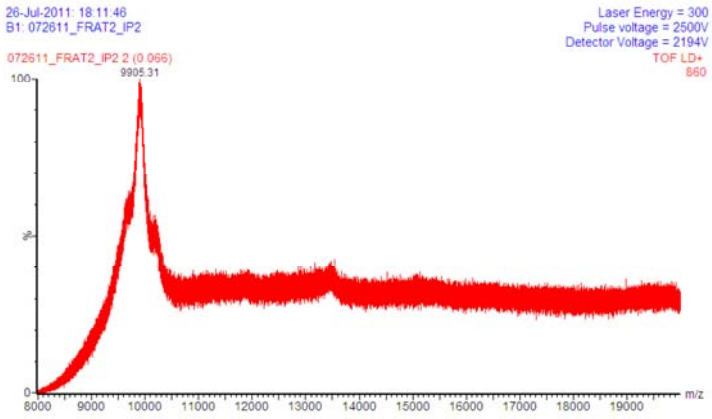
Supplementary Fig. 7. All DEDD2 RNAs detected in the K562 RNA-Seq data. DEDD2 RNA-Seq transcript 2 encodes the DEDD2-SEP. (Note: RNAs not to scale.)



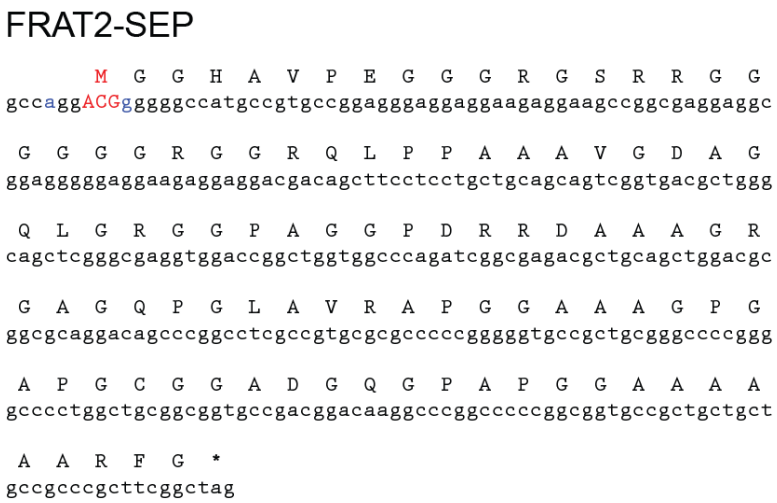
Supplementary Fig. 8. *H2AFx-SEP-FLAG* sORF expressed in HeLa cells, then detected with anti-FLAG antibody (followed by anti-mouse AlexaFluor 488, green).



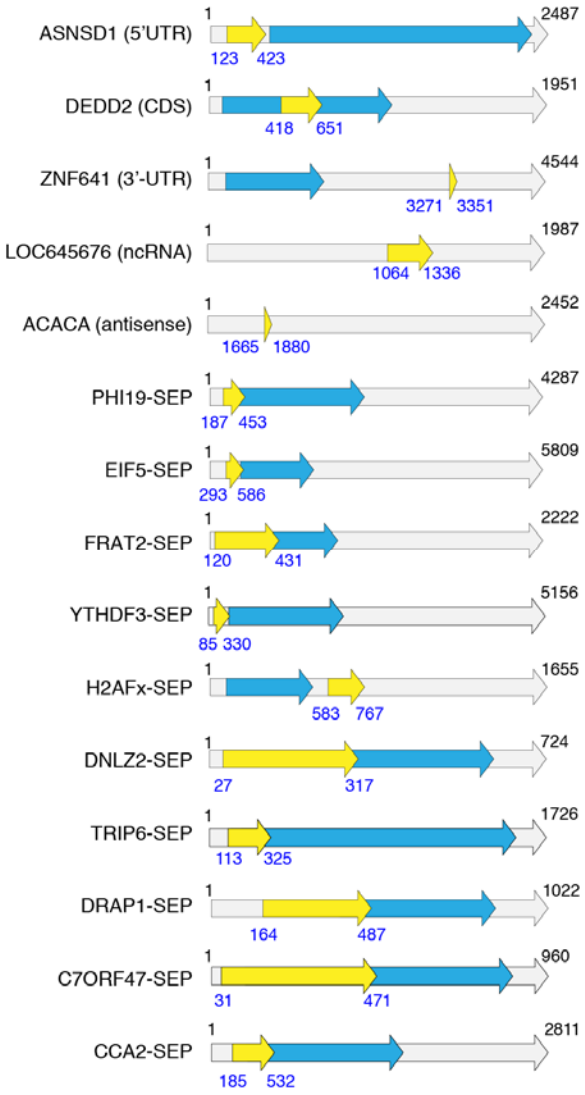
Supplementary Fig. 9. DEDD2-SEP-FLAG (green, detected by anti-FLAG immunofluorescence) is localized to the mitochondria in MEF (A) and COS7 (B) cells.



Supplementary Fig. 10. MALDI-MS of immunoprecipitated FRAT2-SEP-FLAG provides a polypeptide with a molecular weight of 9905, which identifies an ACG initiation codon with methionine as the first amino acid.



Supplementary Fig. 11. FRAT2-SEP sequence. An ACG triplet embedded in a Kozak consensus sequence was identified as the *FRAT2-SEP* initiation codon (red) by determining the molecular weight of immunoprecipitated FRAT2-SEP-FLAG using MALDI-MS.



Supplementary Fig. 12. Lengths of the SEP-encoding RNAs (black numbers) and sORFs (blue numbers).