



Discovery and Characterization of Novel smORF-Encoded Polypeptides (SEPs)

Citation

Ma, Jiao. 2016. Discovery and Characterization of Novel smORF-Encoded Polypeptides (SEPs). Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:26718718>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

Discovery and Characterization of Novel smORF- Encoded Polypeptides (SEPs)

A dissertation presented

by

Jiao Ma

to

The Department of Chemistry and Chemical Biology

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemistry

Harvard University

Cambridge, Massachusetts

December 2015

© 2015 – Jiao Ma

All rights reserved.

Discovery and Characterization of Novel smORF-Encoded Polypeptides (SEPs)

Abstract

Peptides and small proteins have essential physiological roles including metabolism (insulin), sleep (orexin), and stress (corticotropin-releasing hormone). Recent exploration of the human genome and proteome has revealed the existence of hundreds to thousands of short open reading frames (sORFs); however, the extent to which sORFs are translated into polypeptides is unknown. In line with the current convention, a protein-coding short ORF is defined to be a small ORF or smORF; the protein product as a smORF-encoded polypeptide is called a SEP; and a sORF or smORF upstream from an open reading frame (i.e. in the 5'-UTR) is called an upstream ORF or uORF. The identification of smORFs and SEPs have prompted efforts to determine the regulation and biological functions for these molecules. My thesis research focused on improving SEP discovery and the characterization of functional SEPs.

The discovery of novel SEPs contributes to our understanding of the composition of the human genome and proteome. My colleagues and I developed and utilized a proteogenomics strategy, which integrates genomics (RNA-Seq) with proteomics, to discover 86 novel human SEPs, the largest number of validated SEPs described at the time. Our findings indicated that SEPs are a large, unappreciated, peptide family. Moreover, our approach was far from optimized and we felt that there were likely many additional SEPs in the human genome. One goal of my

thesis work was to improve the SEP discovery methodology to find more human SEPs. My efforts led to the discovery of an over 300 SEPs in cell lines and human tissue.

A second goal of my thesis work was to identify and characterize functional SEPs. To do this I identified the SEPs that are most highly conserved throughout evolution with a program called PhyloCSF. PhyloCSF identifies which SEPs are evolutionary conserved to provide evidence for function. Seven out of the 300 plus SEPs had PhyloCSF scores that indicate that they have been conserved throughout evolution. These seven SEPs included an interesting SEP called SLC35A4-SEP that is generated from a uORF in the SLC35A4 gene. The SLC35A4-SEP had contained a transmembrane domain and analysis of cells revealed the mitochondrial localization of this SEP.

Further characterization of SCL35A4 indicated that this polypeptide interacts with members of the ATP synthase complex. Though this interaction requires further validation the putative interactions suggested a role for SLC35A4-SEP in cellular energetics. Overexpression or knockout of SLC35A4-SEP affected cellular respiration. Ongoing work is testing to see if SLC35A4-SEP also effects mitochondrial membrane potential and structure of ATP-synthase. More generally, this approach highlights how I can begin to identify functional SEPs using a combination of computational and experimental methods. And my work on another functional SEP called NoBody indicates that this strategy is general.

Acknowledgement

First, I would like to acknowledge Prof. Alan Saghatelian. His scientific expertise and kind support have made my Ph.D. years an incredibly fruitful and fun experience. Alan's guidance allowed me to grow into a better scientist; he taught me to think critically and pose interesting scientific questions, to be able to see the big picture while at the same time attending to the details. What I have learned in the past four and a half years will always stay with me as I continue to pursue a scientific career. I feel truly honored to have been given this opportunity to be a part of the Saghatelian lab, involved in cutting-edge research in the field of proteogenomics. I particularly enjoyed the collaborative, friendly and open atmosphere in the Saghatelian lab, which exposed me to various interesting projects through collaborations and led me to explore new techniques.

I would like to thank members of my Dissertation Advisory Committee, Prof. David Liu and Prof. Stuart Schreiber, for taking time out of their busy schedules to serve on my committee and provide me with scientific guidance throughout my Ph.D.

I also owe a great deal of thanks to Prof. Sarah Slavoff, Dr. Adam Schwaid and Dr. Tejia Zhang for teaching me about working in the lab, with great patience and guidance from the beginning of my thesis project in SEP discovery. And I would like to thank past Saghatelian lab members: Dr. Nawaporn "Yui" Vinayavekhin, Dr. Edwin Homan, Dr. Anna Mari Lone, Dr. Amanda McFedries, Prof. Arthur Tinoco, Prof. Yun-Gon Kim, Dr. Andrew Mitchell, for welcoming me into the lab and their indispensable help with different aspects of my projects, general techniques, and scientific discussions. I would like to thank Bogdan Budnik and John Neveu for advice on mass spectrometry and proteomics, Dr. Irwin Jungreis for sharing his

expertise in computational biology and our collaboration on PhyloCSF analysis, which he pioneered. I owe special thanks to Carl Ward, who joined the lab as an undergraduate student, who contributed a great deal to the SEP data analysis and generated in-house scripts that we still use as essential tools in the project today.

I feel extremely privileged to be able to have spent the last year and a half at the Salk Institute to complete my thesis work. I would like to thank my Salk colleagues in the Saghatelian lab, from whom I learn every day. I thank Joan Vaughan and Cindy Donaldson for their contributions to every aspect of my projects and their extensive expertise in peptide biology, which dramatically improved my SEP discovery platform. They helped greatly in setting up the new lab and made our transition smooth. The SEP subgroup, Dr. Qian Chu, Dr. Thomas Martinez, Dr. Meric Erikci Ertunc and Annie Rathore, provided me with excellent scientific conversations about various topics, both relevant and irrelevant to our research, which not only pushed forward the project greatly but made the day-to-day experience at work enjoyable. I owe special thanks to Dr. Jolene Diedrich and Dr. Max Shokhirev. Jolene helped us set up the mass spectrometry, and taught me a great deal about proteomics ranging from instrumentation to data analysis. Without her professionalism and patience, my thesis work would have been impossible. I would like to thank Max for sharing his expertise in bioinformatics, which has become an essential part in my research. I greatly enjoyed our hours of discussion on projects and Max's expertise and optimism toward tackling difficult scientific questions really advanced the project. For everyone I have encountered in my academic career or elsewhere, all my friends from across the world, I wish to thank them for all their support and, most importantly, their friendship.

I would also like to thank my American family, Joani Benoit, Vince Vonada, Frances Covey, Frances Vonada, and Ella Vonada for welcoming me to the U.S. when I arrived in Seattle

eight years ago and accepting me into their family; showing me the culture and sharing their love. I truly appreciate the long lasting friendship that they shared.

I will always be grateful to the many mentors in my life who helped with my initial transition to the United States and my undergraduate advisor Prof. Xiaosong Li who inspired me to pursue a graduate career and has always encouraged me and provided me with opportunities and guidance when I needed it the most.

Most importantly, I would not be here today without the hard work and sacrifices from my mother. I am forever indebted to her for everything she did and has continued to do for me. I hope I have made her proud.

Table of Contents

Abstract	
Acknowledgement	
Table of Contents	
List of Figures	
List of Tables	

Chapter 1. The Discovery and Characterization of A Novel Class of Functional Polypeptides

1.1. Introduction.....	2
1.2. Discovery of SEPs	4
1.3. SEP Characteristics.....	11
1.4. Mechanistic investigation	14
1.5. Future Directions	17
1.6. References.....	19

Chapter 2. Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue

2.1. Introduction.....	25
2.2. Results and Discussion	27
2.2.1. Impact of Different Workflows on SEP Discovery	27
2.2.2. Biological and Technical Replicates Increase the Number of SEPs Discovered	29
2.2.3. Using Targeted LC-MS/MS to Rapidly Validate Novel SEPs	34
2.2.4. Overview of 195 Newly Identified SEPs.....	38
2.2.5. SEPs Are Found in Additional Cell Lines and Some Show a Cell-Specific Distribution	43
2.2.6. SEPs Are in Human Tissue.....	46
2.3. Conclusions.....	47
2.4. Materials and Methods.....	48
2.4.1. Cell Culture.....	48
2.4.2. Tissue Sample	48

2.4.3. Peptidome Isolation from Cell Culture	49
2.4.4. Peptidome Isolation form Tissue	50
2.4.5. ERLIC Fractionation (20,21)	50
2.4.6. LC-MS/MS Analysis	51
2.4.7. Data Processing.....	51
2.4.8. RNA-seq Library Preparation and Transcriptome Assembly	52
2.4.9. Skyline Targeted MRM LC-MS/MS Peptidomics	53
2.5. References.....	54

Chapter 3. Improved Identification of Protien-Coding smORF by an Optimized Proteogenomics Platform

3.1. Introduction.....	58
3.2. Results and Discussions.....	60
3.2.1. Enrichment technique can improve number of SEPs detected	60
3.2.2. Isolating SEPs from cells	64
3.2.3. Assessing the impact of mass spectrometry parameters and SEP detection.....	66
3.2.4. Label-free SEP quantitation.....	68
3.2.5. Analysis of newly discovered SEPs.....	72
3.3. Conclusions.....	76
3.4. Materials and Methods.....	77
3.4.1. Cell Culture.....	77
3.4.2. Testing Various Methods for SEP Enrichment.....	78
3.4.3. Testing Different Methods for SEP Extraction.....	79
3.4.4. Digestion and Sample Preparation for LC-MS/MS	79
3.4.5. Q-Exactive LC-MS/MS analysis.	80
3.4.6. Orbitrap Fusion Tribrid LC-MS/MS Analysis.....	81
3.4.7. Data analysis to Identify Annotated and Non-annotated SEPs.....	82
3.4.8. Arsenite Treatment Experiments	83
3.4.9. Raising SLC35A4-SEP Antibody.....	84
3.4.10. Western Blot	85

3.4.11. PhyloCSF analysis of SEPs with Evolutionary Signature	85
3.5. References.....	86

Chapter 4. The Polypeptide NoBody Regulates Cellular P-body Number

4.1 Introduction.....	91
4.2. Results.....	92
4.2.1. Discovery and conservation analysis of NoBody	92
4.2.2. NoBody interacts with proteins involved in 5'-3' mRNA decay	94
4.2.3. Sequence dependence of the NoBody-EDC4 interaction	98
4.2.4. NoBody inversely regulates P-body numbers.....	102
4.2.5. NoBody action and impact on P-bodies.....	109
4.3. Discussion	113
4.4. Materials and Methods.....	115
4.4.1. Cell culture and transfection	115
4.4.2. Retrovirus production and transduction.....	116
4.4.3. Western blotting.....	116
4.4.4. Cloning and genetic constructs	117
4.4.5. Antibodies	118
4.4.6. Peptidomics and LOC550643 SEP identification.....	118
4.4.7. Conservation analysis	118
4.4.8. Co-immunoprecipitation and proteomics	119
4.4.9. NoBody photo-cross-linking.....	121
4.4.10. Microarray analysis and qRT-PCR.....	121
4.4.11. Gene Set Enrichment Analysis (GSEA)	122
4.4.12. Immunofluorescence.....	123
4.4.13. Confocal microscopy	123
4.5. References.....	124

Chapter 5. A Novel Mitochondrial SEP, SLC35A4-SEP

5.1. Introduction.....	131
5.2. Results and Discussion	132
5.2.1. SLC35A4 gene expression and conservation analysis.....	132
5.2.2. SLC35A4 SEP detection and cellular localization	134
5.2.3. SLC35A4 SEP enriches mitochondrial proteins involved in respiration chain.....	136
5.2.4. CRISPR/Cas9-mediated SLC35A4 SEP knockout and cellular RNA profiling	139
5.3.5. SLC35A4 SEP enhances mitochondrial respiration	141
5.3. Conclusion	144
5.4. Materials and Methods.....	145
5.4.1. Cell culture and transfection.....	145
5.4.2. Peptidomics and SLC35A4 SEP identification.....	145
5.4.3. Immunofluorescence and confocal microscopy.....	145
5.4.4. Co-immunoprecipitation and proteomics	146
5.4.5. Western blot.....	148
5.4.6. Knockout of SLC35A4 SEP by CRISPR/Cas9-mediated genome editing.....	149
5.4.7. RNA profiling and Gene Set Enrichment Analysis	149
5.4.8. Mitochondrial respiration assay.....	150
5.5. References.....	151

Appendix Chapter 1. Peptidomic Discovery of Short Open Reading Frame-Encoded Peptides in Human Cells

A 1.1. Introduction.....	156
A 1.2. Results.....	157
A 1.2.1. Discovering SEPs Encoded by Annotated Transcripts.....	157
A 1.2.2. SEPs are derived from Unannotated Transcripts	162
A 1.2.3. SEP Translation is Initiated at Non-AUG Codons.....	164
A 1.2.4. Supporting SEP length assignments	166
A 1.2.5. Cellular Concentrations of SEPs.....	167
A 1.2.6. Heterologous Expression of SEPs.....	169

A 1.2.7. SEPs Exhibit Subcellular Localization	171
A 1.2.8. Non-AUG Start Codons Enable Bicistronic Expression.....	172
A 1.2.9. A Small Subset of lincRNAs encode SEPs.....	175
A 1.3. Discussion	176
A 1.4. Methods.....	180
A 1.4.1. Cloning and mutagenesis	180
A 1.4.2. Cell culture.....	180
A 1.4.3. Isolation and processing of polypeptides	180
A 1.4.4. Offline electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) fractionation of polypeptide fraction	181
A 1.4.5. LC-MS/MS analysis.....	182
A 1.4.6. Data processing	182
A 1.4.7. RNA-Seq library preparation, alignment, and transcriptome assembly	183
A 1.4.8. Peptide synthesis, purification and concentration determination.....	184
A 1.4.9. SEP analysis by PAGE	184
A 1.4.10. Confirmation of the existence of full-length SEPs	185
A 1.4.11. Absolute quantification of SEPs	185
A 1.4.12. Imaging SEPs by immunofluorescence	186
A 1.4.13. Determinations of the FRAT2-SEP start codon by immunoprecipitation and MALDI-MS	187
A 1.4.14. Confirmation of the FRAT2-SEP initiation codon, Kozak sequence, and bicistronic expression by immunoblotting.....	188
A 1.4.15. Annotation of SEPs in Table A1.1	189
Appendix Chapter 2. Chemoproteomic Discovery of Cysteine Containing Human sORFs	
A 2.1. Introduction.....	196
A 2.2. Results and Discussion.....	197
A 2.2.1. Isolation of Cysteine Containing SEPs	197
A 2.2.2. Validation of Cysteine SEP Labeling	201
A 2.2.3. Novel ccSEPs.....	203
A 2.3. Conclusion	205

A 2.4. Methods.....	206
A 2.4.1. Cell culture.....	206
A 2.4.2. Isolation of polypeptides.....	206
A 2.4.3. MudPIT-LC-MS/MS analysis.....	206
A 2.4.4. MS data analysis	209
A 2.5. References.....	211

List of Figures

Chapter 1.

Figure 1.1. Schematic of proteins and s(m)ORF-Encoded Polypeptides (SEPs) expression

Figure 1.2. Discovery and characterization of Pri/Tal SEP

Figure 1.3. SEP discovery workflow applying both the ribosome profiling technique and mass spectrometry

Figure 1.4. SEPs detected in K562 cells that are derived from RefSeq transcripts and their locations

Figure 1.5. SEPs subcellular localization and their involvement in protein-protein interaction

Figure 1.6. SEP conservation

Chapter 2.

Figure 2.1. The workflows tested in the discovery of novel human SEPs

Figure 2.2. Biological and technical replicates lead to the discovery of novel SEPs

Figure 2.3. Total number of SEPs detected in K562 cells using PAGE+LC-MS/MS workflow after performing an additional six technical replicates

Figure 2.4. Alignment for ASNSD1-SEP shows protein-coding signature

Figure 2.5. Validating SEPs with targeted mass spectrometry

Figure 2.6. MS/MS spectra for PRR3-SEP

Figure 2.7. Overview of 195 novel SEPs identified in K562 cells

Figure 2.8. The characteristics of the 36 SEPs in K562 cells validated by Skyline-MRM

Figure 2.9. SEP derived from MCF10A and MDAMB231 cell lines

Figure 2.10. The Characteristics of SEPs detected in MCF10 and MDAMB231 cell lines

Figure 2.11. Discovery of 25 tumor derived SEPs

Chapter 3.

Figure 3.1. SEP discovery workflow and enrichment techniques

Figure 3.2. Comparison among different extraction methods

Figure 3.3. Comparison of MS/MS spectra acquired by different instruments and different instrument settings

Figure 3.4. Changes in SEP and protein expression level upon arsenite treatment

Figure 3.5. SEPs can be further sub-categorized based on their sequences

Chapter 4.

Figure 4.1. LOC550643/LINC01420/NoBody genomic locus and expression

Figure 4.2. The *LOC550643* gene encodes the NoBody peptide in a short open reading frame (sORF)

Figure 4.3. Peptidomic identification of NoBody in K562, HEK293T, and MDA-MB-231 cells

Figure 4.4. NoBody immunoprecipitation enriches a complex of proteins involved in 5'-3' mRNA decay

Figure 4.5. Sequence and domain dependence of the NoBody-EDC4 interaction

Figure 4.6. Photo-cross-linking evidence for direct physical interaction of NoBody with EDC4

Figure 4.7. NoBody dissociates P-bodies via interaction with EDC4, and absence of NoBody increases P-body numbers

Figure 4.8. NoBody disrupts P-bodies and GFP-Dcp2 P-body localization

Figure 4.9. P-body dissociation requires translated NoBody peptide, and does not depend on the *LOC550643* gene

Figure 4.10. Increased P-body numbers are a specific effect of *LOC550643* silencing

Figure 4.11. NoBody co-localizes with P-bodies in a subpopulation of cells at very low expression levels

Figure 4.12. GSEA match for Dcp2 silencing

Figure 4.13. GSEA matches and top gene changes for NoBody silencing

Chapter 5.

Figure 5.1. SLC35A4 gene structure, expression and conservation

Figure 5.2. SLC35A4 SEP detection and localization

Figure 5.3. SLC35A4 SEP enriches mitochondrial proteins

Figure 5.4. SLC35A4 SEP knockout by CRISPR/Cas9-mediated genome editing

Figure 5.5. Effect of SLC35A4 SEP expression on the cellular oxygen consumption rate

Figure 5.6. Western blot analysis of mitochondrial respiratory chain proteins

Appendix Chapter 1.

Figure A1.1. Discovering SEPs

Figure A1.2. Overview of SEPs

Figure A1.3. SEP-encoding sequences are under stronger evolutionary selection than the introns of known coding genes

Figure A1.4. Length distribution for SEPs

Figure A1.5. Confirmation of the presence of full-length SEPs in the K562 lysates by isotope-dilution mass spectrometry (IDMS)

Figure A1.6. SEP quantitation

Figure A1.7. Expression of SEPs

Figure A1.8. All DEDD2 RNAs detected in K562 RNA-seq data

Figure A1.9. H2AFx-SEP-FALG sORF expressed in HeLa cells

Figure A1.10. MALDI-MS of immunoprecipitated FRAT2-SEP-FLAG

Figure A1.11. FRAT2-SEP sequence

Figure A1.12. Characterization of the non-AUG initiation codon of the FRAT2-SEP sORF

Appendix Chapter 2.

Figure A2.1. Workflow for identifying ccSEPs

Figure A2.2. Validation of site of labeling and cellular expression of newly discovered ccSEPs

Figure A2.3. TSP SEP expression was confirmed by western blot

Figure A2.4. ccSEP overview

Figure A2.5. Workflow for isolation, enrichment and orthogonal proteolysis of cysteine SEPs from K562 lysates for LC-MS/MS identification

List of Tables

Table 2.1. A list of 36 SEPs detected in K562 cells that were validated by Skyline-MRM

Table 2.2. Total number of SEPs discovered from K562, MCF10A, MDAMB231 and tumor samples

Table 3.1. Full list of 37 non-Uniprot SEPs detected

Table 5.1. A full list of 106 identified proteins enriched in SLC35A4 SEP-FLAG immunoprecipitated samples

Table A1.1. Full list of identified SEPs

Table A1.2. Quantification of SEP trypsin peptides by IDMS

Table A2.1. Detected peptides and the start codon and length of their corresponding ccSEPs

Table A3.1. Full list of detected SEPs in Chapter 2

Table A3.2. Full list of detected SEPs in Chapter 3

Chapter 1

The Discovery and Characterization of A Novel Class of Functional Polypeptides

This chapter was adapted from: Chu Q, Ma J, Saghatelian A. Identification and characterization of sORF-encoded polypeptides. *Crit Rev Biochem Mol Biol.* 2015;50(2):134-41

1.1. Introduction

Studies over the past few decades have revealed several unprecedented classes of biologically active molecules in the genome with regulatory roles in diverse physiological processes (1-4). In particular, recent work has revealed a novel class of bioactive peptides in a variety of organisms that are derived from short open reading frames (sORFs) or small open reading frames (smORFs) (5-7) (Figure 1.1).

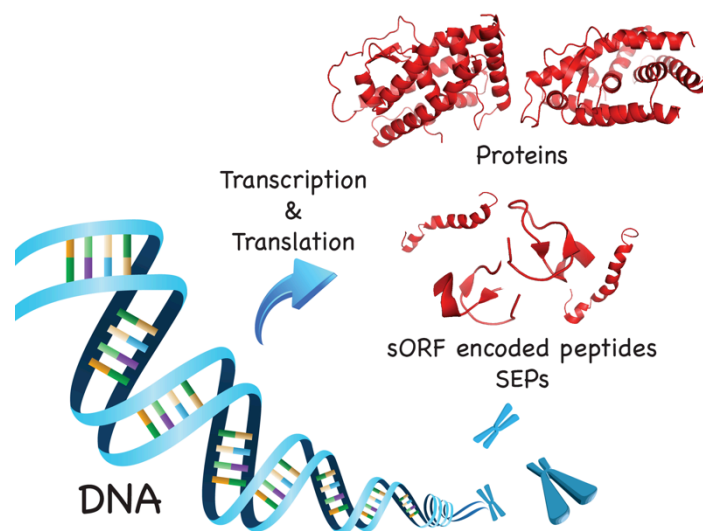


Figure 1.1. Schematic of proteins and s(m)ORF-Encoded Polypeptides (SEPs) expression.

Unlike classical peptide hormones and neuropeptides, which are translated as larger precursor proteins followed by limited proteolytic processing (8, 9), these s(m)ORF-encoded polypeptides (SEPs) are short peptides encoded directly from s(m)ORFs (10-13). A small number of well-studied SEPs have indicated that these polypeptides may act as important regulators in many fundamental biological processes, such as metabolism (14), development (11, 13), and cell death (15), but little is known about the biological activities, regulation, or even total number of SEPs. Therefore, discovery and functional characterization of SEPs will expand

our knowledge of the composition of the genome and proteome, and provide fundamental insights into the molecular biology of cells.

Although satisfactory classifications for these small peptides have not been clearly defined, we consider SEPs as generally less than 150 amino acids in length because we have found a number of non-annotated SEPs in this length range (6, 7). The small size of SEPs hampers their discovery by both computational and experimental approaches (16, 17). On the one hand, it is challenging to apply bioinformatics methods to predict expression from s(m)ORFs by simply grafting widely-used gene prediction algorithms for large proteins. These programs usually assess coding potential of ORFs (i.e. protein-coding regions) by a number of stringent criteria, which recognize certain patterns in the transcripts, such as promoter sequences, polyadenylation signals, AUG-start codon usage, and sequence conservation (18-20). However, these features are not as rich in s(m)ORFs as in long protein coding genes, resulting in a high false positive rate to distinguish between coding and non-coding s(m)ORFs. On the other hand, non-annotated SEPs are not in standard proteomics databases and therefore cannot be discovered by direct detection. Even for SEPs that are in these databases, their small size and lower abundance make them more difficult to detect. In addition, many recently identified SEPs are derived from s(m)ORFs with non-AUG codons, which makes the identification process even more challenging (21, 22). With improved strategies in computational approach, proteomics and next-generation sequencing, there have been great advances to address these challenges, and this has resulted in identification and validation of hundreds of new SEPs.

In this review, we describe various strategies to discover and identify SEPs, and overview several characteristics of SEPs based on recent proteomics results in the K562 human leukemia

cell line. In addition, for SEPs that are discovered through these global identification strategies, we will discuss functional approaches to investigate the biology of these polypeptides.

1.2. Discovery of SEPs

As mentioned, SEPs have been found in several different organisms, including bacteria (23), plants (24-26), yeast (27), worms (28), flies (10, 12) and humans (15). Screening studies looking for key regulators of certain phenotypes resulted in serendipitous discovery of a few short bioactive peptides encoded directly by s(m)ORFs, which revealed the existence of this class of non-annotated genes.

The discovery of a SEP in *E. coli*, for example, began with research aimed at understanding the role of the sugar transport small RNA (SgrS), a non-coding RNA. The expression of SgrS is inversely correlated with glucose flux into the cell and SgrS was shown to operate through an RNA-dependent mechanism. The 3' end of the SgrS RNA sequence is able to bind *ptsG* mRNA that encodes the major *E. coli* glucose transporter and inhibit translation of this protein, but the role of the 5' end of the SgrS sequence, upstream of the nucleotides involved in base pairing with the *ptsG* mRNA, remained a mystery (29, 30). Recent work revealed that the 5' part of SgrS encodes a SEP, a 43-amino acid peptide referred to as SgrT, which inhibits glucose influx by directly binding and inhibiting the glucose transporter and therefore plays a central role in cellular metabolism (23).

Work in flies has also revealed important SEPs. The discovery of the shortest known SEP, the 11-amino acid peptide encoded by the polished rice (*pri*) gene, also began with studies that looked for regulators responsible for developmental defects in *Drosophila* legs. As a result, the

pri gene was identified from a transcript that had been previously classified as a putative non-coding RNA (10, 31) (Figure 1.2).

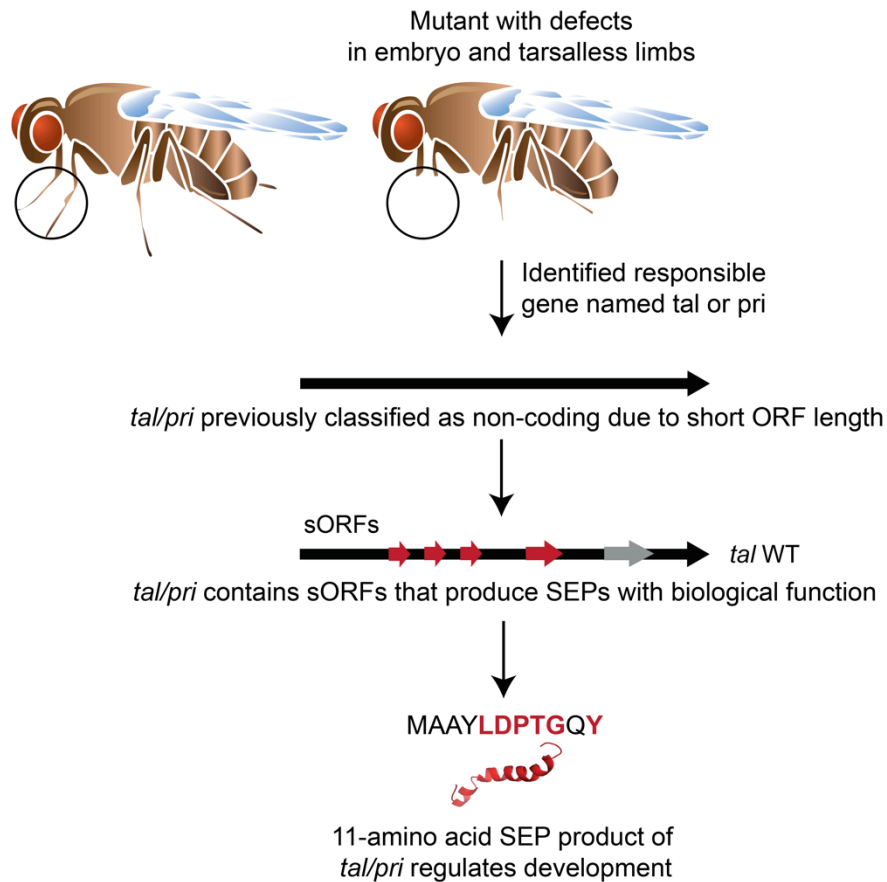


Figure 1.2. Discovery and characterization of Pri/Tal SEP. *Polished rice (pri)* /*Tarsal-less (tal)* transcript was initially classified as non-coding RNA. It contains several open reading frames (ORFs) smaller than 50 amino acids in length. An ORF coding for an 11 amino acid-long peptide has a highly conserved motif (amino acids labeled in red) and mediates the function of the gene. Deletion of this ORF leads to abnormal differentiation of *Drosophila* legs. This is one example that demonstrates: 1) SEPs can arise from a polycistronic messenger or a putative non-coding RNA; 2) SEPs can have crucial biological function such as epidermal differentiation in this case.

The Pri SEP has been demonstrated to trigger N-terminal truncation of a transcription factor, Shavenbaby (Svb). Upon cleavage, Svb converts from a repressor to an activator, and in turn contributes to epidermal differentiation in *Drosophila* (11). Moreover, subsequent

phylogenetic analysis revealed that *pri* belongs to an ancient gene family that can be traced back at least 440 million years. This nucleotide-level conservation indicates that the functional role of this SEP has probably been conserved through insect evolution as well (10).

Functional SEPs are also present in humans. Humanin was discovered during a screen for cDNAs from the nervous system that prevented cell death by the amyloid precursor protein (APP) in an effort to discover new genes that could prevent Alzheimer's disease. In this work, a plasmid containing a cDNA library was introduced into mammalian cells and these cells were then induced to produce APP, which leads to the death of most cells. Plasmids from surviving cells were then isolated, amplified, and the screen repeated an additional four times. One cDNA was particularly effective at preventing apoptosis in this screen, and further analysis revealed that the functional element of this gene was a 75-bp open reading frame (ORF) encoding a 24-amino acid peptide, which was named humanin. The RNA that encodes humanin is the mitochondrial 16s ribosomal RNA, which was thought to be a non-coding RNA (15). Subsequent work reported that humanin prevents cell death via a protein-protein interaction (PPI) with Bcl-2-associated X protein (Bax) that prevents Bax activation (32, 33).

The serendipitous discovery of these SEPs suggests that translation of s(m)ORFs and production of small bioactive peptides in the proteome are much more pervasive than previously appreciated. s(m)ORFs are a common feature in the genome, located mostly within sequences of non-coding RNAs and 5'-UTR of protein coding genes (34, 35). The coding potential of s(m)ORFs has previously been neglected mainly due to the difficulty in distinguishing them from non-coding genes by prevalent bioinformatics methods, as well as the challenge to detect their translation products from a bulk of peptide sequences though current proteomics approaches (36, 37). With the purpose of looking for new SEPs that are biologically relevant,

extensive efforts have been performed to re-evaluate the coding possibility of s(m)ORFs by using improved computational and experimental strategies. For example, Pauli and colleagues revisited the zebrafish transcripts by integrating ribosome profiling data, and identified 700 novel protein-coding transcripts from non-annotated translated ORFs, of which 81% were conserved in other vertebrates. Among these new ORFs, 28 secreted peptides were further isolated which contained putative signal sequences but lacked the predicted transmembrane domains. Follow-up studies with one of these s(m)ORFs, referred to as *toddler*, indicated that it is able to produce a 58-amino acid polypeptide. The toddler SEP was demonstrated to activate G-protein-coupled signaling by binding to the APJ/Apelin receptor and consequently promote cell movement during zebrafish gastrulation (13).

The second example involves the discovery of two functional SEPs of 28 and 29 amino acids that are encoded by putative non-coding RNA *pncr003:2L*. Inspired from the evidence that the *pri* s(m)ORF was initially misannotated as a non-coding RNA, a pool of all polyadenylated, polysome-associated putative non-coding RNA in *Drosophila* was re-examined using an improved bioinformatics approach, which resulted in two s(m)ORFs, *pncr003:2L*, driven by strong Kozak sequences. Subsequent studies indicated that *pncr003:2L* peptides are expressed in somatic and cardiac muscles, where they regulate muscle contraction by modulating Ca^{2+} trafficking. More importantly, these two SEPs were conserved across evolution as they showed structural and functional homology with s(m)ORFs *sarcophilin (sln)* and *phospholamban (pln)* in vertebrates, where they play a role in regulating Ca^{2+} transport into ER following muscle contraction (12).

These examples along with recent computational analyses have suggested that many s(m)ORFs are translated, however, their coding potential in general has not been interrogated

systematically and experimental evidence that these s(m)ORFs are able to generate stable polypeptides is still lacking (38). One reasonable strategy to address this issue is to take advantage of ribosome profiling analyses (39, 40) (Figure 1.3).

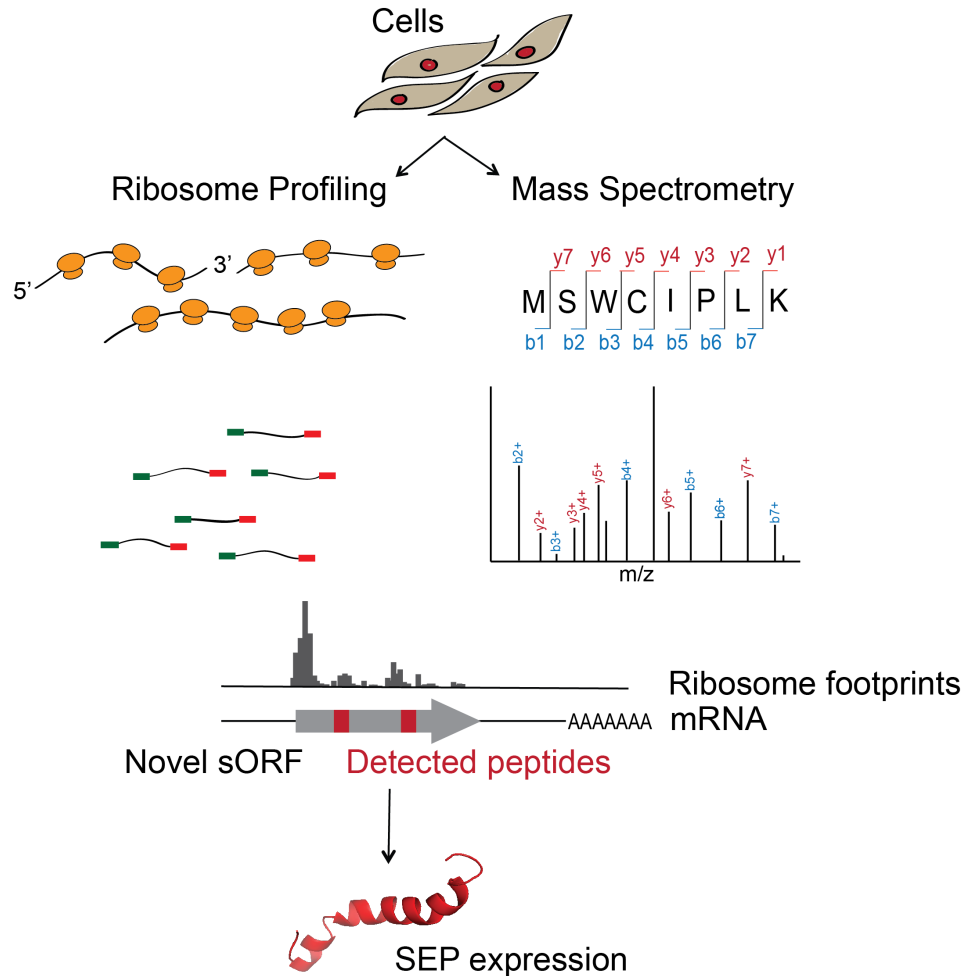


Figure 1.3. SEP discovery workflow applying both the ribosome profiling technique and mass spectrometry. mRNA transcripts are isolated from cells and followed by a ribosome profiling experiment. The ribosome footprints are mapped to the genome and *de novo* assembled into transcripts, leading to the discovery of novel s(m)ORFs with coding potential. Simultaneously, the small proteome is extracted from the cell lysate and analyzed by mass spectrometry. Detected peptides are mapped to the transcriptome from which the peptides are translated. Ribosome profiling and mass spectrometry analysis together lead to the discovery of novel genes coding for SEPs.

Ribosome profiling is an emerging sequencing technique that can provide a global snapshot of all mRNAs being actively translated (41, 42). Whereas RNA-seq sequences all of the transcripts present in a sample; ribosome profiling sequences ribosome-protected mRNA fragments that map back to the genome and *de novo* assemble into transcripts. The combination of these techniques enables the identification of translation start sites and alternative splicing forms to provide a more comprehensive coverage of the transcriptome. In addition, the amount of normalized mRNA reads in sequencing results is proportional to ribosome distribution and density on corresponding transcripts, which could provide quantitative information about translation as well (43). Therefore, ribosome profiling has become increasingly popular to assess coding potential of ORFs in the genome, especially those previously unannotated or considered as non-coding RNAs (40). As a result, plenty of new transcripts have been identified which consist of either novel exons, alternative initiation and splicing sites of annotated genes or entire new ORFs that had been classified as non-coding regions in the transcriptome, including 5'UTRs and long non-coding RNAs (lncRNAs) (44, 45). Among them, quite a few are s(m)ORFs that encode for SEPs.

Ribosome profiling has revealed that there could be pervasive translation of s(m)ORFs outside annotated protein coding genes (45). First, scanning 40S ribosomal subunits and other non-specific protein binders, as well as non-productive binding to single ribosomes could contribute to footprints but not translation (46). Second, different interpretations of ribosome profiling data, especially for non-annotated transcripts, could result in completely opposite conclusions (47). Third, improved ribosome profiling approaches need to be developed specifically for s(m)ORF analyses, since s(m)ORFs that encode SEPs are much shorter and smaller in size, potentially making traditional ribosome profiling less suitable (48). Therefore,

demonstrating protein-coding potential of s(m)ORFs by detecting experimentally their stable protein products becomes extremely necessary and important.

Recent advances in mass spectrometry (MS) proteomics provide a powerful tool to discover SEPs from cell and tissue lysates (49). These MS experiments differ from ribosome profiling because they are able to detect polypeptides translated from s(m)ORFs and thereby validate the protein-coding potential of the s(m)ORF (Figure 1.3). For example, Oyama and colleagues developed a proteomics approach in an attempt to discover novel (non-annotated) coding sequences (CDS) in mammalian cells. The key step in their approach was the generation of their own protein database through the ‘6-way translation’ of annotated RNA sequences in the RefSeq database. By using the entire RNA sequences instead of just those regions thought to be coding, this protein database included any proteins found in 5’UTRs or 3’UTRs, as well as identified frame shifted variants of known genes. This approach identifies peptides and proteins that would be missed by traditional proteomics experiments that rely on annotated protein genes. As a result, a total of four SEPs have been discovered in K562 human leukemia and HEK293 cell lines with a length distribution of 88-148 amino acids (50). To improve on these results, we utilized next-generation RNA sequencing (RNA-Seq) to identify all possible protein-coding mRNA transcripts, including non-annotated transcripts (i.e. transcripts that exist but are not in the NCBI RefSeq database). The RNA-Seq data was translated into all possible reading frames to create a database that is expected to contain all of the polypeptide sequences that could theoretically be produced in the cell. Using this database, we identified more than 300 additional human SEPs from several mammalian cell lines (K562, MCF10A, MDAMB231 cells) as well as human tissue samples, demonstrating the prevalence of this class of polypeptides. In addition, a

few SEPs were detected in every sample we analyzed, which indicates that SEPs might be ubiquitous and serve fundamental (i.e. housekeeping) roles (6, 7).

1.3. SEP Characteristics

So far, we have identified 285 novel SEPs from the K562 human leukemia cell line through the peptidomic approach discussed above. In order to obtain a better understanding of SEPs, we examined several global properties of the molecules we discovered, such as length distribution, start codon usage as well as locations in the genome, and found that they possess many unique characteristics.

Our proteomics analysis using trypsin-digested samples detects small pieces of peptide fragments from a bulk of peptide samples, however, it is not able to obtain full protein-level SEP sequence coverage and in particular the N- and C-terminus of SEPs are missing in most cases. Therefore, we have to assign start and stop codons for each SEP in the corresponding s(m)ORF to determine its length. For example, we assigned the first downstream in-frame stop codons in SEP-encoding s(m)ORFs as stop sites. Likewise, the closest upstream in-frame AUG was assumed to be the start codon. If no upstream AUG was present, the initiation codon was considered to be an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (51). The near-cognate codon usually has a single nucleotide difference from AUG (e.g. CUG), which has been shown as a translation-initiating site in ribosome profiling experiments. In a few cases, neither of these conditions was met and the codon immediately following an upstream stop codon was defined as the start site. By doing this, we intended to determine the maximum SEP length, while not biasing the analysis toward shorter SEP lengths. Using this approach, we determined the SEPs to range between 8 and 149 amino acids in length,

with the majority (>90%) being smaller than 100 amino acids long. In particular, we found that the SEP length could be fitted into a Gaussian distribution with a population centroid between 26-50 amino acids (6, 7).

Another interesting feature of SEPs involves the preponderance of non-canonical translation start sites. Two thirds of the detected SEPs (190/285) do not initiate at AUG codons, among which 56 SEPs start translation from a near-cognate non-AUG codon. These near-cognate codons differ from AUG by a single base and in most cases are embedded within a Kozak sequence, which increases translation initiation. Several lines of evidence indicate that SEPs can be translated using non-AUG start codons. First, ribosome profiling has revealed liberal use of non-AUG initiation (52). Second, transient expression of non-AUG containing SEPs in HEK293 cells produced full-length SEPs, which verifies that s(m)ORFs with non-AUG start codons are translated. Third, we validated the start codon for a single SEP as well as the requirement of a Kozak sequence by site-directed mutagenesis. In this experiment, we found that an ACG start codon was used instead of an AUG. Mutation of the ACG to an AUG resulted in increased translation, while mutation of the Kozak sequence inhibited translation, demonstrating the requirement for the Kozak sequence to be present for the initiation at ACG (6). Together these data provided strong evidence that SEPs can be produced from non-AUG, near cognate, initiation codons.

Next, we analyzed locations of these SEP-encoding s(m)ORFs in the genome and found that more than 70% (201/285) SEP producing RNA transcripts are not annotated in the RefSeq database, which highlights the importance of the custom protein database derived from RNA-Seq data in our approach and indicates that our strategy is able to provide superior coverage of small SEPs in the entire genome. Also the remaining 84 SEPs that are encoded from annotated 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 (Figure 1.4). The locations of these s(m)ORFs mirror the distribution obtained from ribosome profiling indicating that our proteomics coverage achieves the necessary breadth and depth to reveal global properties of s(m)ORFs (40, 53).

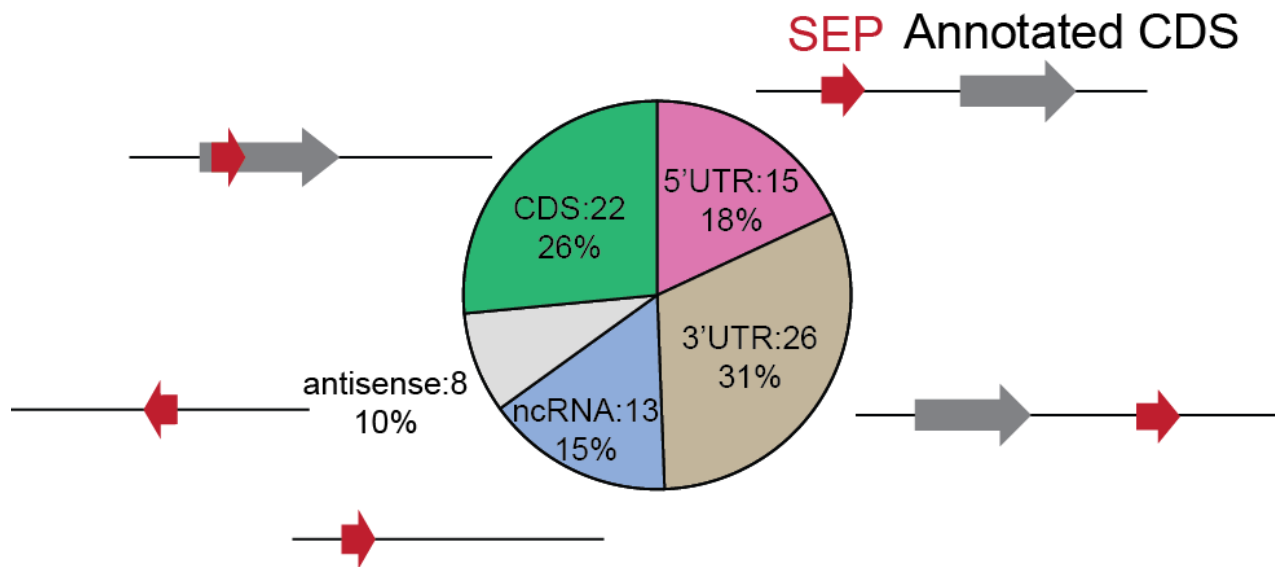


Figure 1.4. SEPs detected in K562 cells that are derived from RefSeq transcripts and their locations. The s(m)ORFs are found on coding RNAs at the 5'UTR, 3'UTR and CDS and on non-coding RNAs. Gray arrows represent annotated protein coding ORFs, and red arrows represent s(m)ORF encoding SEPs and their relative locations to the annotated ORFs.

Furthermore, by carefully examining SEP producing transcripts, we noticed that several SEPs are translated from alternative splicing of annotated protein coding genes. For example, the DEDD2-SEP encoding s(m)ORF is a frameshifted sequence within the main CDS of the DEDD2 transcript, which normally couldn't be translated according to a traditional ribosome scanning mechanism. It turns out that the DEDD2 RNA has two splicing variants. One is for

full-length DEDD2 protein expression and the other is a truncated mRNA sequence wherein the first start codon is that of the DEDD2-SEP s(m)ORF (6). Therefore, alternative splicing of annotated protein coding genes is one of the SEP producing mechanisms, which might be functionally relevant as well since higher organisms are able to create multiple functions from single genes for genetic efficiency.

Recent studies on aminoacyl tRNA synthetases (AARSs) indicated that alternative splicing events occur in all human AARSs, which retain only noncatalytic domains by splicing out catalytic domains at the mRNA level. Some of these catalytically inactive splice variants fall into the SEP regime in terms of their small size. Interestingly, each of the AARS variants demonstrated novel regulatory activities spanning a variety of physiological processes, including cell differentiation and proliferation, transcriptional regulation, and inflammatory response, which are distinct from the original aminoacylation function (54). In addition to the fact that alternative splicing is able to produce functional SEPs, SEPs can also be generated *via* polycistronic translation of a given RNA transcript. For example, the aforementioned *pri (tal)* mRNA is a polycistronic transcript with four individual s(m)ORFs, which produces three 11-aa peptides and one 32-aa peptide with a conserved LDPTGX_Y motif in all of them. The expression of four SEP isoforms has been verified *in vitro* and *in vivo* and all of them exert the same biological functions (10) (Figure 1.2).

1.4. Mechanistic investigation

There are some SEPs with demonstrated biological activities. The majority of bioactive SEPs were discovered as a consequence of phenotypic screens. These unbiased methods were able to identify SEPs that could affect a phenotype. A number of approaches have been taken to

understand the molecular mechanisms underlying the biological functions of these SEPs. The functional characterization and mechanistic investigation of SEPs will provide valuable information about SEP biology. If the SEP is known to contribute to a certain phenotype, it may be connected to other known regulators that lead to the same phenotype. For example, the 11 amino acid-long Pri SEP has been shown to play a role in *Drosophila* embryogenesis, as embryos that lack *pri* gene display prominent defects in trichome formation, though the molecular mechanism was not clear at that time. It is known that a transcription factor Shavenbaby (Svb) is responsible for trichome formation as well as direct regulation of downstream gene expression. So it is likely that Pri SEP works together with Svb in *Drosophila* embryo development. Indeed, studies on a *pri* loss-of-function mutant indicated that *pri* is specifically required for the expression of Svb regulating genes, and subsequent work revealed that Pri SEP exerts its function by cleavage of the N-terminus of the Svb protein, which converts it from a transcription activator to a repressor (11). Therefore, study of potential links between SEPs and other regulators that share the same phenotype is an efficient strategy to investigate SEP functions at the molecular level.

Another strategy is based on the hypothesis that SEPs exert their biological function through interacting with other proteins or biomolecules. Therefore, if the binding protein(s) could be identified, we could use this information to develop and test SEP functions according to the known roles of the interaction partners. For example, co-immunoprecipitation of MRI-2-SEP from HEK293 cells has yielded two proteins, Ku70 and Ku80, which are heterodimeric proteins involving in the non-homologous end joining (NHEJ) pathway of DNA double-strand break (DSB) repair. Therefore, we hypothesized that MRI-2-SEP may also play a role in NHEJ pathway through protein-protein interactions with the Ku70/Ku80 heterodimer. Subsequent

studies confirmed this assumption and showed that MRI-2 accumulates in the nucleus upon Ku overexpression and induction of DSBs, and enhances the rate of NHEJ *in vitro* (55) (Figure 1.5). Last but not least, gene expression profiling is also a useful tool to analyze and elucidate SEP functions. This strategy provides global gene expression patterns for control cells and cells with overexpression and/or knockdown of a given SEP. Any gene or gene sets that exhibit different expression levels between the two conditions might be regulated by the SEP. In particular, if the same genes change in overexpression and knockdown studies, this would indicate that SEP regulation of the genes is specific and proteins encoded by these genes may have potential roles relevant to SEP activity.

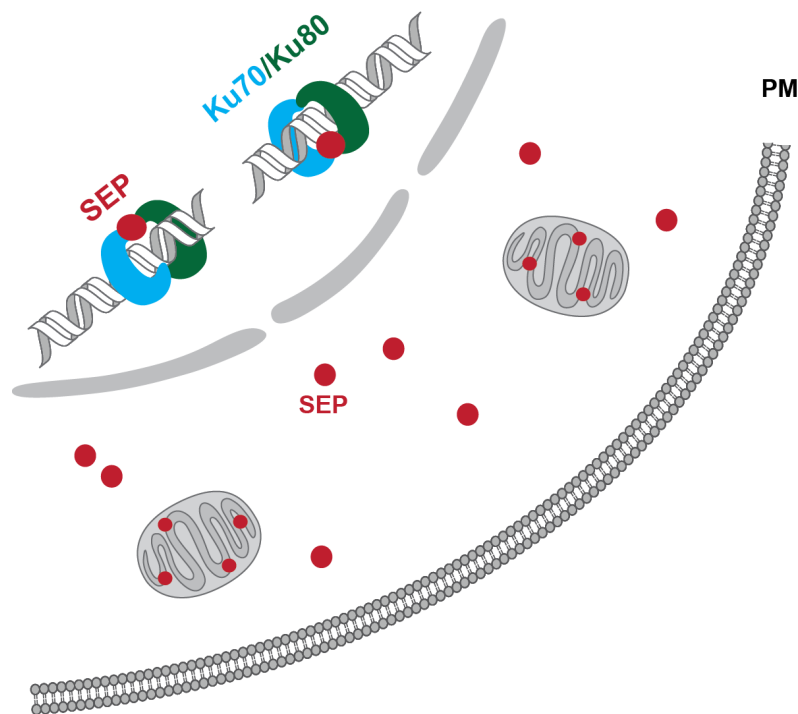


Figure 1.5. SEPs subcellular localization and their involvement in protein-protein interaction. SEPs are expressed at various subcellular locations such as in the nucleus, mitochondria, as well as the cytosol, suggesting that SEPs can have molecular activities in the cell. In one example (Slavoff *et al.*), a SEP participates in protein-protein interaction with Ku70/Ku80 complex and plays a regulatory role in the DNA non-homologous end joining (NHEJ) repair process.

1.5. Future Directions

The discovery of biological active SEPs indicates that the human proteome is significantly more complex than previously appreciated. As an emerging field of research, we believe that we have only begun to explore the breadth and diversity of this exciting new family of polypeptides. Future efforts will move towards greater SEP detection and better understanding of their roles in a variety of biological processes. In particular, several questions need to be addressed.

How many SEPs in total are in the proteome? Current computational and experimental approaches have identified hundreds of SEPs, however, due to the limitation of each approach, it is entirely possible that there are many more as-yet-undiscovered polypeptides encoded from s(m)ORFs. Therefore, improved strategies are required to detect more SEPs. For example, can more accurate algorithms for s(m)ORF identification and coding potential prediction be developed? And can proteomics be improved for better peptide detection? In addition, most of the previous studies focused on intracellular SEPs. However, secreted SEPs may be equally important in terms of their regulatory roles as many bioactive peptides are secreted and act as signaling molecules to trigger a variety of biological activities (56, 57).

In addition to improve detection, quantitative methods for SEP levels also need to be optimized. Can different SEP levels among cells and tissues in different biological states be measured? One challenge with SEPs is that they are shorter and less abundant than average proteins so fewer peptides are detected by proteomics and their detection can be stochastic (detection in 25-50% of all runs). After improvement of detection sensitivity, the next step will be to start to quantify SEPs between biological samples (e.g. disease versus normal tissue) to determine which SEPs are important in different biological processes. In addition, quantitative

analysis can be coupled to anatomical SEP profiling to help identify SEPs with tissue-specific functions (58, 59). In particular, the distinct expression of SEPs in disease cells/tissues may eventually impact medicine and human health by revealing novel pathways for pathogenesis.

Current studies have revealed a few unique features of SEPs, such as pervasive expression starting from non-AUG codons, but a more comprehensive knowledge of SEP functions— biochemical, cellular and physiological— will help reveal any general roles for these polypeptides. For example, the combination of ribosome profiling with proteomics can help define the exact boundaries and start codons for the s(m)ORFs that produce SEPs. If certain SEPs that are co-regulated, for example, share a non-AUG start codon that might suggest a more global method for their regulation. Moreover, it is unlikely that all SEPs are biologically active as some may represent translational noise. Conservation analysis is a useful tool to predict active SEPs but it may not be the only way (Figure 1.6).

ASNSD1-SEP

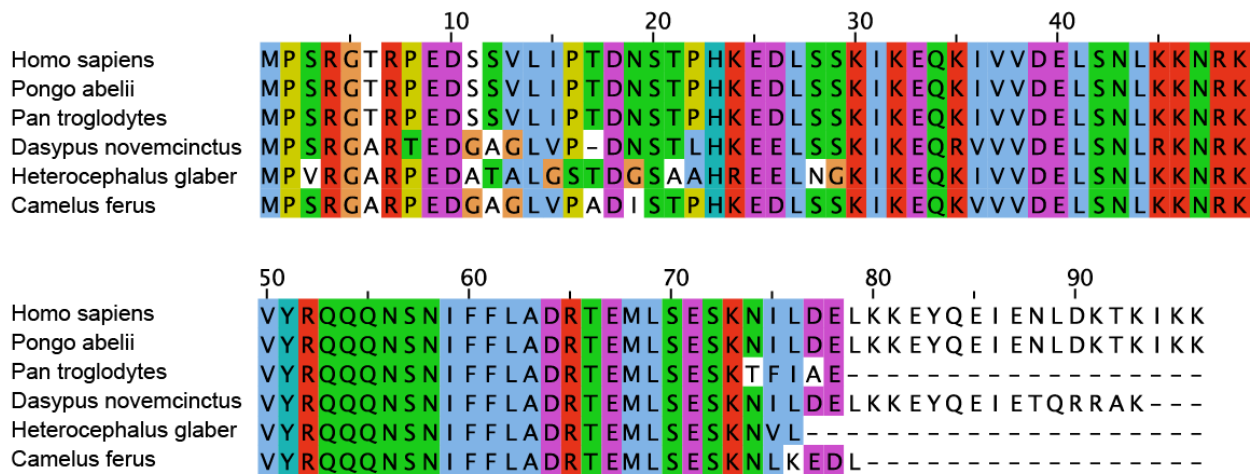


Figure 1.6. SEP conservation. ASNSD1-SEP is a 96 amino acid-long SEP detected in K562, MCF10A and MDAMB231 cell lines. It is highly conserved across mammals, indicating its potential biological function.

Data on the half-life and abundance of SEPs might reveal those SEPs that are long-lived and abundant, which might have the best chance to partake in cellular or physiological functions. Therefore, methods to measure the half-lives of SEPs will be of great value in studying their function. In addition, though SEPs are essentially small proteins, the understanding of their regulation is nowhere near as far along as other proteins. The *Sgrs* gene for example regulates its SEP in the presence of excess glucose by transcriptional regulation, but it is possible, even likely, that other SEPs are regulated at post-translational level. If so, how can these post-translational modifications be discovered and what are their roles in SEP biology?

Several studies revealed that SEPs play pivotal roles in a variety of cellular processes, and aberrant regulation of these bioactive polypeptides leads to developmental defects as well as pathogenesis. As more and more SEPs have been identified, functional and mechanistic exploration becomes increasingly necessary. Knowledge of their modes of action and roles in biology will make tremendous contributions to a new layer of regulation in our genome and proteome and could potentially offer novel therapeutic interventions to human diseases.

1.6. References

1. V. Ambros, The functions of animal microRNAs. *Nature* **431**, 350-355 (2004).
2. M. G. Vander Heiden, Targeting cancer metabolism: a therapeutic window opens. *Nat Rev Drug Discov* **10**, 671-684 (2011).
3. A. Fatica, I. Bozzoni, Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* **15**, 7-21 (2014).
4. M. M. Yore *et al.*, Discovery of a Class of Endogenous Mammalian Lipids with Anti-Diabetic and Anti-inflammatory Effects. *Cell* **159**, 318-332 (2014).
5. Y. Hashimoto, T. Kondo, Y. Kageyama, Lilliputians get into the limelight: Novel class of small peptide genes in morphogenesis. *Dev Growth Differ* **50**, S269-S276 (2008).

6. S. A. Slavoff *et al.*, Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**, 59-64 (2013).
7. J. Ma *et al.*, Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* **13**, 1757-1765 (2014).
8. J. J. Holst, The physiology of glucagon-like peptide 1. *Physiol Rev* **87**, 1409-1439 (2007).
9. D. F. Steiner, The proprotein convertases. *Curr Opin Chem Biol* **2**, 31-39 (1998).
10. M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**, e106 (2007).
11. T. Kondo *et al.*, Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science* **329**, 336-339 (2010).
12. E. G. Magny *et al.*, Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341**, 1116-1120 (2013).
13. A. Pauli *et al.*, Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343**, 1248636 (2014).
14. X. Dong *et al.*, Zm908p11, encoded by a short open reading frame (sORF) gene, functions in pollen tube growth as a profilin ligand in maize. *J Exp Bot* **64**, 2359-2372 (2013).
15. Y. Hashimoto *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 USA* **98**, 6336-6341 (2001).
16. M. E. Dinger, K. C. Pang, T. R. Mercer, J. S. Mattick, Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Comput Biol* **4**, e1000176 (2008).
17. S. J. Andrews, J. A. Rothnagel, Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* **15**, 193-204 (2014).
18. M. Q. Zhang, Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet* **3**, 698-709 (2002).
19. M. R. Brent, R. Guigo, Recent advances in gene structure prediction. *Curr Opin Struct Biol* **14**, 264-272 (2004).
20. M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-282 (2011).

21. I. P. Ivanov, A. E. Firth, A. M. Michel, J. F. Atkins, P. V. Baranov, Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res* **39**, 4220-4234 (2011).
22. G. Menschaert *et al.*, Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics* **12**, 1780-1790 (2013).
23. C. S. Wadler, C. K. Vanderpool, 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).
24. H. Rohrig, J. Schmidt, E. Miklashevichs, J. Schell, M. John, Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc Natl Acad Sci USA* **99**, 1915-1920 (2002).
25. X. H. Yang *et al.*, Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res* **21**, 634-641 (2011).
26. B. De Coninck *et al.*, Mining the genome of *Arabidopsis thaliana* as a basis for the identification of novel bioactive peptides involved in oxidative stress tolerance. *J Exp Bot* **64**, 5297-5307 (2013).
27. J. P. Kastenmayer *et al.*, Functional genomics of genes with small open reading frames (sORFs) in *S-cerevisiae*. *Genome Res* **16**, 365-373 (2006).
28. C. A. Gleason, Q. L. Liu, V. M. Williamson, 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).
29. K. Maki, T. Morita, H. Otaka, H. Aiba, A minimal base-pairing region of a bacterial small RNA SgrS required for translational repression of ptsG mRNA. *Mol Microbiol* **76**, 782-792 (2010).
30. J. B. Rice, C. K. Vanderpool, The small RNA SgrS controls sugar-phosphate accumulation by regulating multiple PTS genes. *Nucleic Acids Res* **39**, 3806-3819 (2011).
31. T. Kondo *et al.*, Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* **9**, 660-665 (2007).
32. B. Guo *et al.*, Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **423**, 456-461 (2003).

33. D. Zhai *et al.*, Humanin binds and nullifies Bid activity by blocking its activation of Bax and Bak. *J Biol Chem* **280**, 15815-15824 (2005).
34. S. E. Calvo, D. J. Pagliarini, V. K. Mootha, Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* **106**, 7507-7512 (2009).
35. E. Ladoukakis, V. Pereira, E. G. Magny, A. Eyre-Walker, J. P. Couso, Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* **12**, R118 (2011).
36. M. Falth *et al.*, SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol Cell Proteomics* **5**, 998-1005 (2006).
37. J. Crappe *et al.*, Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* **14**, 648 (2013).
38. B. Vanderperre *et al.*, Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS One* **8**, e70698 (2013).
39. A. A. Bazzini *et al.*, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *Embo J* **33**, 981-993 (2014).
40. J. E. Smith *et al.*, Translation of Small Open Reading Frames within Unannotated RNA Transcripts in *Saccharomyces cerevisiae*. *Cell Rep* **7**, 1858-1866 (2014).
41. N. T. Ingolia, Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* **15**, 205-213 (2014).
42. N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218-223 (2009).
43. N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, J. S. Weissman, The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**, 1534-1550 (2012).
44. G. L. Chew *et al.*, Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140**, 2828-2834 (2013).
45. N. T. Ingolia *et al.*, Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep* **8**, 1365-1379 (2014).
46. B. A. Wilson, J. Masel, Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol* **3**, 1245-1252 (2011).

47. M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, E. S. Lander, Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240-251 (2013).
48. J. L. Aspden *et al.*, Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife* **3**, e03528 (2014).
49. J. T. Ferguson, C. D. Wenger, W. W. Metcalf, N. L. Kelleher, Top-down proteomics reveals novel protein forms expressed in *Methanosarcina acetivorans*. *J Am Soc Mass Spectrom* **20**, 1743-1750 (2009).
50. M. Oyama *et al.*, Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol Cell Proteomics* **6**, 1000-1006 (2007).
51. M. Kozak, Point Mutations Define a Sequence Flanking the Aug Initiator Codon That Modulates Translation by Eukaryotic Ribosomes. *Cell* **44**, 283-292 (1986).
52. P. Van Damme, D. Gawron, W. Van Criekinge, G. Menschaert, N-terminal Proteomics and Ribosome Profiling Provide a Comprehensive View of the Alternative Translation Initiation Landscape in Mice and Men. *Mol Cell Proteomics* **13**, 1245-1261 (2014).
53. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
54. W. S. Lo *et al.*, Human tRNA synthetase catalytic nulls with diverse functions. *Science* **345**, 328-332 (2014).
55. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J Biol Chem* **289**, 10950-10957 (2014).
56. A. R. Saltiel, C. R. Kahn, Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799-806 (2001).
57. A. N. van den Pol, Neuropeptide transmission in brain circuits. *Neuron* **76**, 98-115 (2012).
58. R. N. Margolis, R. M. Evans, B. W. O'Malley, N. A. Consortium, The Nuclear Receptor Signaling Atlas: development of a functional atlas of nuclear receptors. *Mol Endocrinol* **19**, 2433-2436 (2005).
59. J. B. Regard, I. T. Sato, S. R. Coughlin, Anatomical profiling of G protein-coupled receptor expression. *Cell* **135**, 561-571 (2008).

Chapter 2

Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue

This chapter was adapted from: Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, Budnik BA, Kellis M, Saghatelian A. Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *J Proteome Res.* 2014 Mar 7;13(3):1757-65

2.1. Introduction

Modern transcriptome profiling methods such as tiling arrays (1) and whole transcriptome shotgun sequencing (RNA-Seq) (2) have revealed that a larger number of RNAs are produced from the genome than previously thought (3-6). Furthermore, subsequent analysis of these non-annotated transcripts has demonstrated the existence of functional non-coding RNAs, such as long intergenic non-coding RNAs (LINC)s (7, 8). The identification of additional RNAs also raises the possibility that there may also exist additional non-annotated protein-coding RNAs. The computational prediction of open reading frames (ORFs) (i.e. protein-coding regions) relies on a number of stringent criteria to avoid false discovery, such as a length cutoff, AUG start codon usage, and sequence conservation (9, 10). These criteria are not perfect, and several types of ORFs are often missed, including ORFs that use non-AUG initiation codons as well as short ORFs (sORFs) that fall below the typical length cutoff of a 100 codons (i.e. a 100 amino acid polypeptide) (11, 12). Firth and colleagues, for example, utilized a new search algorithm to reanalyze the mouse genome and predicted an additional 3000 protein-coding sORFs(13), which would correspond to an 10% increase in the size of the mouse genome (14).

More recently, direct experimental evidence for the existence of non-AUG initiation sites and protein-coding sORFs has begun to emerge. Ribosome profiling methods, which footprint the location of the ribosome on RNAs to identify protein-coding regions, revealed the existence of a number of non-annotated protein-coding sORFs in the mouse genome (11). In these experiments, the addition of the drug cycloheximide freezes the ribosome on start codons, and when cycloheximide is used in combination with ribosome profiling the start codons of ORFs can also be identified (15). This analysis led to the observation that while AUG is the most

common codon used (~45% of the time), CUG and GUG are also used frequently(11), which contradicts the dogma that translation initiation is restricted to AUG. Thus, ribosome profiling indicates that cells often use non-AUG start codons and reveal the existence of non-annotated protein-coding sORFs, both of which would likely be missed by classical algorithms for predicting protein-coding regions in the genome.

In addition to ribosome profiling, mass spectrometry (MS) peptidomics and proteomics experiments have recently been implemented in the discovery of sORF-encoded peptides (SEPs) (12, 16). These MS experiments differ from ribosome profiling because they detect polypeptide generated from a sORF and therefore validate the protein-coding potential of the sORF by demonstrating the production of a stable protein product. Because of transcript amplification and number of reads per sequencing experiment ribosome profiling is more sensitive and will identify the largest number of sORFs, but the bias of MS towards more abundant proteins (17) means that peptidomics and proteomics will likely identify the most abundant cellular SEPs, which might be the SEPs most likely to be functional. Slavoff and colleagues developed and utilized a peptidomics-based strategy for the detection of novel human SEPs (12). These studies were based on initial observations by Yamamoto and colleagues who identified four SEPs in K562 cells (defined here as less than 150 amino acids in length) (18). To improve on these results, Slavoff and co-workers utilized next-generation RNA sequencing (RNA-Seq) to identify all possible protein-coding mRNA transcripts, including non-annotated transcripts (i.e. transcripts that exist but are not in the NCBI RefSeq database). The RNA-Seq data was translated into all possible reading frames to create a database that should contain all of polypeptide sequences that could theoretically be produced in the cell. Using this database, Slavoff and colleagues identified 90 human SEPs in these K562 cells and 86 of these SEPs were

novel (12). This work indicated that SEPs represent a large class of non-annotated cellular polypeptides. Recent work from others has also supported this conclusion, with Vanderperre and colleagues having characterized 1259 non-annotated polypeptides (19), the largest number reported to date using an elegant combination of bioinformatics and mass spectrometry.

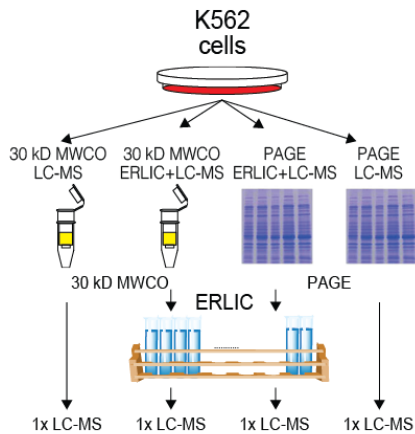
Our goal here is to 1) determine whether we can identify a workflow that provides the easiest route for SEP detection, 2) determine whether SEPs exist in other cell lines, and 3) determine whether we can find SEPs in human tissues, specifically a human tumor sample. Our results identify several workflows for SEP discovery, and demonstrate that SEPs are ubiquitous and present in multiple cell lines and human tissues.

2.2. Results and Discussion

2.2.1. Impact of Different Workflows on SEP Discovery

Our first goal was to determine whether changes to the reported SEP-discovery workflow would lead to the identification of additional SEPs in K562 cells, and whether any particular workflow is superior to others. In the reported workflow (12), SEPs are separated from larger proteins with a 30 kDa molecular weight cutoff (MWCO) filter, and the ≤ 30 kDa fraction then undergoes electrostatic repulsion hydrophilic interaction chromatography (ERLIC) (20, 21) followed by LC-MS/MS (Figure 2.1). This workflow led to the identification of 90 SEPs, 86 of which were novel, in the commonly used K562 leukemia cell line (12). We refer to this workflow as MWCO + ERLIC + LC-MS/MS. More recently, Vanderperre and colleagues have identified 1259 non-annotated polypeptides (19). Below, total SEPs refer to the total number of SEPs discovered and novel SEPs refer to any SEPs from these groups that was not identified in Slavoff, et. al. or Vanderperre, et. al. manuscripts.

A



B

Workflow	Total SEPs Detected	Novel SEPs Detected
MWCO + LC-MS	13	6
MWCO + ERLIC + LC-MS	90	86
PAGE + ERLIC + LC-MS	94	80
PAGE + LC-MS	19	7

Figure 2.1. The workflows tested in the discovery of novel human SEPs. (A) A schematic of the four different SEP discovery workflows used: MWCO+LC-MS; MWCO+ERLIC+LC-MS; PAGE+ERLIC+LC-MS; and PAGE+LC-MS. The K562 peptidome is separated by size using a 30kD MWCO filter (MWCO) or polyacrylamide gel electrophoresis (PAGE), and then analyzed directly by LC-MS (first and last lane) or fractionated by ERLIC prior to LC-MS analysis (middle lanes). (B) The number of total SEP and novel SEPs identified in K562 cells using each of the four different SEP discovery workflows.

Three additional workflows were tested here: MWCO + LC-MS/MS; PAGE + ERLIC + LC-MS/MS; and PAGE + LC-MS/MS. In these workflows, MWCO indicates fractionation using a 30 kDa MWCO filter, while PAGE refers to molecular weight fractionation using a 16% Tricine polyacrylamide gel, where the region between 2-15 kDa is analyzed by LC-MS/MS. We used K562 cells in these experiments. All of these workflows led to the discovery of novel human SEPs, though the number of SEPs and the ease of the different methods varied.

We began by comparing the MWCO + LC-MS/MS and the PAGE + LC-MS/MS workflows. These workflows differ in their approach to peptidome isolation by using a 30 kDa MWCO filter or the excising the 2-15 kDa portion of a 16% Tricine gel. After separation, the \leq 30 kDa fraction is treated with trypsin and then analyzed by LC-MS/MS. The 2-15 kDa gel slice

undergoes an in-gel trypsin digest followed by LC-MS/MS analysis. SEPs are identified using a custom K562 database generated from RNA-Seq data that will account for polypeptides produced from previously non-annotated protein-coding sORFs. We identified 13 SEPs using the MWCO + LC-MS/MS workflow with a single LC-MS/MS run. Of these 13 SEPs, six were novel, while seven were identified before (Figure 2.1B). In comparison, the PAGE + LC-MS/MS workflow identified 19 SEPs, with seven of these being novel. These results indicate that both MWCO and PAGE fractionation are able to identify similar number of total SEPs (13 vs 19) per LC-MS/MS run (Figure 2.1B). None of the novel SEPs discovered by these two methods overlapped, resulting in the discovery of an additional 13 human SEPs (six from MWCO and seven from PAGE).

Next we analyzed the K562 sample by PAGE + ERLIC + LC-MS/MS (Figure 2.1A). In this approach, we subject the sample to ERLIC after an in-gel trypsin digestion. The ERLIC fractionated samples are then analyzed by LC-MS/MS and new SEPs identified by analysis of the K562 database. This analysis led to the identification of 94 SEPs, and 80 novel SEPs (Figure 2.1B). Thus, the two workflows that utilize ERLIC identify 90-94 SEPs per run while workflows without ERLIC identified 13-19 SEPs per run. As expected, increased fractionation results in better coverage and there is no substantial difference between different methods of peptidome size fractionation (i.e. PAGE or MWCO).

2.2.2. Biological and Technical Replicates Increase the Number of SEPs Discovered

The preliminary data revealed that there is little overlap between the different workflows. We hypothesize that the low natural abundance of SEPs and shotgun peptidomics, which is inherently stochastic (17), results in the low overlap among samples. Indeed, the Yates lab has

demonstrated that in complex mixtures that data-dependent data acquisition doesn't completely sample all peptides in a sample, and therefore does not provide information on all ions (17). Based on models of this process, they determined that for yeast cell soluble lysate that 10 replicates are required to achieve 95% saturation of the proteome (17). In addition, most SEPs are short (<100 amino acids) such that they do not generate many tryptic peptides that can be used to identify a SEP. In most cases, we only detect a single peptide for each SEP identified and if this peptide is missed due to inefficient ionization or low abundance then the entire SEP and sORF is overlooked. Together, these factors contribute to the variable detection of SEPs. If SEP detection were stochastic, then biological and technical replicates would be expected to show little overlap in the SEPs identified per LC-MS/MS run, but each replicate analysis would yield additional SEPs.

We repeated the PAGE + LC-MS/MS for three additional K562 samples for a total of four biological replicates (which includes the sample from Figure 2.1). An average of 22 SEPs were detected per run with a range between 11-37 SEPs in each sample (Figure 2.2A). Of the 87 total SEPs identified, 26 overlapped with previously detected SEPs and 61 were novel, for an average of 15 new human SEPs per run. Many of the novel SEPs were only identified in a single sample. These results bring the total number of novel SEPs detected here to 147 (80 from PAGE + ERLIC + LC-MS (Figure 2.1), 6 from MWCO + LC-MS (Figure 2.1), and 61 from four PAGE + LC-MS biological replicates (Figure 2.2)). The lack of overlap between samples is consistent with our earlier observations and supports the idea that SEP detection is variable, as predicted from proteomics studies (17).

A

Biological Replicate	Total SEPs Detected	Novel SEPs Detected
1	19	7
2	11	8
3	20	18
4	37	28
Total	87	61

B

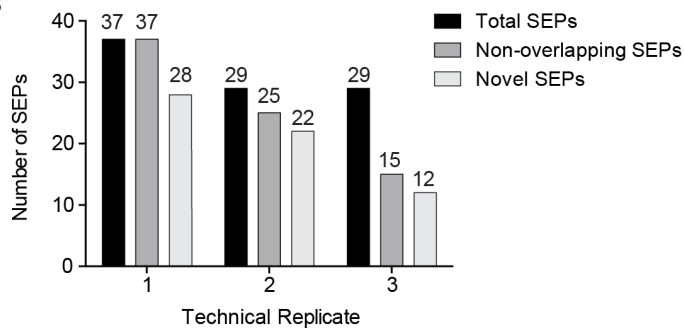


Figure 2.2. Biological and technical replicates lead to the discovery of novel SEPs. (A) The number of SEPs detected in four biological replicates of K562 cells. Each of these samples was analyzed using the PAGE+LC-MS SEP discovery workflow. For each replicate, the detected SEPs include the total number of SEPs identified as well as the novel SEPs that were characterized for the first time. (B) Three technical replicates were of biological replicate #4 from part A was performed using the PAGE+LC-MS workflow with K562 peptidome. The total number of SEPs detected in each run (black), non-overlapping SEPs (gray; SEPs that were not present in either of the other two technical replicates), and novel SEPs (light gray; SEPs that were not detected in any other analysis).

Next, we tested the impact of performing technical replicates. We analyzed biological replicate #4 (Figure 2.2A)— where we identified 37 total SEPs in the first run— two more times to provide a total of three technical replicates. In the three runs, we identified 37, 29 and 29 SEPs in each run (Figure 2.2B). Of the 29 SEPs identified in the second run, 25 were not detected in the first run (i.e. non-overlapping SEPs), and of the 29 detected in the third run, 15 were not detected in the first or second runs (Figure 2.2B). The number of novel SEPs identified per run decreases from 28-12 as more runs are performed but there are still a substantial amount of novel SEPs discovered even after three technical replicates. This result affirms the hypothesis that SEP detection is stochastic and demonstrates the value in performing biological and/or technical replicates to increase the number of SEPs discovered. Additionally, we also performed five more technical replicates (using biological replicate #3 from Figure 2.2A) and detected 48 SEPs (with 32 of these being novel SEPs) (Figure 2.3). At this point, we had identified a total of 195 novel

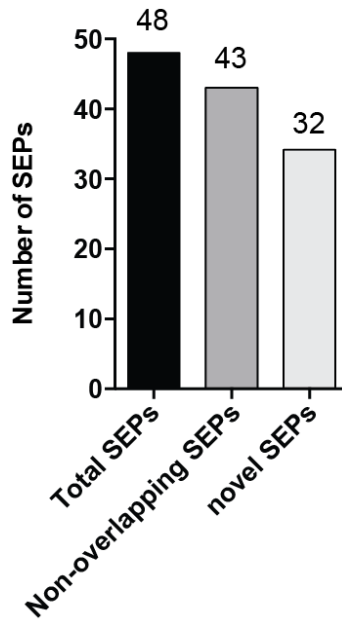


Figure 2.3 Total number of SEPs detected in K562 cells using PAGE+LC-MS/MS workflow after performing an additional six technical replicates.

SEPs (i.e. not identified in Slavoff et. al. or Vanderperre et. al.) in K562 cells through a combination of different workflows, biological and technical replicates.

Three to four biological/technical replicates matches the total number and novel SEPs identified through an ERLIC fractionation; however, we analyzed a total of 25 ERLIC fractions by LC-MS/MS. Thus, it seems more efficient to perform multiple technical and or biological replicates when wanting to identify more SEPs, as predicted from similar conclusions made with data-dependent proteomics experiments (17). Finally, a handful of SEPs were detected among biological and/or technical replicates repeatedly such as ASNSD1-SEP and CIR1-SEP.

ASNSD1-SEP is the most frequently SEP and therefore it is likely to have high cellular concentration and stability. ASNSD1-SEP also shows an unmistakable evolutionary signature of protein coding regions (Figure 2.4), as measured across 29 eutherian mammals by PhyloCSF (28), suggesting that this SEP has undergone positive selection during evolution. In total, 195

novel SEPs represents a greater than 200% increase from the previous study and also the largest number of SEPs ever reported from a single cell line.

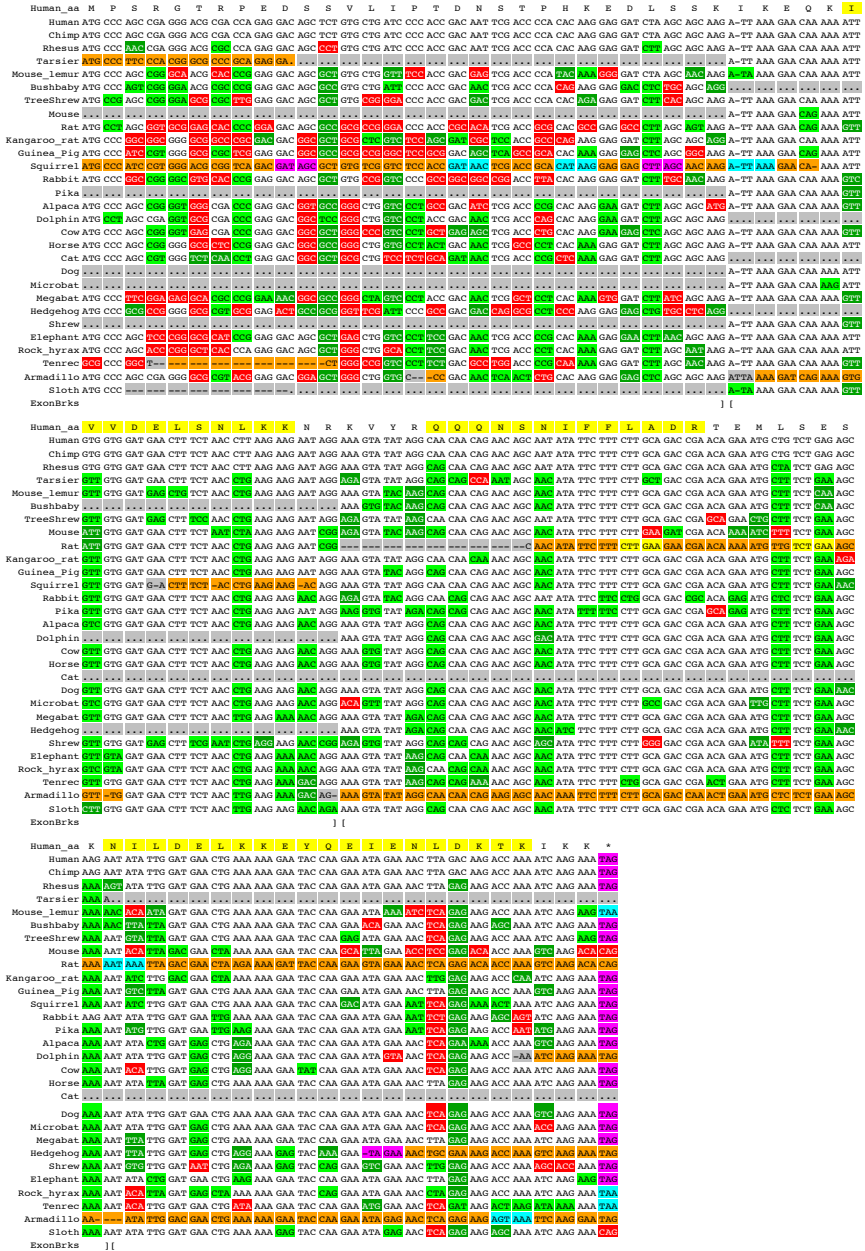


Figure 2.4. Alignment for ASNSD1-SEP shows protein-coding signature. Alignment for ASNSD1-SEP across 29 eutherian mammals, color-coded by CodAlignView

Figure 2.4. (Continued) ("CodAlignView: a tool for visualizing protein-coding constraint", I Jungreis, M Lin, M Kellis, in preparation). The amino acid sequences of the four tryptic peptides detected, NILDELKK, IVVDELSNLKK, QQQNSNIFFLADR, and EYQEIENLDKTK, are highlighted in yellow. The high concentrations of synonymous substitutions (light green) and conservative amino acid changes (dark green), and relatively low concentrations of radical amino acid changes (red) and frame-shifted regions (orange) is characteristic of protein-coding regions. The region's evolutionary coding potential as measured by per-codon PhyloCSF score, 4.315, is higher than 99.97% of non-coding regions, implying that it has been functional at the amino acid level in much of the eutherian mammal tree.

2.2.3. Using Targeted LC-MS/MS to Rapidly Validate Novel SEPs

In the majority of cases, a single peptide is used to identify a SEP. Analysis of our data showed that only 7 out of the 195 novel SEPs had more than one unique peptide to support the protein-coding potential of the sORF. To obtain additional data to support the identification of a novel protein-coding sORF, we previously relied on molecular biology (12). We cloned the candidate protein-coding sORFs and tested whether they produce SEPs in mammalian cells to ensure that the newly identified sORF actually code for proteins. While successful, this approach is time consuming and does not provide the necessary throughput to validate large numbers of SEPs easily. We decided to use mass spectrometry instead of molecular biology to increase the throughput and provide more evidence for the endogenous detection of SEPs. Specifically, our aim was to use targeted MRM LC-MS/MS to characterize additional peptides from sORFs. This approach would afford more than one peptide from the sORF and in doing so would provide the necessary data to validate the sORF, and should increase throughput.

Skyline, a program designed to identify the best tryptic peptides from an ORF for targeted multiple-reaction monitoring (MRM) experiments (27), was used to define the MRM transitions for peptides derived from 105 SEPs. These SEPs include the 81 from the PAGE +

ERLIC + LC-MS and seven from MCWO + LC-MS (Figure 2.1), as well as 17 SEPs from the biological replicates #1 and #2, which were identified by PAGE + LC-MS (Figure 2.2), for a total of 105 SEPs. Trypsin digestion of these 105 SEPs resulted in 224 tryptic peptides and over 700 transitions were predicted by Skyline and monitored by targeted MRM LC-MS/MS. The total number of SEPs was capped at 105 in this targeted MRM LC-MS experiment due to the total number of MRM transitions that could be easily monitored per run. This experiment confirmed the presence of 62 peptides out of the possible 224 (27%) and the identification of these peptides resulted in having at least two peptides identified for 36 out of the 105 SEPs (34%) (Table 2.1). Skyline analysis of PRR3-SEP (Figure 2.5), for example, identified MRM transitions for four tryptic peptides and a targeted LC-MS/MS using these transitions identified the existence of two out of four of these peptides (Figure 2.5, Figure 2.6 for MS/MS of PRR3-SEP peptide that we detected). Along, with the PRR3-SEP peptide identified during shotgun peptidomics, we now have a total of three peptides identified from the PRR3-SEP, which provides the necessary confirmation that the PRR3 sORF is protein-coding.

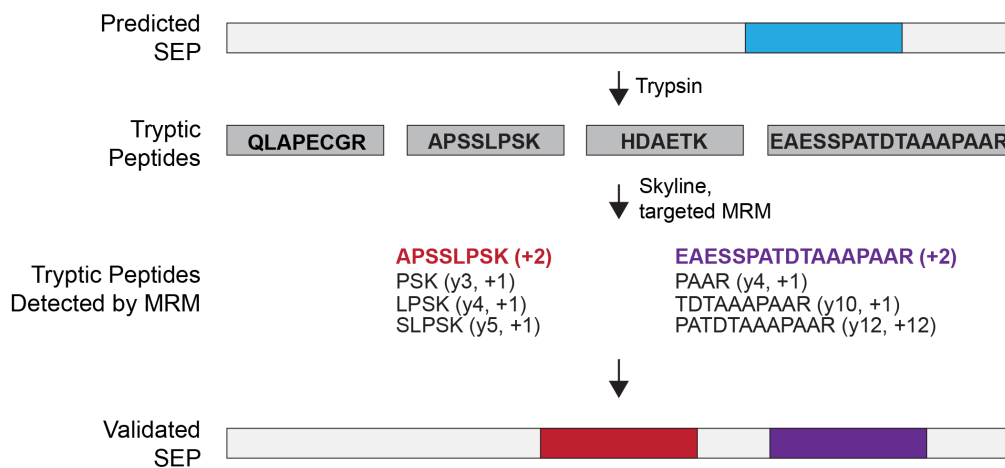


Figure 2.5. Validating SEPs with targeted mass spectrometry. Analysis of PRR3-SEP by Skyline and subsequent MRM targeted LC-MS identifies additional peptides from

Figure 2.5. (Continued) this SEP. The tryptic peptide (blue box) that was detected in the original shotgun proteomics experiment led to the initial identification of the PRR3-SEP. To identify additional peptides from PRR3-SEP, Skyline is used to predict MRM transitions for four tryptic peptides from PRR3-SEP and this information is fed into a targeted LC-MS experiment. This experiment identified peptides for two out of the four peptides and provided an additional two peptides (red and purple boxes) to validate this PRR3-SEP.

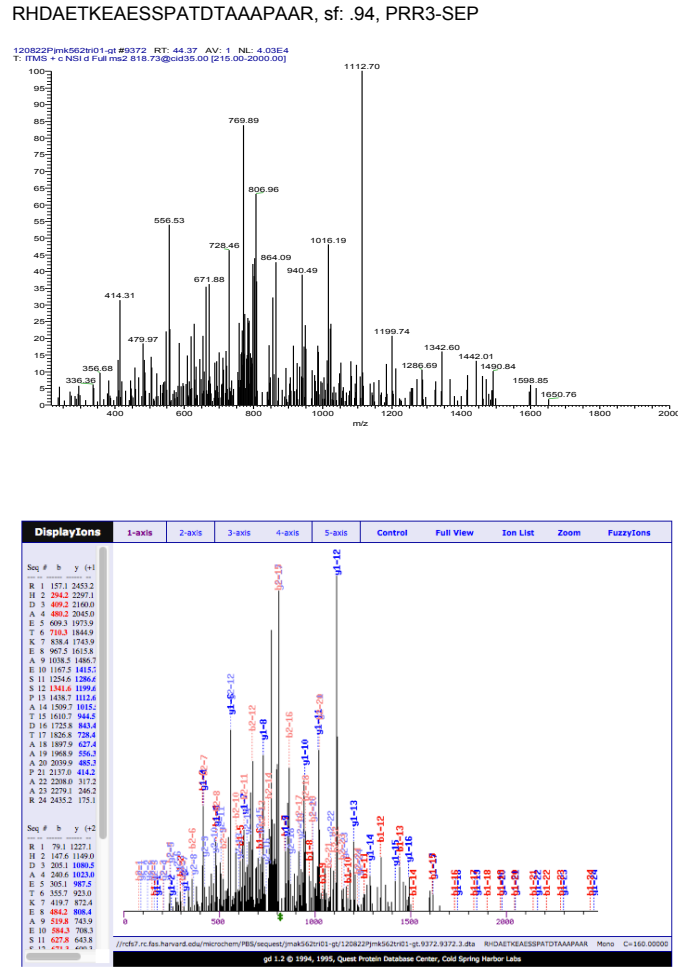


Figure 2.6. MS/MS spectra (raw spectrum: top, SEQUEST annotated spectrum: bottom) for the detected peptide: RHDAETKEAESSPATDTAAAPAAR for PRR3-SEP, with SF score of 0.94.

Detected peptide by shotgun	Detected peptide by Skyline-MRM	Predicted SEP length	SEP sequence
IVYGDIRK KIVYGDIR	IVYGDIR	41	LYQRREELKLQWRSFNLDKIVYGDIRK EGI LNALETMGSSF
ADAPAVSPESPQK	KPPFQPR ADAPAVSPESPQK	101	MSRGGFQGQEQEPLDMFFWVNEISGEITYPPQKADAPAVSPE SPQKKPPFQPRSVQEAPCSPQGGPPAQRPALAPPSKPSLKDS GSRNPCPSAPTWARPKPEE
KNFFISK	ALAYIGAR	20	MPKNFFISKIKALAYIGARC
EAESSPATDTAAAPAAR RHDAETKEAESSPATDTAA APAAR	APSSLPSK EAESSPATDTAAAPA AR	93	MPGRDGGRLAPECGRRRCSPQPSGSGMAGASALLPTDPP EAETESPRAPSSLPSKSRCSHCRRHDAETKEAESSPATDTAA APAARAGRDRWR
FSPLELAAGGVR	FSPLELAAGGVR LLLPCPCCFG	127	MAGGDPAAAGRAVGAAKAKRAAGICGDTAAPALSGHPPDCLPE SPGRGGLTCSQSRFSPLELAAGGVRGCGAQLVTGVCSEP SWSPGEGAGPGEQQSLDVEAAEALQAGPASPRCRLLLPCPC CFG
YLDLHER	QIFHSNK	33	IRKYLDLHERQIFHSNKHNNIINCKAQTTTRVK
QGSLVQQVALR	EVLEQLMK	66	MMKNQWMEIGHHLHQPSRFLHLKHEILPTVQGSLSVQQVAL RLEVLLRKEVLEQLMKMIL
RNLVVVINL	DPMNAK TLEVVIS	71	MNVRNVRNRLVVVINLFIYITGFMSLRDPMNAKSVGRTFVVAIN LLYIKDFILVRNPMNVQNVGRTLEVVIS
IYNRLYFLEK	VNFCNK	24	VNFCNKASWDLKIIYNRLYFLEKF
EGGWRQVEGTGTPK	DSAKPIR QVEGTGTPK	75	MTTGERDSAKPIRATATRQEDRSPEGGWRQVEGTGTPKSKQ GSRVLAQEETQHPEAVPQRADPKGASASPLRRQ
ELSQYLK	MDELSQYLK	42	MDELSQYLKVIPLSTVLDVILLQLSFLLYIANFLSLGSLP
SRQVDQEVRRSSR	QVDQEVRR	24	SQHFGRRQVDQEVRRSSRTAWPRW
ENIPDITK	DFVFNLSK	32	DFVFNLSKILVENRPAFVNENIPDITKPKHF
VAEIIIER	TYTCLLR LVSHGINLALIFSIWK	116	RVAEIIIERLVSHGINLALIFSIWKLKENHFHCRKSFFKYLlyPR EISLYLPPQAVISCFREWNPPCPSIFWFLGNSLWVSPWLGIL SWEQILSCLMCLHSPKTYTCLLRA
AIVVARVVTIPK	MVAIVVAR	43	MVAIVVARVVTIPKIMHQPVLSFLNFHVPLYTFMVSYYDLSLV
GDFLNLNLR AGDFLNLNLR GFLAGYVVAK TLRDYLQLLR NQLESQQR RVEDEVNSGVGQDGSLLSS PFLK	AGDFLNLNLR IGISYQFCK NQLESQQR NQLESQQR GFLAGYVVAK	41 103	AGDFLNLNLRIGISYQFCKFSPINYYFFLFSPCLLYGILLDIS MADDKDSLPLKDLAFLKNQLESQQRVEDEVNSGVGQDGS LLSSPFLKFLAGYVVAKLRAVAVLGFVAVGTCTGIYAAQAYAVP NVEKTLRDYLQLLRKGPD
RLLFAGK IRLLFAGK	GTTFSWVIR	22	GTTFSWVIRLLFAGKLNYSMS
GLIENPALIR	LMQEGK QGLIENPALIR	100	QRVQAERLAIRARLKREYLLQYNDPNRQGLIENPALIRWAYAR TTNVYPNFRPTPKNSLMGALCGFGPLIFIYIIKTERDRKEKLM QEGKLDRTVHLSY
QNIKGLENILQK	GLENILQK VVTLTQSSENQR	81	IWSRVVTLTQSSENQRQNIKGLENILQKEAATCVDNGLFMPL LLSVDLVQETCSGDGCEGMRIDITPVSTCLFITLL
SWLTPVAGK	MAPLGLK	26	MAPLGLKDPLSSWLTPVAGKLVMAVS
HALPLLK	QEFHALPLLK NSTNFFLLIK	52	NSTNFFLLIKQRSFGGFIADKRGKDGKCSRFLSFHKQEFHA LPLLKQRKE
GAGILLLR	TGAGILLLR	44	TGAGILLLRWLTHWLLAGSLRSPGVPLHVLLHGLMMWHEPH SV
KQNSLIANMEK	SGYINR SCLHSIK GEEAAEEK MQIEATR GQTLFSSTK QNSLIANMEK	114	LIKQNSLIANMEKVLVWVWMDQTSNIPLSQSLIQSKGQTLFS STKNEKGEEAAEEKFEASRVWLMRFKERSCLHSIKMQIEATR ADEEGTASDPEDPAKLIDKSGYINRFTM

Table 2.1. A list of 36 SEPs detected in K562 cells that were validated by Skyline-MRM.

KNEFLK	DHVLFFK	27	IIFKNEFLKDHVLFFKSIFSSYFCYC
AEIILK	VFDLQDF MAEIIILK	17	MAEIIILKAKVFDLQDF
TPLLAYIQ	TPLLAYIQDTSAF	49	MNLEMEKKAGLFQRVDLSELDSTIELCCIFCGSSKTPLLAYIQDTSAF
HAFNLNR	HALFLNLR NLQTPGAVGEDK	54	HAFNLNLRRAIPSPQSNLNERPQVQLLHSPDLLSTPRNLQTPGAVGEDK
EVEGAVSR	QSEVMSQK IFNNHTLIK	42	TQEVEGAVSRDCITALQPQKQSEVMSQKQTTKIFNNHTLIK
RKPLYTIGWNL	DFTSHQLER	64	SSGKGSNSQRDFTSHQLERLSSKRQNIKRVGKNAEKRPPLYTIGWNLNWYSHYKQHGSSKN
KINALLK	GNILLSNK	50	SQPPLKCLIKINALLKGNILLSNKCGCVFYHTSILRKCWTSEYHKTGN
FQPPHHVQSSPDVK	GLSFQPPHHVQSSPDVK	32	ESCEPTEQKGLSFQPPHHVQSSPDVKSQFWF
ARDQYGHLIPTK	KPSFSPR DQYGHLIPTK GSCHFLSQVGGWGI	61	MCAEIEEGAEGVTARDQYGHLIPTKVASGPQGLSGARKPSFSPRLRGSCHFLSQVGGWGI
LAFIFLPDR	NDLAFIFLPDR	65	AKIVPLHSSLGDRVRPCLKTKQTKEFRNDLAFIFLPDRQCIHQDGLTGNQVLAPLLAGKEHEVF

Table 2.1. (Continued) A list of 36 SEPs detected in K562 cells that were validated by Skyline-MRM. The peptides detected by shotgun proteomics were shown in blue, the additional peptides validated in Skyline-MRM were shown in red, the overlapping peptides detected in both shotgun method and Skyline-MRM were highlighted in purple.

Using molecular biology and peptide synthesis we had previously validated 17 out of 86 novel SEPs (20%) by expression or co-elution over several weeks (12). Here, we validated 36 out of a 105 SEPs (34%) by identifying a second peptide in approximately a week. Out of these 36 validated SEPs, 32 were novel. Thus, using Skyline (27) to define MRM transitions for SEP tryptic peptides and targeted MRM LC-MS/MS to validate SEPs provides a much more facile and efficient approach.

2.2.4. Overview of 195 Newly Identified SEPs

We analyzed the length distribution, codon usage, and source of RNA to determine whether the 195 newly identified SEPs in K562 cells differ substantially from the 86 SEPs we had previously identified (12). The length distribution for the SEPs was determined by using

AUG-to-stop or upstream stop-to-stop (i.e. distance between two in frame stop codons that encompass the sORF). We did not try to predict alternative start codons for the length distribution because we did not want to bias the analysis towards shorter lengths. The SEPs range between 8 and 134 amino acids long, with the majority (>90%) of new SEPs being less than < 100 amino acids long (Figure 2.7A).

We assigned initiation codons to sORFs using a simple set of criteria. An upstream in-frame AUG was assumed to be the start if present; otherwise the initiation codon is assigned to an in-frame near-cognate codon (NCC), which differs from AUG by a single base. NCCs were commonly found in ribosome profiling experiments (11) and our previous SEP discovery effort (12), so this result is consistent with what has already been observed. If neither of these criteria was met, no start codon was assigned. Many of these SEPs (~70%) do not appear to initiate with an AUG codon (Figure 2.7B).

Lastly, we tried to account for the RNAs that are responsible for producing these SEPs. First, we determined the RNAs in the RefSeq database that produce SEPs, and we refer to this pool of RNAs as “annotated RNAs”. These RefSeq RNAs are primarily mRNAs, which already contain a protein-coding ORF. Slightly over a quarter of all the SEPs we discovered, 47 in total, are derived from RefSeq RNAs (Figure 2.7C) A breakdown of the distribution of these SEPs on the RNAs reveal that a majority are found on the 3'-UTR. We counted sORFs in the 3'-UTR only if there is an additional stop codon between the start of the sORF and the stop codon of the upstream ORF and to avoid identifying read-through products (29, 30). In addition, we did not identify any splice acceptor sites at the 5'-end of the 3'-UTR sORFs (31) indicating that these are not alternative exons.

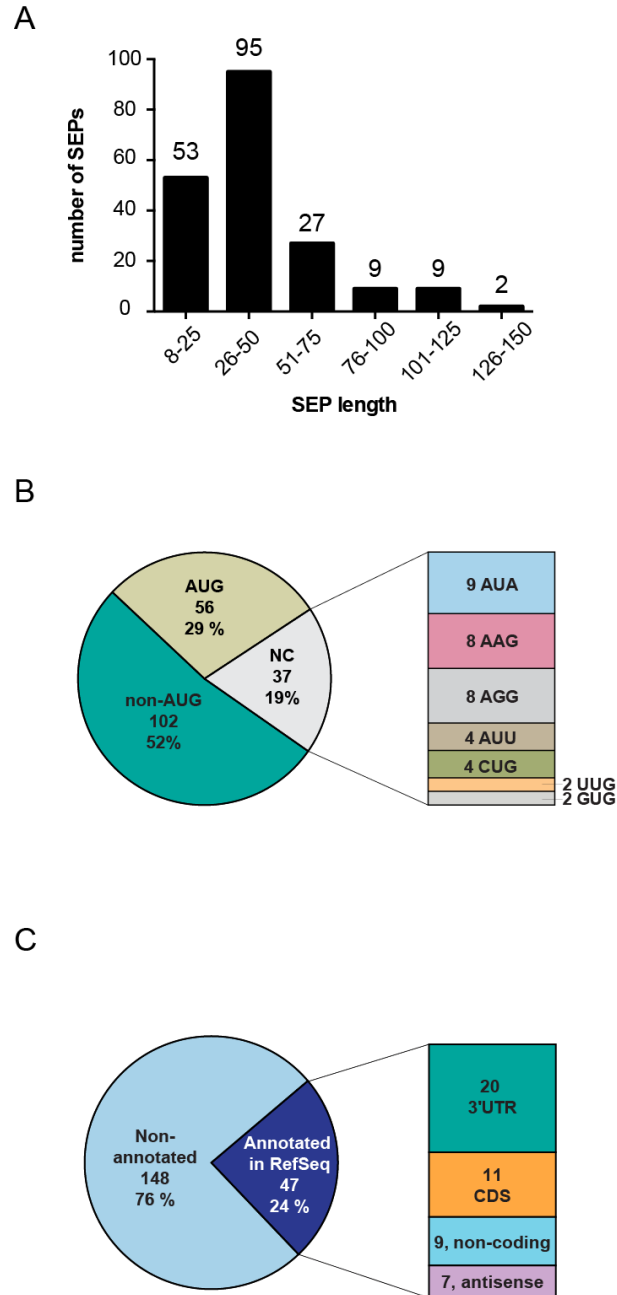


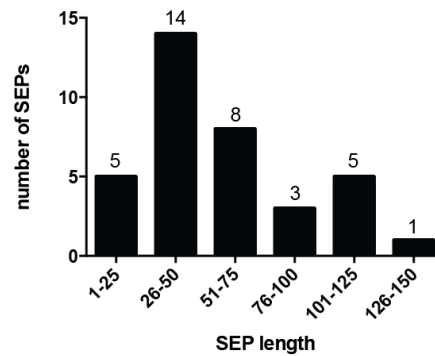
Figure 2.7. Overview of 195 novel SEPs identified in K562 cells. (A) The length of each SEP was determined using a defined set of criteria (see methods) and the length distribution reveals that the majority (> 90%) of SEPs discovered are between 8-100 amino acids. (B) SEPs utilize AUG, near cognate codons (i.e. one base away from AUG), and unknown codons to initiate translations. (C) SEPs are primarily derived from non-annotated RNAs (i.e. not found in RefSeq database) but RefSeq RNAs do account for the production of 24% of these SEPs. For the RefSeq-RNAs the sORFs are found on coding RNAs at the 3'-UTR, CDS and on non-coding RNAs such as antisense RNAs and non-coding RNAs.

SEPs are also produced from sORFs regions that are frame shifted within the coding sequence (CDS) of the longer ORF. These SEPs are likely produced from RNA splice forms that can only translate the sORF to produce the SEP because there is no plausible mechanism to explain the production of the ORF and sORF from the same mRNA (12). Since splice forms are difficult to distinguish by RNA-Seq further experimentation is necessary to validate that some SEPs are produced from a splice form of a known annotated RNA. The remaining sORFs are found in the 5'-UTR of RNAs (two SEPs in this study are generated from 5'-UTR of RefSeq annotated RNAs and these SEPs were detected previously in the study by Vanderperre et al.) , non-coding RNAs, and antisense RNAs (i.e. reverse-complement of known RNAs), which are produced at sites of transcription (32, 33). The discovery of a protein-coding sequence within a RNA that is annotated as non-coding reveals a weakness in common algorithms that assign protein-coding genes (9). The small number of sORFs in the 5'-UTR of RefSeq RNAs is the biggest difference between this set of SEPs and the previously reported set (12), where the majority of RefSeq sORFs we found were in the 5'-UTR. There could be several reasons for this, including the possibility that sORFs in the 5'-UTR produce the most abundant SEPs and therefore we and others already discovered the majority of them. Transcripts that are not part of the RefSeq database are considered to be “non-annotated”. We identified 148 SEPs that were generated from these non-annotated transcripts in the K562 RNA-seq database. Thus, there are still mRNAs and protein-coding genes that remain to be discovered.

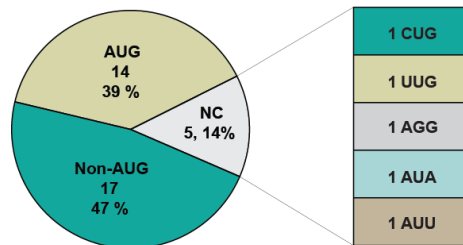
We also measured the lengths, initiation codon usage, and RNA source for the 36 MRM-validated SEPs from this set of 195 SEPs to determine whether MRM targeting is enriching for a particular class of SEPs. We find a similar distribution for SEP length, start codon usage and SEP mRNA RefSeq annotation for the 36 MRM-validated SEPs (Figure 2.8) as we do for the

195 SEPs (Figure 2.7), indicating that no bias is introduced during the targeted MRM experiment, and further supporting the use of Skyline-targeted MRM as a general, rapid, approach for the high-throughput validation of SEPs.

A



B



C

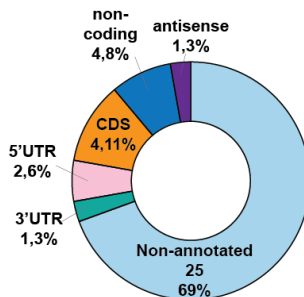


Figure 2.8. The characteristics of the 36 SEPs in K562 cells validated by Skyline-MRM. (A) The length distribution of the SEPs, (B) the start codon usage of the SEPs, (C) the SEPs mRNA annotation by RefSeq.

2.2.5. SEPs Are Found in Additional Cell Lines and Some Show a Cell-Specific Distribution

To ascertain whether SEPs are found in other cell lines and whether some SEPs are specific to certain cell lines, we profiled the MCF10A and MDAMB231 cell lines. These are breast cancer cell lines that differ in their invasiveness, with MDAMB231 being invasive (34). Invasiveness is a measure of the ability of a cell line to tunnel through a matrix in cell culture, and is thought to model the aggressiveness of the cancer cell line (35).

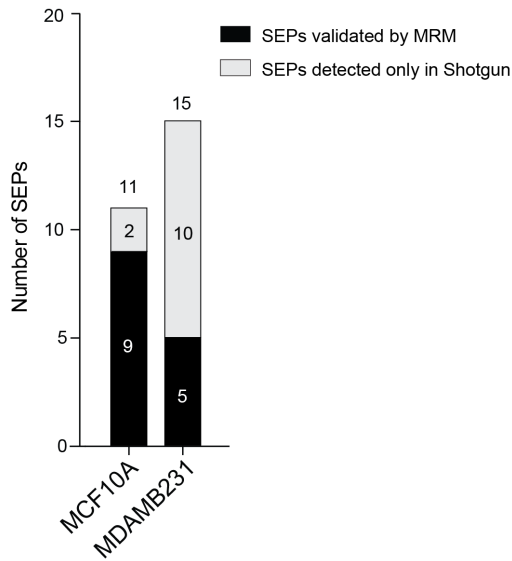
We obtained RNA-Seq data for these cell lines, assembled this data into a transcriptome, and then translated these sequences into a custom protein database. Analysis of MCF10A and MDAMB231 by PAGE + LC-MS/MS led to the identification of 12 and 17 SEPs, respectively (Figure 2.9A, Figure 2.10). Analysis of these SEPs by Skyline followed by a targeted MRM LC-MS/MS experiment validated 14 of these SEPs (out of 29)—nine in the MCF10A cell line and five in the MDAMB231 cell line (Figure 2.9B).

These 14 SEPs were targeted MRM LC-MS in both cell lines (MCF10A and MDAMB231) to determine whether any of these SEPs are specific to either cell line. Out of the 14 SEPs targeted, 12 are present in the MCF10A and MDAMB231 cell lines, while two SEPs are found only in the MDAMB231 sample (Figure 2.9C). Together, these experiments demonstrate that SEPs are found in additional (i.e. not K562) cell lines, and that some SEPs might be specific to particular cell lines.

A



B



C

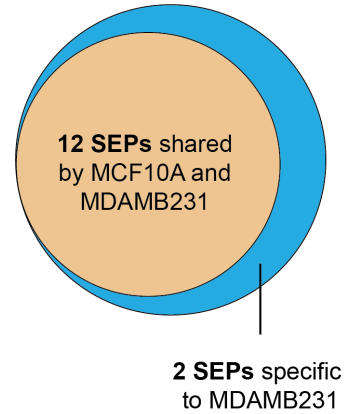


Figure 2.9. SEP derived from MCF10A and MDAMB231 cell lines. (A) The steps in the discovery and validation of SEPs from these cell lines. (B) A total of nine and five SEPs were validated using by MRM in the MCF10A and MDAMB231 cell lines, respectively. (C) These 14 validated SEPs were targeted in MCF10A and MDAMB231 and while 12 SEPs were found in both cell lines, two SEPs, TASP1-SEP and CAMD8-SEP, were specific to the MDAMB231 cell line.

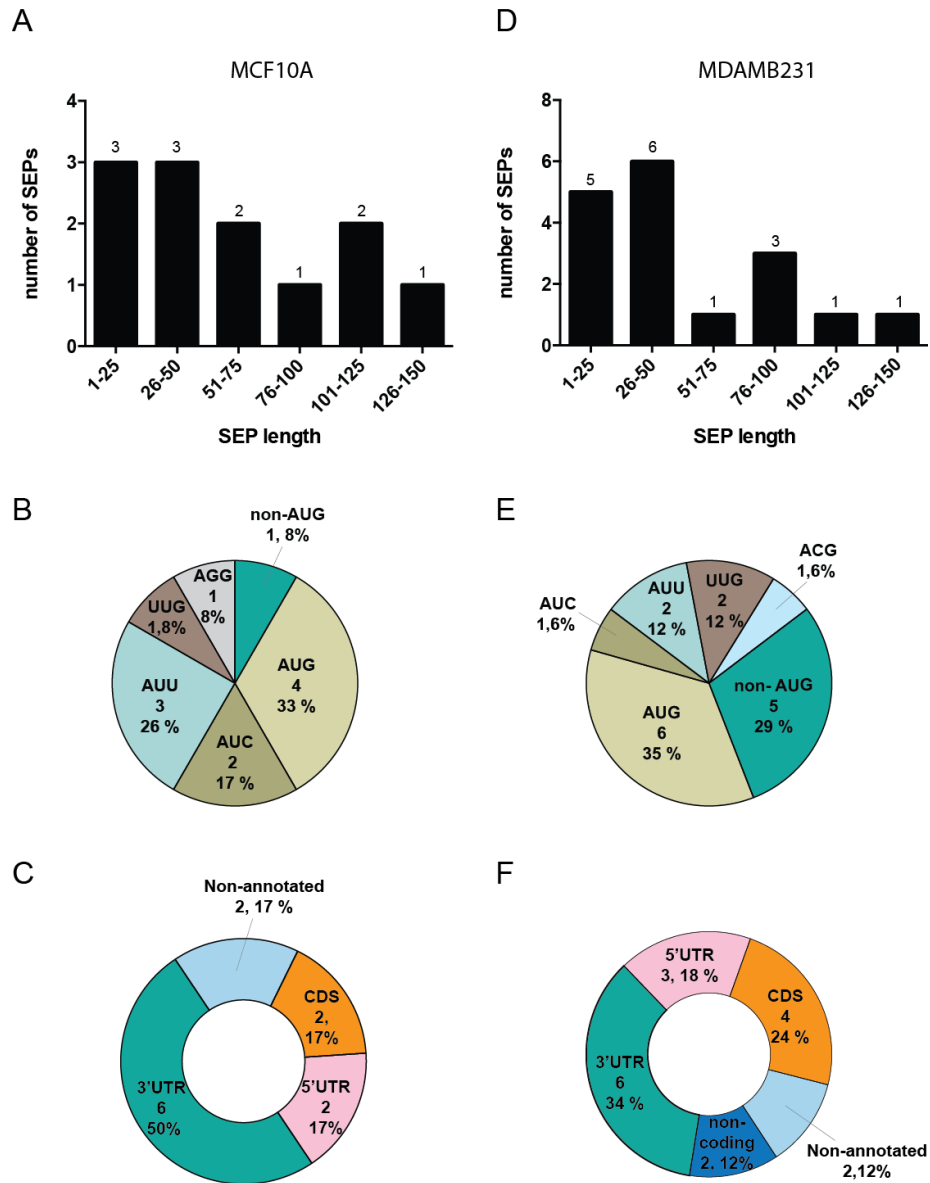


Figure 2.10. The characteristics of SEPs detected in MCF10A and MDAMB231 cell lines. (A) The length distribution of the SEPs, (B) the start codon usage of the SEPs, (C) the SEPs mRNA annotation by RefSeq in MCF10A cells. Similarly (D) The length distribution of the SEPs, (E) the start codon usage of the SEPs, (F) the SEPs mRNA annotation by RefSeq in MDAMB231 cells.

2.2.6. SEPs Are in Human Tissue

To determine whether we could find SEPs in human tissue, we used the protein database generated from K562 cells (this was the largest database we had) and analyzed a human breast cancer tissue biopsy by PAGE + LC-MS/MS. This analysis yielded 25 SEPs, 22 of which were novel and three that were also found in K562 cells. One SEP found on the MYBL2 RNA (MYBL2-SEP) was found in every sample we analyzed (tumor sample, MCF10A, MDAMB231, and K562 cell lines) indicating that some SEPs are ubiquitous and may serve broad biological roles.

These newly identified 25 tissue-derived SEPs (tdSEPs) were then analyzed to estimate the lengths of the sORFs, their initiation codon usage, and whether the RNAs that produce these SEPs are annotated or non-annotated. The SEP length for these tdSEPs varies between 15-138 amino acids, the percentage of AUG usage is 24%, and most are derived from non-annotated RNAs (80%), which is consistent with data obtained from cell lines (i.e. K562, MCF10A, and MDAMB231) (Figure 2.11). These data support the idea that SEPs are ubiquitous and found in tissues as well, which further enhances the interest in this class of polypeptides.

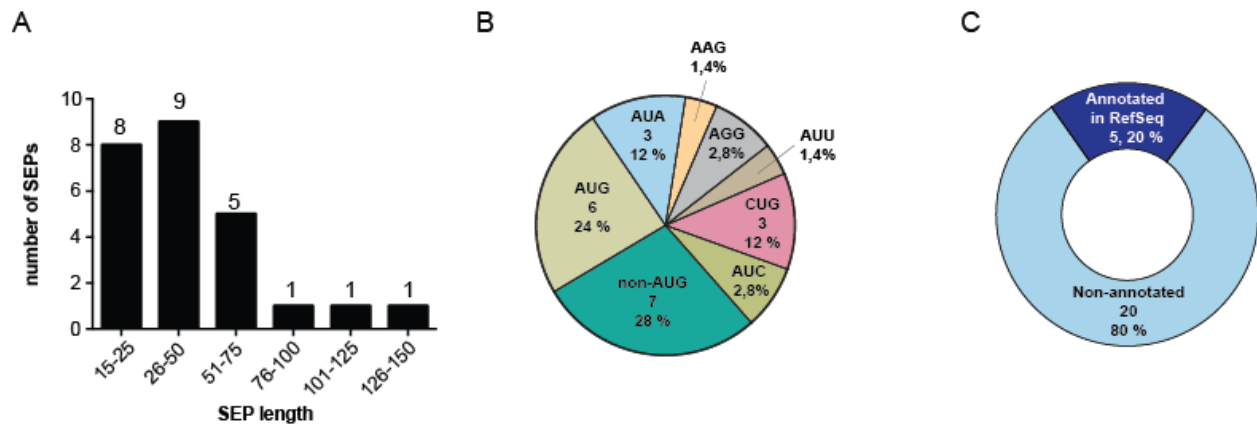


Figure 2.11. Discovery of 25 tumor derived SEPs (tdSEPs). (A) The length distribution (A), initiation codon usage (B) and RNA source (C) of tdSEPs were similar to the distributions seen for SEPs derived from cell lines.

2.3. Conclusions

We tested several parameters for our SEP discovery workflow and determined that running replicates (technical/biological) is the most efficient way to detect more SEPs. In total, we describe the discovery of an additional 237 human SEPs here (Table 2.2), demonstrating the prevalence of this class of polypeptides. With an increasing number of SEPs discovered through our shotgun profiling it became obvious that our previous approach for validation would not suffice and therefore we utilized a targeted MRM LC-MS/MS approach that relies on Skyline (27) to rapidly identify multiple peptides from a single SEP/sORF. Through the analysis of additional cell lines and a tumor biopsy, we also find that SEPs are ubiquitous and that at least some SEPs are specific to a cell line. This effort provides the necessary evidence for us to begin to start large-scale SEP profiling experiments. These experiments could be done by differentially profiling SEPs in disease models to identify SEPs that might cause disease or can serve as a biomarker.

Cell Line	Number of SEPs Detected	Number of Novel SEPs
K562	257	195
MCF10A	12	9
MDAMB231	17	11
Tumor	25	22
Total	311	237

Table 2.2. Total number of SEPs discovered from K562, MCF10A, MDAMB231 and tumor samples.

2.4. Materials and Methods

2.4.1. Cell Culture

K562 cells were grown in RPMI1640 medium supplemented with 10% FBS, penicillin and streptomycin at a density of $1-10 \times 10^5$ cells/ml. MCF10A cells were grown in MEGM complete medium (Life Technologies) and MDAMB231 cells were grown in DMEM medium supplemented with 10% FBS, penicillin and streptomycin. All cells were grown at 37°C under an atmosphere of 5% CO₂.

2.4.2. Tissue Sample

Tissue was obtained from the Massachusetts General Hospital (MGH) Department of Pathology as a de-identified sample. This was done in accordance with all of the rules and regulations of the Harvard IRB.

2.4.3. Peptidome Isolation from Cell Culture

Aliquots of K562, MCF10A and MDAMB231 cells (2×10^8 cells per experiment) were placed in Falcon tubes, washed three times with cold PBS, pelleted, and transferred into 1.5 ml Protein LoBind Tubes (Eppendorf). Boiling water (300 μ l) was directly added to the cell pellet and the cells were boiled for an additional 20 min. This step eliminates protease activity to maintain the integrity of the peptidome for subsequent LC-MS analysis. After cooling the samples on ice, the cells were sonicated on ice for 20 bursts at output level 2 with a 30% duty cycle (Branson Sonifier 250; Ultrasonic Converter). Acetic acid was added to the cell lysate until the final concentration of acetic acid was 0.25% by volume. The sample was then centrifuged at 14,000 x g for 20 min at 4 °C to precipitate large proteins and reduce the complexity of the sample. The supernatant was passed through a 30-kDa molecular weight cutoff (MWCO) filter and the small proteins and polypeptides were isolated in the flow-through. An aliquot of the flow-through was taken for a BCA assay to measure the protein concentration. The remaining sample was then evaporated to dryness at low temperature in a SpeedVac and used for LC-MS analysis.

In cases where PAGE analysis was used, this supernatant was loaded onto a 16% Tricine gel (Novex, 1.0 mm) and run at 120 V for 80 min instead of being passed through a MWCO filter. This gel was stained with Coomassie blue and then destained using standard protocols. Dual Xtra Standards (Bio-Rad) was used as the molecular weight marker and the gel was sectioned below the 15 kDa marker to afford three sections: 2-5kDa, 5-10kDa and 10-15kDa. Each gel slice was placed in 1.5 ml Protein LoBind Tubes (Eppendorf) and washed with 1 ml of 50% HPLC grade acetonitrile in water three times.

2.4.4. Peptidome Isolation form Tissue

Frozen human breast tumor sample (~200 mg) was immersed in boiling water (200 μ L) for 10 minutes. This step denatures proteins and eliminates proteolytic activity. The aqueous fraction was collected and saved in a clean tube, and the tissue was dounce-homogenized in 500 μ L of ice-cold acetic acid (0.5% v/v). The aqueous fraction and the homogenate were combined and centrifuged at 20,000 x g for 20 min at 4 °C. The supernatant was transferred to a new Lo-Bind tube and evaporated to dryness at low temperature in a SpeedVac. The dried sample was suspended in PBS and loading dye, followed by separation in a 16% Tricine gel (Novex, 1.0 mm). The excised gel bands (<15kDa) were analyzed by LC-MS/MS as described below.

2.4.5. ERLIC Fractionation (20,21)

After trypsin digest the samples were dried in a speed vac and suspended in ERLIC buffer A (90% acetonitrile 0.1% acetic acid). Samples were then fractionated using an HPLC (Agilent 1200 HPLC) equipped with an ERLIC column (PolyWAX LP Column, 200 x 2.1 mm, 5 μ m, 300 Å (PolyLC Inc)). Samples were separated using a stepwise gradient with the following steps: 0-5 min., 0%B; 5-15 min., 0-8% B; 15-45 min., 8-35% B; 45-55 min., 35-75% B; 55-60 min., 75-100% B; 60-70 min., 100% B (A: 90% acetonitrile, 0.1% formic acid; B: 30% acetonitrile, 0.1% formic Acid). An automated fraction collector was used to collect 25 equivalent fractions that were concentrated then analyzed by LC-MS/MS.

2.4.6. LC-MS/MS Analysis

ERLIC samples were digested prior to ERLIC and did not require any additional sample PREPL prior to LC-MS. Gel slices from PAGE separation were extracted and then digested with trypsin overnight. The resulting peptide mixture was separated from any residual gel slices and analyzed on an Orbitrap Velos Hybrid Ion Trap Mass Spectrometer (Thermo Fisher Scientific). Regions between 395 –1600 m/z ions were collected at 60K resolving power for the MS1 and this data was used to trigger MS/MS in the ion trap for the top 20 ions in the MS1 (i.e. Top 20 experiment). Active dynamic exclusion of 500 ions for 90 sec was used throughout the LC-MS/MS method. Samples were trapped for 15 minutes with flow rate of 2 $\mu\text{l}/\text{min}$ on a trapping column 100 micron ID packed for 5cm in-house with 5 μm Magic C18 AQ beads (Waters) and eluted onto 20 cm x 75 micron ID analytical column (New Objective) packed in-house with 3 μm Magic C18 AQ beads (Waters). Peptides were eluted with 300 nl flow rate using a NanoAcquity pump (Waters) using a binary gradient of 2-32% B over 90 minutes (A: 0.1% formic acid in water; B: 0.1% formic acid in acetonitrile).

2.4.7. Data Processing

The SEQUEST algorithm (22, 23) was used to analyze the acquired MS/MS spectra using a database derived from three-frame translation from the RNA-Seq data for that cell line. RNA-Seq data from K562, MCF10A or MDAMB231 cell lines was assembled into a transcriptome using Cufflinks (24) and then translated in three (forward) frames *in silico*. The search against this database was performed using the following parameters: variable modifications, oxidation (Met), N-acetylation, semi-tryptic 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 at 1%. For this purpose the Sf score of greater than 0.7 was the required with a mass accuracy of less than 3.5 ppm. After analysis, the data was filtered based on a combination of the preliminary score, the cross-correlation and the differential between the scores for the highest scoring protein and the second highest scoring protein. A list of peptides passed the search criteria were then searched against the Uniprot human (SwissProt) protein database using a string-searching algorithm. Peptides found to be identical and overlap with part of annotated proteins were eliminated from the list. The remaining peptides were then searched one more time against non-redundant human protein sequences using the Basic Local Alignment Search Tool (BLAST) (25, 26). Peptides that were identical or different by one amino acid from the nearest protein match were discarded. Peptides with more than two missed cleavages were also removed at this point. The final list of peptides, candidate SEPs, were searched against Human Reference (RefSeq) RNA sequences using BLAST to assess their location relative to the annotated transcripts, which can be categorized into 5'UTR, 3'UTR, CDS, and non-coding. If the peptides had no match in the RefSeq RNA sequences, then they were derived from RNAs that were present in the RNA-Seq data that had not been annotated in RefSeq (i.e. non-annotated RNAs).

2.4.8. RNA-seq Library Preparation and Transcriptome Assembly

Total RNA (3,000 ng) was purified from MCF10A and MDAMB231 cell lines using RNeasy Kit (QIAGEN) according to the instructions provided by the manufacturer. cDNA libraries with paired-end, indexed adapters were created using the Illumina TruSeq RNA sample prep kit. Two libraries were pooled and sequenced on a single lane of a HiSeq2000 machine.

RNA-Seq reads were aligned to the human genome (hg19) using TopHat (version V2.0.4), and transcriptome assembly was performed using Cufflinks (version V2.0.2)(24).

2.4.9. Skyline Targeted MRM LC-MS/MS Peptidomics

Sequences for SEPs were submitted in FASTA format to Skyline (version. 2.1.0.4936) (27) for analysis. The goal was to identify peptides from these sequences that are most amenable for targeted proteomics using multiple reaction monitoring (MRM). Skyline predicts transitions for each peptide and we use all of transitions in a targeted MRM experiment to identify the presence or absence of the peptide. We must detect at least three transitions for a given peptide to determine that it is present in the sample. The output from Skyline is imported directly into a targeted method for analysis with a TSQ Quantum Ultra™ Triple Stage Quadrupole Mass Spectrometer (i.e. a triple quad (QQQ), Thermo Fisher Scientific). Peptide samples were analyzed using the TSQ with a 90-minute gradient and targeted multiple reaction monitor (MRM) tandem mass spectrometry using the aforementioned Skyline method. Samples were trapped for 15 minutes with flow rate of 2 µl/min on a trapping column 100 micron ID packed for 5 cm in-house with 5µm Magic C18 AQ beads (Waters) and eluted with a gradient to 20 cm 75 micron ID analytical column (New Objective) packed in-house with 3 µm Magic C18 AQ beads (Waters). Peptides were eluted with 300 nl flow rate using a NanoAcquity pump (Waters) using a binary gradient of 2-32% B over 180 minutes (A: 0.1% formic acid in water; B: 0.1% formic acid in acetonitrile).

2.5. References

1. P. Bertone *et al.*, Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246 (2004).
2. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63 (2009).
3. J. M. Johnson, S. Edwards, D. Shoemaker, E. E. Schadt, Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *TRENDS in Genetics* **21**, 93-102 (2005).
4. P. Kapranov *et al.*, RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488 (2007).
5. U. Nagalakshmi *et al.*, The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).
6. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515 (2010).
7. A. M. Khalil *et al.*, Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences* **106**, 11667-11672 (2009).
8. I. A. Mitchell Guttman *et al.*, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).
9. W. Gish, D. J. States, Identification of protein coding regions by database similarity search. *Nature genetics* **3**, 266-272 (1993).
10. A. V. Kochetov, AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. *Bioinformatics* **21**, 837-840 (2005).
11. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
12. S. A. Slavoff *et al.*, Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature chemical biology*, (2012).

13. M. C. Frith *et al.*, The abundance of short proteins in the mammalian proteome. *PLoS genetics* **2**, e52 (2006).
14. J. L. Guénet, The mouse genome. *Genome Research* **15**, 1729-1740 (2005).
15. N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, J. S. Weissman, The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *nature protocols* **7**, 1534-1550 (2012).
16. A. G. Schwaid *et al.*, Chemoproteomic discovery of cysteine-containing human sORFs. *Journal of the American Chemical Society*, (2013).
17. H. Liu, R. G. Sadygov, J. R. Yates, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry* **76**, 4193-4201 (2004).
18. M. Oyama *et al.*, Diversity of translation start sites may define increased complexity of the human short ORFeome. *Molecular & Cellular Proteomics* **6**, 1000-1006 (2007).
19. B. Vanderperre *et al.*, Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS one* **8**, e70698 (2013).
20. P. Hao, H. Zhang, S. K. Sze, Application of electrostatic repulsion hydrophilic interaction chromatography to the characterization of proteome, glycoproteome, and phosphoproteome using nano LC-MS/MS. *Methods in molecular biology* **790**, 305-318 (2011).
21. P. Hao *et al.*, Enhanced separation and characterization of deamidated peptides with RP-ERLIC-based multidimensional chromatography coupled with tandem mass spectrometry. *Journal of proteome research* **11**, 1804-1811 (2012).
22. C. L. Gatlin, J. K. Eng, S. T. Cross, J. C. Detter, J. R. Yates, Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Analytical chemistry* **72**, 757-763 (2000).
23. M. J. MacCoss, C. C. Wu, J. R. Yates, Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical chemistry* **74**, 5593-5599 (2002).
24. C. Trapnell *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578 (2012).
25. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).

26. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
27. B. MacLean *et al.*, Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968 (2010).
28. M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-i282 (2011).
29. H. S. Chittum *et al.*, Rabbit β -globin is extended beyond its UGA stop codon by multiple suppressions and translational reading gaps. *Biochemistry* **37**, 10866-10870 (1998).
30. S. Tork, I. Hatin, J. P. Rousset, C. Fabret, The major 5' determinant in stop codon read-through involves two adjacent adenines. *Nucleic acids research* **32**, 415-421 (2004).
31. S. L. Salzberg, A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Computer applications in the biosciences: CABIOS* **13**, 365-376 (1997).
32. P. Gunning, J. Leavitt, G. Muscat, S.-Y. Ng, L. Kedes, A human beta-actin expression vector system directs high-level accumulation of antisense transcripts. *Proceedings of the National Academy of Sciences* **84**, 4831-4835 (1987).
33. S. Katayama *et al.*, Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566 (2005).
34. N. Jessani, Y. Liu, M. Humphrey, B. F. Cravatt, Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. *Proceedings of the National Academy of Sciences* **99**, 10335-10340 (2002).
35. A. Albini *et al.*, A rapid in vitro assay for quantitating the invasive potential of tumor cells. *Cancer research* **47**, 3239-3245 (1987).

Chapter 3

Improved Identification of Protein-Coding smORFs by an Optimized Proteogenomics Platform

This chapter was adapted from Ma, J.; Diedrich, JK.; Jungreis, I.; Donaldson, C.; Vaughan, J.; Kellis, M.; Yates, JR.; and Saghatelian, A. Improved Identification of smORF-Encoded Polypeptides. *Analytical Chemistry* (Submitted)

3.1. Introduction

A search for genes that could protect neurons from death caused by the neurotoxic peptide abeta, a key molecule in Alzheimer's disease, identified a non-annotated anti-apoptotic 24-amino acid peptide called humanin (1). Unlike classical neuropeptides, which are generated from proteolysis of longer prohormones (2-6), humanin is produced directly from the ribosomal translation of a short open reading frame (sORF). Because of this difference in production, humanin was distinct from neuropeptides and peptide hormones. Though sORFs were previously known, it was unclear whether they were translated. Humanin indicated that some sORFs are translated, and that the peptides generated from these sORFs are biologically active.

Later on, a protein-coding sORF called tarsal-less or polished rice (tal/pri) was discovered in flies (7, 8). Deletion of tal/pri resulted in loss of segmentation of the embryo, and a truncated limb and a missing tarsus in the adult fly. Tal/pri contains several protein-coding sORFs that produce three 11 amino acid or a 32 amino acid peptide. Along with the discovery of protein-coding sORFs in bacteria (9-11), plants (12), and other eukaryotes (12-19), tal/pri and humanin indicated that genomes might harbor many protein-coding sORFs (< 100 codons), some of which are biologically active (20-22). The identification of several protein-coding sORFs led to them being called small ORFs or smORFs, to distinguish them from sORFs that are not translated. And the smORF-encoded polypeptides are termed SEPs to differentiate these polypeptides from neuropeptides and peptide hormones that are generated by proteolysis of the prohormones.

The discovery of non-annotated smORFs is necessary to understand the protein-coding potential of genomes, and is a crucial step in the characterization of these novel genes.

Traditional computational methods for assigning protein-coding ORFs are not as robust for identifying smORFs. But newer, more reliable, computational strategies have identified hundreds to thousands of non-annotated smORFs in genomes (*13, 14, 20, 23*). Empirical methods for smORF discovery have relied on genomics or proteomics methods, or combinations of both. Genomics methods are mostly based on ribosome sequencing (Ribo-Seq) (*13, 24*), which footprints the position of the ribosome on RNAs. The presence of ribosome-bound RNA is interpreted as the active translation of that particular region of the RNA. Ribo-Seq has led to the discovery of non-annotated smORFs in the fly and mouse genomes (*13, 24, 25*). Ribo-Seq provides excellent coverage of the ORFeome and reports on sights of active RNA translation.

Though proteomics techniques are not as comprehensive as Ribo-Seq methods, they validate that the SEPs are stable and abundant enough to be detected. The discovery of SEPs and smORFs requires the combination of proteomics with next generation RNA-Seq, also termed proteogenomics (*26, 27*). We and others have successfully applied proteogenomics methods for SEP identification. In this approach, the proteome is enriched for low molecular weight peptides, a fraction presumably containing most of the SEPs, and then this fraction is digested and analyzed by liquid chromatography-mass spectrometry (LC-MS) (*16, 18, 19, 26*). The resulting dataset is then interrogated using a protein database that is generated from the three-frame translation of the RNA-Seq data. Removal of known proteins identifies non-annotated SEPs, which simultaneously reveals a new smORF. Other methods such as the computational prediction of new ORFs followed by mass spectrometry have also successfully identified new SEPs.

Because SEPs are short and not very abundant, we typically only observe a single peptide from a SEP. In some cases, the quality of the MS2 peptide is not sufficient to confidently

identify a new SEP. We previously improved the number of SEPs we detected by varying methods for proteome fractionation but because of the lower abundance of SEPs we found variability in the SEPs detected in replicate experiments. There is known stochasticity in proteomics identifications, which is likely exacerbated because we only identify a single peptide for each SEP. Therefore, we reasoned that SEP detection might require modifications to the protocol and instrument settings from a typical proteomics experiment to improve the robustness of our platform. Here, we tested several enrichment/extraction conditions and varied instrument parameters, including the type of fragmentation, to improve the quality and reproducibility of our SEP profiling experiments. These improvements enabled us to quantify SEPs under different biological conditions. And lastly, we demonstrate how evolutionary analysis of the identified SEPs can be used to determine those SEPs that are most likely functional. Together these improvements provide a robust platform for the identification of SEPs and smORFs with the greatest likelihood of being biologically active.

3.2. Results and Discussions

3.2.1. Enrichment technique can improve number of SEPs detected

To help characterize the functions of smORFs and SEPs, we want to be able detect as many SEPs as possible using proteomics. Our experience has been SEP detection requires an enrichment step because they are of lower abundance and generate fewer tryptic peptides. In a complex mixture such as total cell lysate, detecting small and low abundant proteins is challenging because mass spectrometry is biased towards detecting more abundant species (28). Here, we compare different enrichment methods by the number of known and unknown SEPs detected. As mentioned, SEPs are defined as polypeptides that are less than 150 amino acids.

The total proteome is prepared by boiling cells to inactivate all proteolytic activity and then lysing the cells by sonication. We used three methods to enrich the < 150 kDa proteome: 1) acetic acid precipitation; 2) molecular weight cutoff (MWCO) filtration (30 kDa); or 3) solid phase extraction (SPE). A BCA assay quantified the protein concentrations in each of these enriched samples and an equal amount of total protein was analyzed by SDS-PAGE gel (Figure 3.1B).

The results are clear. The 30 kDa MWCO performed poorly, while the acid precipitation and SPE provided excellent enrichment of the lower molecular weight proteomes. Analysis of total lysate by SDS-PAGE reveals that a majority of the proteome is larger than 25 kDa. Acetic acid precipitation aggregates larger proteins leaving behind lower molecular weight proteins solution. SDS-PAGE of the solution after acetic acid precipitation led to the majority of the signal coming from proteins less than 25 kDa. Previously, we had relied on MWCO filtration to enrich the lower molecular weight proteome. We were surprised that MWCO gave such a poor recovery of proteins less than 25 kDa. The SPE method using a C8 column was originally developed to improve the recovery of peptide hormones for radioimmunoassays or ELISAs by removing larger molecular weight proteins. Applying this method to enrich the lower molecular weight proteins gave excellent results by SDS-PAGE.

Next we demined whether the results we measured by SDS-PAGE correlated with the number of known and unknown polypeptides of < 150 amino acids, SEPs, we can detect using proteomics. Enriched and non-enriched proteome samples were reduced, alkylated, and trypsin digested followed by LC-MS/MS analysis. Samples were analyzed using a 6-hour gradient on a Q-Exactive mass spectrometer set to a top 10 mode. The decoy database searching was used to identify the acquired MS/MS spectra using two databases (Figure 3.1A). One database was

derived from three-frame translation of K562 cell line RNA-seq data; the other is human Uniprot database. First we analyzed the data using the human Uniprot database. This analysis detected 1901, 1072, 170, and 1188 total proteins from non-enriched, acetic acid precipitated, MWCO, and C8 column, respectively.

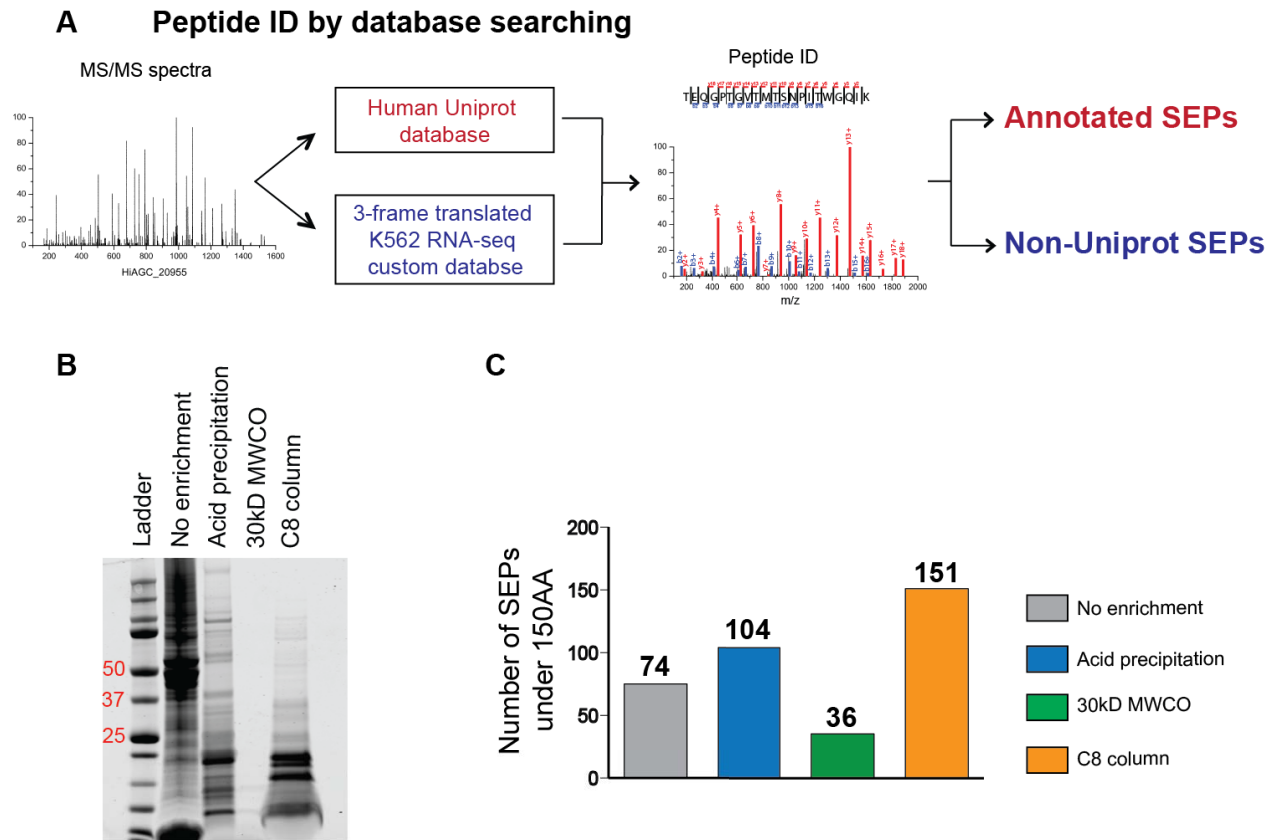


Figure 3.1. SEP discovery workflow and enrichment techniques. (A) Overview of SEP discovery workflow. MS/MS spectra were searched against two databases: Human Uniprot database and 3-frame translated RNA-seq custom database for peptide identification. Peptides that uniquely match to Uniprot protein under 150 amino acids correspond to annotated SEPs. Peptides that are derived from RNA-seq transcripts but do not overlap with Uniprot annotated SEPs are called non-Uniprot SEPs in this study. (B) Analytical gel representing peptidome enrichment efficiency. Cells were lysed by boiling in water and sonication, followed by different enrichment methods: acid precipitation, 30 kDa MWCO filter, C8 column. 30 μ g of enriched protein from each method were run on a SDS-PAGE gel to show recovery and enrichment efficiency. Acid precipitation and C8 column both enriched peptidome under 25 kDa compared to sample that was not enriched. 30 kDa MWCO filter treated sample showed the poorest recovery. (C) Total number of SEPs detected in

Figure 3.1. (Continued) samples treated with different enrichment methods described in (B). The most SEPs (151 SEPs) were detected in C8 column enriched sample followed by acid precipitated sample (104 SEPs). 30 kDa MWCO filter treated sample gave the worst recovery and the least number of SEPs detected, which agree well with the analytical gel analysis.

and SPE enriched samples, respectively. Of these we found 70, 96, 35, 143 SEPs from the non-enriched, acetic acid precipitated, MWCO, and SPE enriched samples, respectively. These represent known or annotated SEPs, but previous work from our lab and others demonstrated the existence of a significant fraction of novel or non-annotated SEPs. To identify novel SEPs by proteomics requires a searchable database that contains potential non-annotated SEP sequences.

We created a database with non-annotated SEP sequences by three-frame translation of RNA-Seq data from K562 cells. The raw sequence data is assembled into transcripts using Cufflinks followed by in silico translation. This database should contain any possible protein sequence generated from RNAs in these cells. In order to capture non-annotated SEPs, we search the MS/MS spectra against the K562 translated RNA-seq database. This analysis provides a list of proteins that are mostly comprised of known (annotated) proteins and SEPs. To identify non-annotated SEPs, we removed all peptides that match to the Uniprot database. The MS/MS spectra for the remaining peptides were subjected to visual inspection to ensure confident peptide identification. Finally, the transcripts that produced the identified peptides were searched in RefSeq and RNA-seq database to identify the smORFs.

After accounting for known and unknown SEPs, we identified 5, 8, 1, and 8 SEPs using the non-enriched, acetic acid precipitated, MWCO, and SPE enriched samples, respectively. The total number of SEPs detected, annotated or novel, from each enrichment technique is shown in Figure 3C. These values correlate with the apparent recovery by SDS-PAGE. In our previous

SEP detection workflows, we had used the 30 kDa MWCO to fractionate the proteome, but these new results indicate that the acetic acid precipitation and C8 SPE methods are superior. These findings will greatly improve the detection and analysis of SEPs using proteomics.

3.2.2. Isolating SEPs from cells

Next, we compared several different methods for isolating SEPs from the lung cancer cell line A549 (i.e. extraction methods) (Figure 3.2). We chose to use another cell line because A549 are adherent cells lines which are much more typical than the non-adherent K562 cells. Second We tested four different extraction methods: 1) water + sonication; 2) lysis buffer + sonication; 3) acetic acid (1N) + HCl (0.1N); or 4) lysis buffer. After extraction, we used SPE to prepare the sample for LC-MS/MS. We searched the proteomics data against the Human Uniprot database and three-frame translated RNA-seq custom database for peptide identification. Samples extracted in the lysis buffer resulted in detecting the most SEPs while acid extraction resulted in the least number of SEPs detected. Overall, the lysis buffer performs better than water or acid alone, likely due to more efficient protein solubilizaation. Boiling did not seem to have a strong effect. The number of SEPs identified with or without boiling are similar, and share over 70 percent of the SEPs detected (Figure 3.2C). The less number of SEPs detected in samples boiled and sonicated could be due to additional manual handling of the samples. Overall, the combination of extracting cell lysate in the lysis buffer and enriched with C8 column provided the highest recovery of small peptidome and the largest number of SEPs detected.

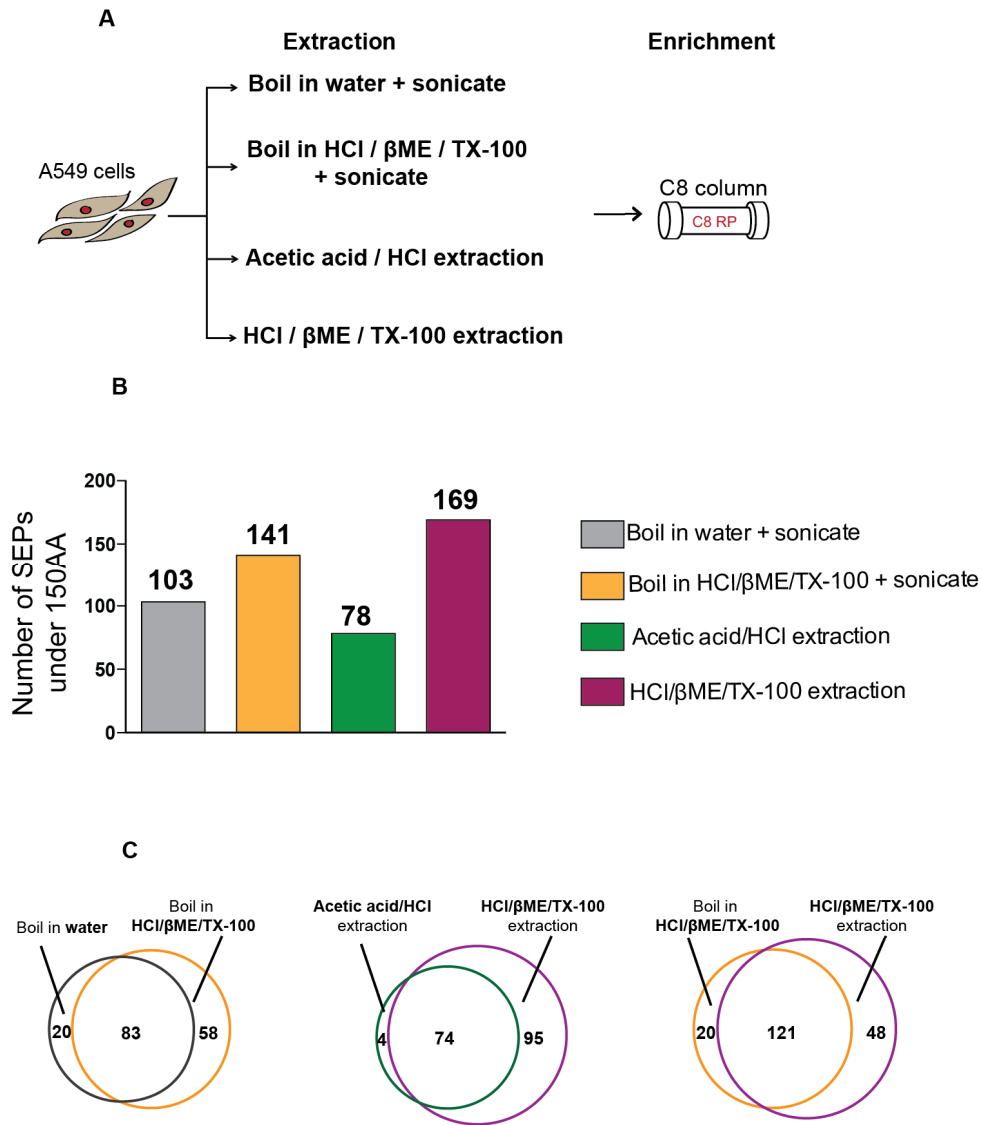


Figure 3.2. Comparison among different extraction methods. (A) Schematic of the sample preparation: Cell lysis was obtained by four different extraction methods: boil the cell pellet in water then sonicate, boil the cell pellet in lysis buffer (50 mM HCl, 0.1% b-ME, 0.05% Triton X-100) then sonicate, extract cells in acid (1 N acetic acid, 0.1 N HCl), or extract cells in lysis buffer. All followed by enrichment using C8 column. (B) Total number of SEPs detected in different extraction methods. (C) Pairwise comparison of the extraction methods and overlap in SEPs detected.

3.2.3. Assessing the impact of mass spectrometry parameters and SEP detection

For SEP discovery, spectral quality is crucial because SEPs are low abundant and only one peptide is detected per SEP in most cases. Therefore, the confidence of the identified peptide is of high importance. Confidence of the peptide identification is dependent on good quality MS/MS spectra with manual validation. High percent coverage of b- and y- ions (5 consecutive ions) and low background (avoid chimeric) are important.

In previous studies, the data was acquired using a Orbitrap Velos hybrid ion trap mass spectrometer (Thermo Fisher Scientific) with CID and low resolution MS/MS spectra acquisition. Low resolution spectra detected in the linear ion trap can often have high background noise, especially for low abundant species such as SEPs. High resolution data, detected in the orbitrap are less sensitive by nature (require more ions for detection than an ion trap) but they are less plagued by the issue of noisy spectra. Prefiltering in our ProLucid database search filters out spectra that contain less than 8 peaks in the MS/MS scan. Thus a very low intensity peak can produce a very sparse MS2 in the orbitrap due to lack of signal but it will be filtered out instead of possibly being missassigned as a SEP. Low resolution data in an iontrap however will be much more sensitive but in the case of very low intensity peaks it will generate a noisy spectrum that may or may not have peaks randomly matched to a SEP, and manual validation of the spectra will be challenging.

In SEP identification and validation sequence coverage is one of the most important parameters to be addressed. HCD is known to provide better sequence coverage than CID. Sequential fragmentation of the peptide within the HCD cell provides increased sequence coverage, provided the HCD energy is adequate for the peptide (29). This is of particular

advantage in the case of SEP detection as the peptide must have high percentage of coverage and a certain number of consecutive ions observed in order to pass manual validation.

Shown in Figure 3.3A and B is a representative SEP that was detected with the Fusion by low resolution CID (A) and high resolution HCD (B). While both of these spectra were good enough quality to pass filters of database search it can be seen that the low resolution CID spectrum is of lower quality and has unannotated peaks. Figure 3.3C shows the same peptide identified with Q Exactive HCD. It can be seen that spectra is very similar in appearance, the quality of the Q Exactive spectra is slightly higher due to the AGC setting and fill times being adjusted to provide increased sensitivity. (Unfortunately AGC is calculated in a slightly different manner in the two instruments so a direct correlation between identical settings could not be made, we simple utilized a Fusion method which is standard in the lab for comparison to the optimized Q Exactive methods.)

Prior to the collection of the above data we also optimized the settings of the Q Exactive to produce data that was most suitable for our purpose. Shown in Figure 3.3D and E is the peptide TEQGPTGVTMTSNPITQGQIK identified by two Q Exactive methods. Figure 3.3E was collected with a standard Q Exactive method utilized in the lab. The spectra provide good y-ion series coverage but detection of b-ions is limited. While this is considered a decent spectrum and we can confidently assign the spectra to the SEP we also wanted to improve upon the method to be able to increase the data quality of the collected spectra, even if it meant sacrificing scan speed. We subsequently acquired data from the same sample but with an increased AGC target and increased maximum fill times for both the MS1 and MS2 scans. Shown in Figure 3.3D is a spectra from the same peptide acquired with sensitive settings. The spectra shows near complete coverage of b- and y- ions and good signal to noise. In Figure 3.3 D, it can be seen that

while we do sacrifice scan speed and collect fewer spectra, the quality of those spectra are higher, and thus a larger percentage of them lead to IDs (and hopefully SEPs). All data presented herein was collected under the “sensitive” settings to insure good spectral quality.

3.2.4. Label-free SEP quantitation

SEPs are expressed at low levels and their expression is stochastic, therefore makes quantitation difficult. Two sets of HEK293 cells were prepared in biological triplicates. One set was treated with 10 μ M sodium arsinite for 24 hours and the other without treatment. Cellular proteins were extracted using the lysis buffer followed by centrifugation 20,000 x g for 20 minutes at 4 °C to remove any insoluble particulates. Then the cell lysates were processed for LC-MS/MS analysis.

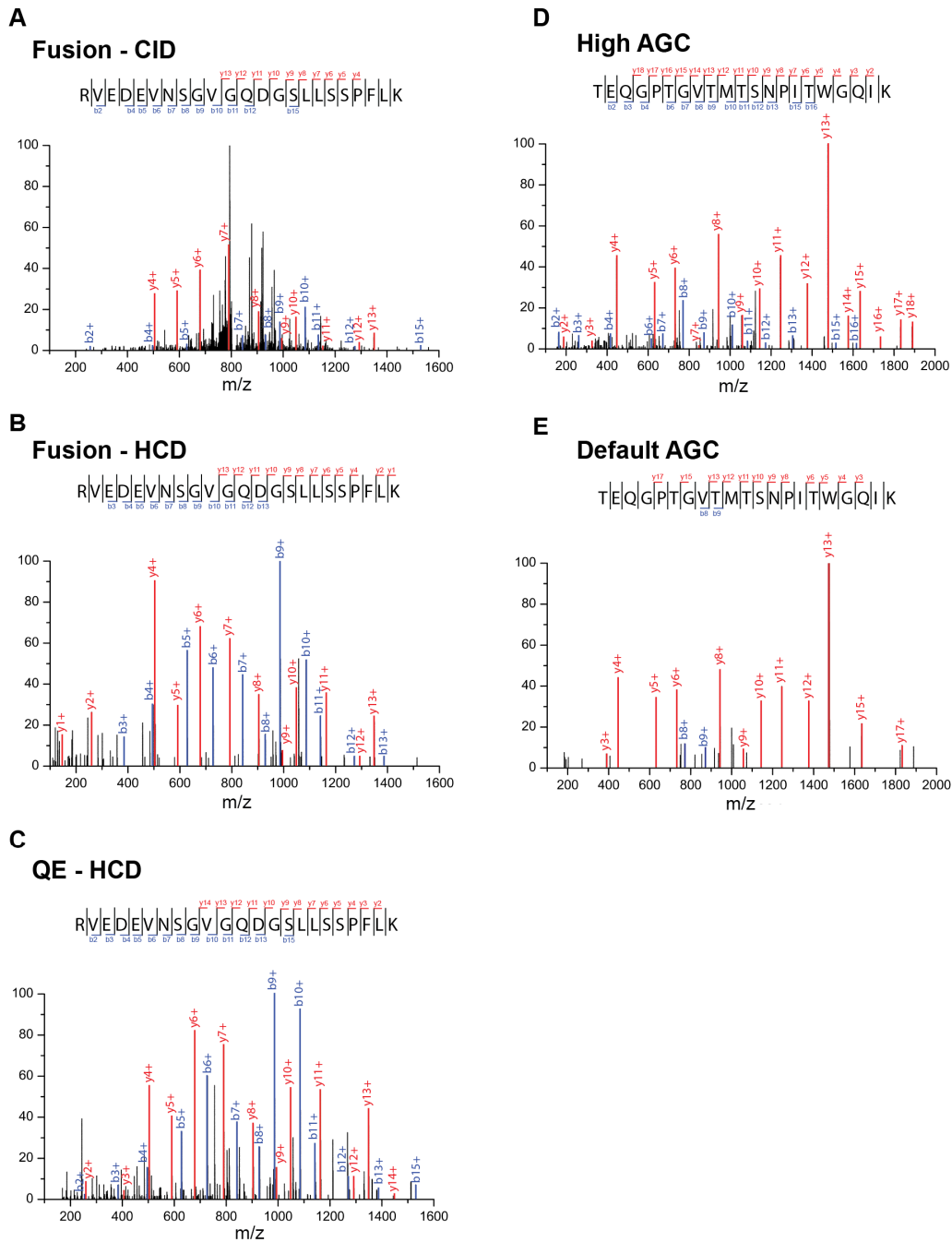


Figure 3.3. Comparison of MS/MS spectra acquired by different instruments and different instrument settings. (A), (B), (C) showing MS/MS spectra identified to be peptide sequence: RVEDEVNSGVGQDGSLLSSPFLK. (A) MS/MS spectrum acquired by Fusion mass spectrometry with low resolution CID method. Spectrum has high background noise and unannotated large peaks. (B) MS/MS spectrum for the same peptide acquired with Fusion mass spectrometry with high resolution HCD method. Peptide has better b- and y- ion coverage and the spectrum has improved resolution with much lower background. This significantly increased the confidence in peptide

Figure 3.3. (Continued) identification. (C) The same peptide MS/MS spectrum acquired using QE instrument with HCD setting. The quality is very comparable with that of Fusion mass spectrometry. It provided near complete b- and y- ion coverage of the peptide with high resolution MS/MS spectrum. (D), (E) compare the peptide TEQGPTGVTMTSNPITQGQIK identified by two Q-Exactive methods. (D) MS/MS spectrum acquired with sensitive setting shows near complete coverage of b- and y- ions and good signal to noise. (E) MS/MS spectrum acquired with standard setting provides good y-ion coverage but loses b- ion coverage. All mass spectrometry data was acquired using QE mass spectrometry with HCD fragmentation method and high AGC setting for confident peptide identification.

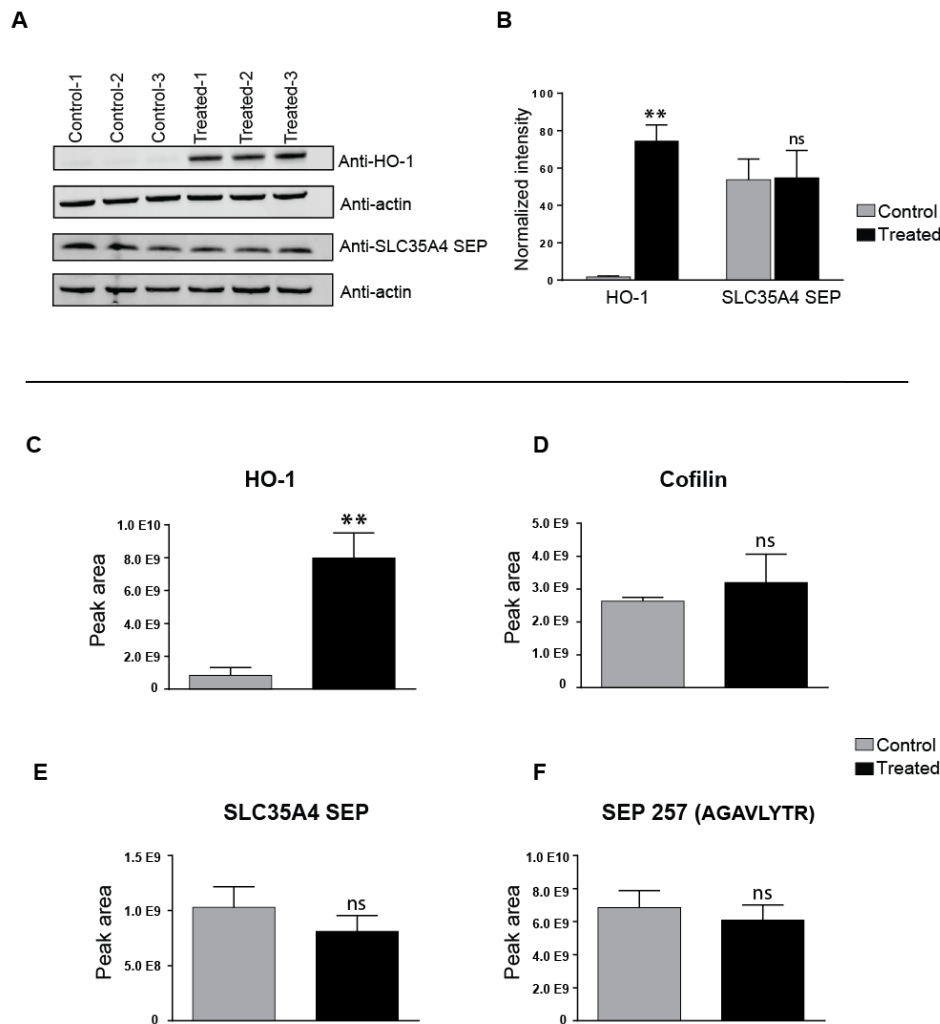


Figure 3.4. Changes in SEP and protein expression level upon arsenite treatment. HEK293 cells were treated with 10 μ M sodium arsenite for 24 hours. (A) Western blot analysis to detect SEP expression level change upon on arsenite treatment

Figure 3.4. (Continued) (10 μ M, 24 hours). Both control and treated samples were prepared in biological triplicates. Heme oxygenase 1 (HO-1) is reported to show up-regulation when treated with arsenite, and this was used as positive control. SLC35A4 SEP expression level change was also blotted using in-house raised antibody against the SEP. (B) The intensity of the bands on the blot from (A) was quantified by LiCor Odyssey CLx and normalized by beta-actin. ** $p < 0.01$, ns: not significant. (C)-(F) Peak area of MS1 of the detected peptide that corresponds to each protein or SEP was extracted using Skyline software. Expression level change in HO-1 and SLC35A4 SEP show agreement with western blot analysis. In addition, cofilin, a small house keeping protein and SEP257 with detected peptide AGAVLYTR are quantified the same way.

Heme oxygenase 1 (HO-1) is known to be up regulated upon arsenite treatment in a previous proteomics study (30). We validated this by western blot showing HO-1 was highly expressed in arsenite treated samples ($p < 0.01$) (Figure 3.4A, B). In order to see whether this treatment affects SEP expression level, we generated an antibody against SLC35A4-SEP. We tested this antibody and it generates a band of the appropriate molecular weight and gives increased signal when SLC35A4 SEP is overexpressed. As mentioned, SLC35A4 mRNA is increased, but we did not detect a change by Western blot using our SCL35A4-SEP antibody. Andreev et al. has shown in their study that SLC35A4 transcript level increased right after arsenite treatment by ribosome profiling (31), however we did not detect a quantitative difference SLC35A4 levels (Figure 3.4A, B). Generating in-house antibody against SEPs for quantitation is extremely time consuming and impractical. We performed label-free quantitative analysis using Skyline software that extracts peak area of the detected peptides from MS1 by retention time and accurate mass. This allows us to quantitate relative protein or SEP expression level between two conditions. HO-1 and SLC35A4 SEP expression level analyzed by Skyline (Figure 3.4C, E) (32) is in strong agreement with the western blot, suggesting that this method is applicable. In addition, we looked at a small 18kDa housekeeping protein, Cofilin, and showed

that the expression level is not affected by arsenite treatment (Figure 3.4D). Finally, we compared another SEP expression level change, a SEP named SEP257 with detected peptide: AGAVLYTR and expressed in all samples (control and treated). This also did not show significant changes in translation upon arsenite treatment (Figure 3.4F). In summary, label-free analysis is a useful and reliable tool in quantitation of SEPs when studying SEP regulation and translational response to drug treatment, without additional labeling of samples. We have shown that the western blot analysis and label-free quantitation are in good agreement for detecting relative translational changes in annotated protein and SEP. Therefore this method can be applied to detect SEPs changes under various conditions when specific antibodies are typically not available.

3.2.5. Analysis of newly discovered SEPs

In this study, we detected 37 SEPs that are not annotated in Uniprot database (Table 1). SEPs are translated from various locations of the annotated transcripts such as 5'UTR (5 SEPs), 3'UTR (2 SEPs) and non-coding RNAs (6 SEPs), which suggest strong coding potential of these genomic regions previously thought to be non-coding. 21 SEPs (55%) initiate with AUG start codon, which the remaining 16 SEPs (45%) do not. This observation is in agreement with previous studies by our and other labs, indicating that a significant portion of SEPs can be translated from non-canonical AUG start codon. There are increasing number of SEPs detected across cell lines and tissues, and a handful of studies have shown that SEPs involve in important biological functions. In order to select which SEPs are most likely to be functional therefore following with biological studies, we performed PhyloCSF analysis.

PhyloCSF is a computational method to detect the evolutionary signature of functional protein-coding regions using substitutions and codon frequencies in a multi-species alignment

(33). This is a useful tool to predict protein-coding regions with conservation from a large database generated by high-throughput transcriptome sequencing such as RNA-seq. There are a total of 6 SEPs (Table 1. Highlighted in red) including ASNSD1 SEP, SLC35A4 SEP and MIEF1 SEPs are detected across cells lines studied (K562, HEK293, HeLa, A549) and also predicted in PhyloCSF with clear evolutionary signature, likely to have important biological function.




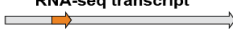
SEP location	RefSeq gene name	detected peptide	start codon	SEP length
 5'UTR	ASNSD1	IVVDELSNLKK QQQNSNIFFLADR QQQNSNIFFLADRTEMLSESK NILDELSNLKKYQEIENLDK	AUG	96
	MKKS	NDDIPEQDSLGLSNLQK VLVVMVPLVGLIHLGWYR SSPVFQIPKNDIPEQDSLGLSNLQK	AUG	63
	MIEF1	YTDRDFYFASIR	AUG	70
	SLC35A4	RVEDEVNSGVGQDGLSSPFLK	AUG	103
	STAR	KDEEPLREEAAAAAAAAAATPPLPHLPGNNAASDIQAVR	AUG	109
 3'UTR	FAM131A	PGAYGLSR	non-AUG	80
	PSIP1	IQVEQTRDELDLTDLSLD	AUG	63
 non-coding	c7orf49	TPANLTACDKDCVLHE	AUG	69
	LOC101928527	MQLVQESEEEK	AUG	54
	TUBA3FP	LLPLGASPAGVVGGLAPPR	AUG	85
	CCT6P1	AQLGVQAFADVLLVIPK	AUG	49
	LOC105372481	TPDSMFLAMLAVVSCASIGSGEPPTGN RDPWYSTVGLLPPVR DPWYSTVGLLPPVR TEQGPTGVTMTSNPITWGQIK	AUG	109
 RNA-seq transcript	SEP transcriptana not annotated in RefSeq	AGPGSEASTEAESGEGSGR DLSQQMQSDLDKADLSAR	AUG	97
		SVIFSSFLQEAASEAYLVGLFEDTNLCAIHAKE	non-AUG	58
		NIILEEGKEILVGDVGGQTVNNLYATFVK	non-AUG	76
		KMGALLESGLAEYLFDKHTLGDSDNES MGALLESGLAEYLFDKHTLGDSDNES	AUG	146
		AGAVLYTR	non-AUG	15
		LILWSCLGTIDYR	AUG	51
		SAAETVTRGGIMLPEK	non-AUG	85
		QSVVIPHIWSSSKP	non-AUG	40
		DPPLPPVPEAGSGAGDKPGPAR	non-AUG	98
		NMITETSQLADCAVLVAAGVGEFEAGISK	AUG	123
		TAFDEAIAELDTLSEESYKDSMLIMQLLR	non-AUG	105
		GIKQGVPDFK	non-AUG	22
		TLSNYNIQKESTLHLVLR	AUG	88
		VVLNSQPQVICPPQPPK	AUG	27
		EDPHVTHLQVAQDVTPEAAQISSEHPQEK	AUG	36
		LKLEAELGNMQLLEDFK	non-AUG	52
		MTNQEAIQDLWQWR	AUG	44
		EESLVMQEEVWRKGN	non-AUG	24
		MRAGVVCVSQAQKDELILEGNDIELVSNAAAIQQATTVK	AUG	113
		RVGAELNLWLLK	non-AUG	42
		GNTFGYLLK	non-AUG	33
		KNSLLDLTLPSSDTR	non-AUG	23
		LQLETEIEALREELLFMK	non-AUG	94
		GYFDSGDYNAK	non-AUG	119
		DGLAPTWSLPCPLLPGPLPPDAALPGAVR	AUG	149
		KYTLPPGVDPQTKVSSSLSPGTLTVEAMPK	non-AUG	85

Table 3.1. Full list of 37 non-Uniprot SEPs detected.

Table 3.1. (Continued) Full list of 37 non-Uniprot SEPs detected. Locations where the SEPs are translated relative to the known RefSeq annotated transcripts are shown (5'UTR, 3'UTR, non-coding). If there is no annotated transcript corresponds to the detected peptide, then SEPs are generated from RNA-seq transcripts. RefSeq gene name, detected peptide(s), start codon, and SEP length are shown for each detected SEPs. Labeled in red are SEPs with clear evolutionary signature across 29 mammals by PhyloCSF analysis.

Next, we looked into SEPs sequence homology to known protein sequences. The majority of the SEPs have unique amino acid sequences that their function needs to be studied individually. One of the SEPs with detected peptide: GYFDSGDYNMAK which is a 119 amino acid long SEP based on RNA-seq with ATG start does not have corresponding RefSeq transcript. When the full SEP sequence is compared to non-redundant human proteins by pBLAST, it has >85% sequence homology to alpha-endosulfine isoforms (Figure 3.5A). This is an interesting case such that it indicates that there exists additional homolog of alpha-endosulfine with potentially related function. Another SEP with detected peptide: TAFDEAIAELDTLSEESKDSMLIMQLLR, which we termed “fusion SEP” (Figure 3.5B). The first half of this SEP has completely unique sequence while the latter part has >90% sequence homology to tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta. Lastly, there are several cases where the SEPs have high sequence homology to a part of long proteins (Figure 3.5C). For example, one of the SEPs with detected peptide: NMITETSQADCAVLIVAAGVGEFEAGISK, 123 amino acid long SEP with ATG start, has high sequence homology to a 462 amino acid long eukaryotic translation elongation factor 1 from residue 49 to 169. The observation for these distinct cases can be taken into account when studying particular SEPs function.

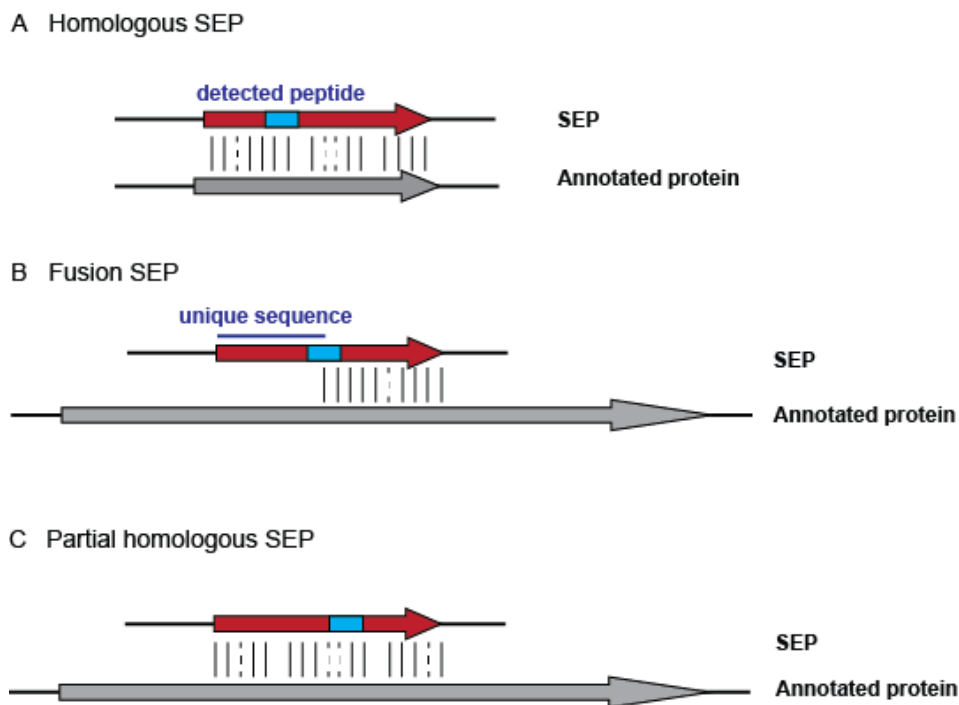


Figure 3.5. SEPs can be further sub-categorized based on their sequences. Aside from unique SEPs that have no sequence homology to known proteins, there are three sub-categories. (A) Homologous SEP: SEP has high sequence homology (>80%) to annotated small proteins. (B) Fusion SEP: a part of the SEP has unique sequence and the other half has high sequence homology (>90%) to known protein. (C) Partial homologous SEP: SEP has high sequence homology (>80%) to part of the known large protein. Solid lines represent matching amino acids, dotted lines represent synonymous changes, and gaps represent mismatches and deletions.

3.3. Conclusions

Enrichment of small peptidome is a key factor for detecting low abundant SEPs. We tested several parameters to improve upon our original workflow in order to determine the most efficient way to enrich small proteins. Different enrichment methods and extraction methods were applied and in conclusion, we determined that cellular proteins extracted with lysis buffer in combination with SPE enrichment led to the largest number of SEPs detected. SEPs are short and therefore only a limited number of tryptic peptides can be detected. Often, we rely on a

single peptide detection to assign a SEP. Improvements in instrumentation using high resolution HCD fragmentation method couple with increased AGC setting allowed higher confidence in peptide identification. SEPs are often expressed in low level and relative quantitation by spectral count is unreliable. Label-free analysis using Skyline (32) is a useful and reliable tool in quantitation of SEPs when studying SEP regulation and translational response to drug treatment, without additional labeling of samples. Finally, PhyloCSF (33) can be used to elucidate which SEPs are most likely biologically functional by assessing evolutionary signature of SEP sequences. Sequence homology of SEPs to known proteins can also aid to determine SEPs' cellular functions. Overall, the optimized SEP discovery platform can be applied widely to differentially profile SEPs in cell lines and tissues under different conditions and reveal underlying biology.

3.4. Materials and Methods

3.4.1. Cell Culture

K562 and A549 cells were maintained in RPMI and F-12K media, respectively. HeLa and HEK293 cells were cultured using DMEM. The media contained 10% fetal bovine serum (FBS). Cells were grown under an atmosphere of 5% CO₂ at 37°C until confluent. Before cells lysis and enrichment of SEPs, the media was removed from adherent cells by aspiration (A549, HeLa, HEK293) or non-adherent cells (K562) by centrifugation. HEPES-buffered saline (pH 7.5) was used to wash the cells to remove residual media and FBS.

3.4.2. Testing Various Methods for SEP Enrichment

We tested three conditions for SEP enrichment: (1) acid precipitation, (2) 30-kDa molecular weight cut off (MWCO) filter, and (3) reverse phase (C8) cartridge enrichment. Cellular proteomes from 4×10^7 cells were extracted by lysis with boiling water. After cooling the samples on ice, the cells were sonicated for 20 bursts at output level 2 with a 30% duty cycle (Branson Sonifier 250; Ultrasonic Converter). For the acid precipitation, addition of acetic acid (final concentration of 0.25% by volume) was followed by centrifugation at $14,000 \times g$ for 20 min at 4°C . This step precipitates larger proteins to reduce the complexity of the supernatant, which is then analyzed for SEPs. For the 30-kDa MWCO, the addition of acetic acid (final concentration of 0.25% by volume) was followed by centrifugation at $14,000 \times g$ for 20 min at 4°C . The supernatant is then passed through a 30-kDa MWCO filter and the flow through is analyzed for SEPs. Lastly, the reverse phase enrichment, the cellular extracts are were centrifuged at $25,000 \times g$ for 30 minutes and supernatants removed and filtered through 5 mM syringe filters followed by enrichment of SEPs using Bond Elute C8 silica cartridges (Agilent Technologies, Santa Clara, CA). Approximately 100 mg sorbent was used per 10 mg total lysate protein. Cartridges were wet with one column volume methanol, equilibrated with two-column volumes triethylammonium formate (TEAF) buffer, pH 3.0 and subsequently the sample applied. The cartridges were then washed with two column volumes TEAF and the SEP enriched fraction eluted by the addition of 75% acetonitrile/25% TEAF pH 3.0 and lyophilized using a Savant Speed-Vac concentrator (Thermo Scientific). BCA protein assay (Thermo Scientific) was used to measure protein concentration of each sample after extraction and enrichment.

3.4.3. Testing Different Methods for SEP Extraction

Four different methods were compared for extraction of SEPs from 4×10^7 total cells: (1) 50 mM HCl, 0.1% b-mercaptoethanol (b-ME), 0.05% Triton X-100 at room temperature (lysis buffer), (2) 1 N acetic acid/0.1 N HCl at room temperature, (3) boiling in water, or (4) boiling in lysis buffer. After extraction using these four methods, the extracts were centrifuged at 25,000 $\times g$ for 30 minutes, and supernatants filtered through 5 mM syringe filters. The flow through was then enriched for SEPs by binding and elution using Bond Elute C8 silica cartridges (Agilent Technologies, Santa Clara, CA). Approximately 100 mg sorbent was used per 10 mg total lysate protein. Cartridges were wet with one column volume methanol, equilibrated with two-column volumes triethylammonium formate (TEAF) buffer, pH 3.0 and subsequently the sample applied. The cartridges were then washed with two column volumes TEAF and the SEP enriched fraction eluted by the addition of 75% acetonitrile/25% TEAF pH 3.0 and lyophilized using a Savant Speed-Vac concentrator (Thermo Scientific). BCA protein assay (Thermo Scientific) was used to measure protein concentration of each sample after extraction and enrichment.

3.4.4. Digestion and Sample Preparation for LC-MS/MS

An aliquot of 100 μg of enriched samples were precipitated with chloroform/methanol extraction. Dried pellets were dissolved in 8 M urea/100 mM TEAB, pH 8.5. Proteins were reduced with 5 mM tris 2-carboxyethylphosphine hydrochloride (TCEP, Sigma-Aldrich) and alkylated with 10 mM iodoacetamide (Sigma-Aldrich). Proteins were digested overnight at 37 $^{\circ}C$ in 2 M urea/100 mM Tris, pH 8.5, with trypsin (Promega). Digestion was stopped with formic acid, 5 % final concentration.

3.4.5. Q-Exactive LC-MS/MS analysis.

Digests were analyzed by LC-MS using an Easy-nLC1000 (Proxeon) and a Q Exactive mass spectrometer (Thermo Scientific). An EASY-Spray column (Thermo Scientific) 25 cm by 75µm packed with PepMap C18 2µm particles was used. Electrospray was performed directly from the tip of the analytical column. Buffer A and B were 0.1 % formic acid in water and acetonitrile, respectively, and the solvent flow rate was 300 nl/min. Each sample was run in triplicate. The digested samples were loaded onto the column using an autosampler, and the samples were desalted online using a trapping column. Peptide separation was performed with 6-hour reverse phase gradient. The gradient increases from 5-22% B over 280 minutes, 22-32% B over 60 minutes, 32-90% B over 10 minutes, followed by a hold at 90% B for 10 minutes. The column was re-equilibrated with buffer A prior to injection.

The Q Exactive was operated in a data dependent mode. Full MS1 scans were collected with mass range of 400 to 1800 m/z at 70k resolution. The 10 most abundant ions per scan were selected for MS/MS with an isolation window of 2 m/z and HCD energy of 25 and resolution of 17.5k. Maximum fill times were 60 and 120 ms for MS and MS/MS scans, respectively. An underfill ratio of 0.1% was utilized for peak selection, dynamic exclusion was enabled for 15s and unassigned and singly charge ions were excluded. Data was collected with default values for AGC target of 1e6 and 5e5 and maximum injection times of 60 and 120ms for MS and MS/MS scans respectively. Data was also collected with sensitive settings for comparison. AGC of MS and MS/MS scans were increased to 5e6 and 5e6 respectively and maximum fill times were increased to 120ms and 500ms. All other parameters remained unchanged.

3.4.6. Orbitrap Fusion Tribrid LC-MS/MS Analysis

C8 enriched samples were analyzed on a Orbitrap Fusion tribrid mass spectrometer (Thermo). The digest was injected directly onto a 50cm, 75um ID column packed with BEH 1.7um C18 resin (Waters). Samples were separated at a flow rate of 200 nl/min on a nLC 1000 (Thermo). Buffers A and B were 0.1% formic acid in water and acetonitrile, respectively. A gradient of 1-22%B over 160 min, an increase to 32%B over 60 min, an increase to 90%B over another 10 min and held at 90%B for a final 10 min of washing was used. Column was re-equilibrated with 20 µl of buffer A prior to the injection of sample. Peptides were eluted directly from the tip of the column and nanosprayed directly into the mass spectrometer by application of 2.5kV voltage at the back of the column. The Orbitrap Fusion was operated in a data dependent mode. Full MS scans were collected in the Orbitrap at 120K resolution with a mass range of 400 to 1500 m/z and an AGC target of 4e5 and maximum fill time of 50ms. The cycle time was set to 3sec, and within this 3sec the most abundant ions per scan were selected for fragmentation by either CID in the ion trap with an AGC target of 1e4 and maximum fill time of 35ms or ions were selected for HCD and detection in the orbitrap with an AGC target of 5e5 and max fill time of 250ms. Collision energy was set to 35 for both CID and HCD and a minimum intensity of 5000 was required for selection. Quadrupole isolation at 1.6 m/z was used, monoisotopic precursor selection was enabled and dynamic exclusion was used with exclusion duration of 10 sec.

3.4.7. Data analysis to Identify Annotated and Non-annotated SEPs

Tandem mass spectra were extracted from raw files using RawExtract 1.9.9.2 and searched with ProLuCID using Integrated Proteomics Pipeline – IP2 (Integrated Proteomics Applications) (34). We used two databases in these searches, a custom database created from the in silico 3-frame translation of RNA-Seq data from K562 cells (custom database), and the UNIPROT Human database. The transcriptome data is deposited on GEO (GSE34740). The search space included all fully-tryptic and half-tryptic peptide candidates. Carbamidomethylation on cysteine was considered as a static modification.

To identify annotated and non-annotated SEPs, data files from technical duplicates were combined and searched by ProLuCID. Data was searched with 50-ppm precursor ion tolerance then filtered to 10-ppm, and 50-ppm fragment ion tolerance with maximum of two internal missed cleavages using either the custom database and UNIPROT Human database. Identified spectra were filtered and grouped into proteins using DTASelect (35, 36). Proteins and SEPs required at least one peptide to be identified with a setting of less than 1% FDR. Unique peptides identified by searching the UNIPROT database that belonged to smORFs of less than a 150 codons were kept and are referred to as ‘Annotated SEPs’.

To identify non-annotated SEPs, data files from technical duplicates were combined and searched by ProLuCID. Data was searched with 50-ppm precursor ion tolerance then filtered to 10-ppm, and 50-ppm fragment ion tolerance with maximum of two internal missed cleavages using only the custom database. The results from the custom database search were then filtered against the UNIPROT human database using a string-searching algorithm to remove any annotated peptides. The remaining peptides are identified from the custom database but are not found in the UNIPROT human database.

The next step is to use these remaining peptides to determine whether there are derived from smORFs of less than 150 codons. To do this, each remaining peptide was searched against Human Reference (RefSeq) RNA sequences using tBLASTn. After identifying an RNA, the downstream in frame stop codon of the smORF is established. The upstream start codon is then determined by looking for an in-frame upstream ATG, or a near cognate start codon (i.e. ACG, AAG, CUG etc.) that is embedded within a Kozak sequence (37). If an ATG or near cognate start codon cannot be found, an upstream in-frame stop codon is identified. If the length between the upstream and downstream stop codons is less than 150 codons the smORF/SEP is retained on our final list. If the peptides did not match to any RNA sequences with the RefSeq RNA database, it means that they were derived from RNAs that were present in the RNA-Seq data that are not annotated in RefSeq (i.e. non-annotated RNAs). For these peptides, we repeat the identification of the start and stop codons using RNAs from the RNA-Seq database as described above.

Also, we visually inspect the MS2 spectra for all of the peptides that are assigned to smORFs to ensure that we have optimal coverage of the sequence and that any key amino acid residues that uniquely distinguish the peptide are present in the data. Then, the entire SEP sequence is searched using the BLAST (pBLAST) algorithm against the non-redundant human protein database to determine whether there is any sequence homology to known proteins.

3.4.8. Arsenite Treatment Experiments

HEK293 cells were grown to ~70% confluence and then treated with 10 μ M sodium arsenite for 24 hours. Cellular proteins were extracted using the lysis buffer followed by centrifugation 20,000 x g for 20 minutes at 4 °C to remove any insoluble particulates. The

concentrations were determined using a Bradford assay and 100 mg was taken forward for digestion and sample preparation (see above) and LC-MS/MS using the Q Exactive. After collection of the data, LC-MS peaks corresponding to two SEPs and two proteins were identified and quantified using Skyline. The retention times for these the peptides corresponding to the two proteins and two peptides were determined using ProLuCID. The corresponding retention time were then used to identify and quantify the proteins using Skyline. The AUC (area under the curve) for the peptide ions was used to determine the relative quantity of each peptides between control and arsenite treated samples. The extraction of the isotopic peaks for each peptide and comparison to the theoretical isotopic distribution at resolution of 60k validated the selected peptide ion that was used for quantitation.

3.4.9. Raising SLC35A4-SEP Antibody

Antisera against SLC35A4 was raised in rabbits against a synthetic peptide fragment encoding Cys³⁴SLC35A4(2-34) coupled to maleimide activated keyhole limpet hemocyanin (ThermoFisher, Waltham MA). The peptide, <Hnt>ADDKDSLPLKDLAFLKNQLES LQRRVEDEVNC<OH>, was synthesized and C18 HPLC purified by RS Synthesis (Louisville, KY); purity was 99.0%. Immunogen was prepared by emulsification of Freund's complete adjuvant-modified *Mycobacterium butyricum* (EMD Millipore, Billerica MA) with an equal volume of phosphate buffered saline (PBS) containing 1.0 mg conjugate/ml for initial injections. For booster injections, incomplete Freund's adjuvant was mixed with an equal of PBS containing 0.5 mg conjugate/ml. For each immunization, an animal received a total of 1 ml emulsion in 20 intradermal sites in the lumbar region. Three individual rabbits were injected every three weeks and were bled one week following booster

injections. Bleeds were screened for titer and specificity; antiserum PBL #7383, 6/25/15 bleed, was used for these studies. All animal procedures were approved by the Institutional Animal Care and Use Committee of the Salk Institute and were conducted in accordance with the National Institutes of Health guidelines.

3.4.10. Western Blot

Control and sodium arsenite treated HEK293 cells were extracted by lysis buffer. Protein concentration was measured using Bradford assay (BioRad). 30 µg of total protein from each sample was loaded on a 4-12 % BisTris gel, 10-well (Bolt, Life Technologies) and run in MES running buffer at 200V for 20 min. Proteins were transferred to PVDF membrane and then blocked at room temperature for 1 hour using LiCor Blocking Buffer. The membrane was then blotted with primary antibody; rabbit anti-beta actin (LiCor) 1:1000 for 1 hour at room temperature; rabbit anti-HO-1 (Cell Signaling) overnight at 4°C; or rabbit anti-SLC35A4 SEP at 1:5000 dilution overnight at 4°C. Wash membrane three time with TBS-T, then blot with secondary antibody: goat anti-rabbit IRDye 800CW (LiCor) at 1:10000 dilution, rock 1 hour at room temperature. Wash membrane three times with TBS-T then scan the membrane using LiCor Odyssey CLx at IR700 and IR800. Buit-in tool in Odyssey CLx was used to quantify the intensity of the bands of interest.

3.4.11. PhyloCSF analysis of SEPs with Evolutionary Signature

PhyloCSF scores (33) were computed for the sORFs encoding each of the detected SEPs using multiple sequence alignments of 29 eutherian mammals. We excluded SEPs with score significantly below 0 or for which the average branch length of the species present in the

alignment was less than one tenth of the branch length of the full 29-mammal tree. We excluded any of the SEPs that overlapped a pseudogene annotated in GENCODE version 23 (38) because non-coding regions of pseudogenes often get a positive PhyloCSF score. We then manually examined the alignments and predicted splice sites using CodAlignView. We excluded any SEPs for which the evolutionary coding signature stopped near a predicted splice site rather than extending all the way to the start or stop codon, since the coding signature of such SEPs could be due to overlap with a different splice variant. We also excluded SEPs for which the evolutionary coding signature extended 5' of the start codon, since the coding signature of such SEPs could be due to a longer ORF. This left 6 SEPs with strong evolutionary signature of being short, protein-coding ORFs with conserved function.

3.5. References

1. Y. Hashimoto *et al.*, A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and A β . *Proc Natl Acad Sci U S A* **98**, 6336-6341 (2001).
2. M. Bliss, *The Discovery of Insulin*. (University of Chicago Press, 2013).
3. M. Bliss, R. Purkis, *The discovery of insulin*. (University of Chicago Press Chicago, 1982).
4. L. De Lecea *et al.*, The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. *Proceedings of the National Academy of Sciences* **95**, 322-327 (1998).
5. T. Sakurai *et al.*, Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92**, 573-585 (1998).

6. W. Vale, J. Spiess, C. Rivier, J. Rivier, Characterization of a 41-residue ovine hypothalamic peptide that stimulates secretion of corticotropin and beta-endorphin. *Science* **213**, 1394-1397 (1981).
7. M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**, e106 (2007).
8. T. Kondo *et al.*, Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology* **9**, 660-665 (2007).
9. M. R. Hemm *et al.*, Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. *Journal of bacteriology* **192**, 46-58 (2010).
10. M. R. Hemm, B. J. Paul, T. D. Schneider, G. Storz, K. E. Rudd, Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular microbiology* **70**, 1487-1501 (2008).
11. C. S. Wadler, C. K. Vanderpool, A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proceedings of the National Academy of Sciences* **104**, 20454-20459 (2007).
12. K. Hanada *et al.*, Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proceedings of the National Academy of Sciences* **110**, 2395-2400 (2013).
13. J. L. Aspden *et al.*, Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife* **3**, e03528 (2014).
14. M. C. Frith *et al.*, The abundance of short proteins in the mammalian proteome. *PLoS Genet* **2**, e52 (2006).
15. J. P. Kastenmayer *et al.*, Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* **16**, 365-373 (2006).
16. J. Ma *et al.*, Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *Journal of proteome research* **13**, 1757-1765 (2014).

17. M. Oyama *et al.*, Diversity of translation start sites may define increased complexity of the human short ORFeome. *Molecular & Cellular Proteomics* **6**, 1000-1006 (2007).
18. S. A. Slavoff *et al.*, Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature chemical biology* **9**, 59-64 (2013).
19. B. Vanderperre *et al.*, Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* **8**, e70698 (2013).
20. E. Ladoukakis, V. Pereira, E. G. Magny, A. Eyre-Walker, J. P. Couso, Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* **12**, R118 (2011).
21. C. Lee *et al.*, The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab* **21**, 443-454 (2015).
22. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J Biol Chem* **289**, 10950-10957 (2014).
23. B. Vanderperre, J. F. Lucier, X. Roucou, HALtORF: a database of predicted out-of-frame alternative open reading frames in human. *Database (Oxford)* **2012**, bas025 (2012).
24. A. A. Bazzini *et al.*, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981-993 (2014).
25. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
26. R. M. Branca *et al.*, HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods* **11**, 59-62 (2014).
27. N. Castellana, V. Bafna, Proteogenomics to discover the full coding content of genomes: a computational perspective. *Journal of proteomics* **73**, 2124-2135 (2010).

28. H. Liu, R. G. Sadygov, J. R. Yates, 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193-4201 (2004).
29. J. K. Diedrich, A. F. Pinto, J. R. Yates, 3rd, Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot. *J Am Soc Mass Spectrom* **24**, 1690-1699 (2013).
30. A. T. Lau, Q. Y. He, J. F. Chiu, A proteome analysis of the arsenite response in cultured lung cells: evidence for in vitro oxidative stress-induced apoptosis. *Biochem J* **382**, 641-650 (2004).
31. D. E. Andreev *et al.*, Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4**, e03971 (2015).
32. B. MacLean *et al.*, Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968 (2010).
33. M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-282 (2011).
34. T. Xu *et al.*, ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J Proteomics* **129**, 16-24 (2015).
35. D. Cociorva, L. T. D, J. R. Yates, Validation of tandem mass spectrometry database search results using DTASelect. *Curr Protoc Bioinformatics* **Chapter 13**, Unit 13 14 (2007).
36. D. L. Tabb, W. H. McDonald, J. R. Yates, 3rd, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **1**, 21-26 (2002).
37. M. Kozak, Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283-292 (1986).
38. J. Harrow *et al.*, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).

Chapter 4

The Polypeptide NoBody Regulates Cellular P-body Number

This chapter was adapted from: Slavoff, SA.; Ma, J.*; Winkler, L.*; Chu, Q.; Liberzon, A.; Narayan, R.; Budnik, BA.; Subramanian, A.; and Saghatelian, A. The Polypeptide NoBody Regulates Cellular P-body Number. *Nature Chemical Biology* (In Revision)

*Authors contributed equally

4.1 Introduction

Genomics and proteomics experiments have discovered numerous translated short open reading frames (sORFs) in human cells and tissues, but the functional capacity of these novel human sORF-encoded polypeptides (SEPs) is unknown. Here, we characterize a ~7-kilodalton SEP named non-annotated P-body dissociating polypeptide (NoBody) in human cells. The first evidence that NoBody is functional came from functional proteomic studies that identified its specific association with the mRNA decapping complex, a group of proteins that remove the 5'-cap from mRNA to promote mRNA degradation. A short binding site on NoBody interacts with the decapping complex by binding to enhancer of decapping protein 4 (EDC4), a recognized mRNA decapping protein. In cells, the mRNA decapping complex is found in solution and also in a ribonucleoprotein granule called an mRNA processing body, or P-body. Imaging experiments reveal that NoBody inversely regulates cellular P-body levels without affecting levels of P-body proteins. Imaging studies at low NoBody expression levels validated direct binding of NoBody to the P-body, which support a model of NoBody binding followed subsequent dissociation of P-body granules. Transcriptomic analysis was also consistent with a role for NoBody in mRNA processing. The discovery and characterization of NoBody reveals a novel mechanism for regulating cellular P-bodies and highlights an essential function for a human SEP.

Translated short open reading frames (sORFs) containing less than 100 codons are becoming accepted as a significant fraction of genomes across evolution (1, 2), and many of these sORFs encode functional polypeptides. In *Drosophila*, for example, sORF-encoded polypeptides (SEPs) serve critical biological roles, including regulating morphogenesis (*tal/pri*) (3-6) and heart rhythm (*sarcolamban*) (7). Over one thousand SEPs have now been reported in

human cells and tissues by proteomics (8-12), but it is unknown whether these genes are functional. To understand the importance of human SEPs, we must identify and characterize as many of these molecules as possible. Such studies will eventually lead to a broader understanding of SEP function, including unique roles for these polypeptides, and in doing so will increase scientific knowledge regarding the functional proteome. Here, we characterize a conserved non-annotated human SEP we have named NoBody.

4.2. Results

4.2.1. Discovery and conservation analysis of NoBody

We detected an unknown SEP translated from the LOC550643 RNA (Figure 4.1) in K562 cells; the first experimental evidence that the LOC550643 gene is protein-coding (Figure 4.2A). The NCBI database annotates LOC550643 RNA as non-coding, but our new data

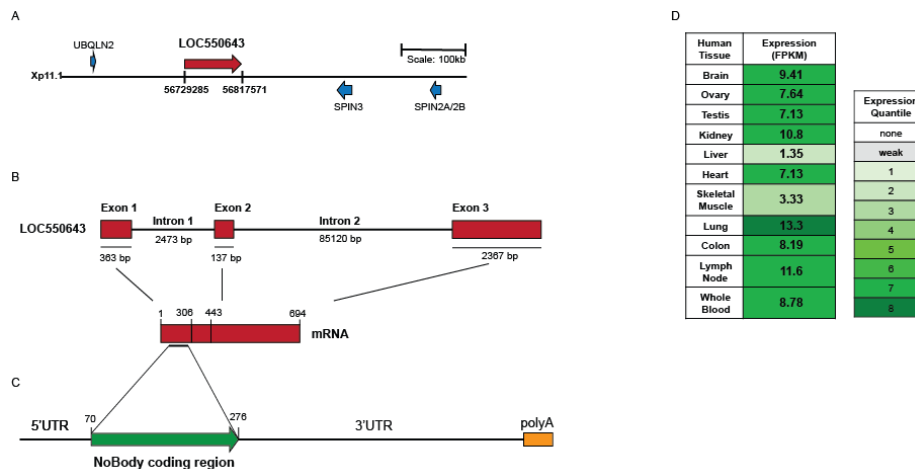


Figure 4.1. LOC550643/LINC01420/NoBody genomic locus and expression. (A) LOC550643/LINC01420/NoBody (red arrow) is coded on the positive strand of the X chromosome at Xp11.1. Genomic coordinates are provided below, and nearby genes are indicated by gene name and blue arrows. Diagram to scale. (B) The NoBody gene consists of 3 exons and 2 long intervening introns. (C) A single 722-nucleotide RNA

Figure 4.1. (Continued) transcript spanning the 3 spliced LOC550643 exons is currently annotated in RefSeq. The transcript is shown as a horizontal black line, NoBody coding sequence is shown as a green arrow, splice junctions are shown as vertical black lines, and the polyA tail is shown as a small orange box. Positions of each feature are provided in the diagram. (D) Expression data for LOC550643 for various human tissues derived from publicly available RNA deep sequencing data, expressed in fragments per kilobase of transcript per million mapped reads. Color scale represents expression quantile range for all human genes sequenced in this dataset; 8th quantile: highest, weak: too low for reliable quantitation (2-8 reads above intergenic background), none: no expression detected.

indicates otherwise. We refer to this 7-kilodalton polypeptide as NoBody (non-annotated P-body dissociating polypeptide) based on the cellular function of this peptide (*vide infra*). NoBody is also detected in HEK293T and MDA-MB-231 cells (Figure 4.3), indicating expression in cell lines from different tissues of origin. We transfected human cells with an expression construct comprising the full-length LOC550643 cDNA with a 3' epitope tag. These cells expressed NoBody, validating that LOC550643 RNA contains a protein-coding sORF (Figure 4.2B). A translated nucleotide BLAST (tBLASTn) search revealed that NoBody is a conserved mammalian gene (Figure 4.2C). A BLAST search of NoBody against the *Drosophila melanogaster*(13) and zebrafish (14-16) genomes with relaxed parameters did not identify any homologous proteins, indicating that NoBody arose in the mammalian lineage. The expression of NoBody in several cell lines and its conservation are indicative of a functional gene. NoBody shows no homology to any functional proteins, and structure analysis (17, 18) predicted no secondary structure. Therefore, we began our characterization of NoBody using an unbiased functional proteomics approach to identify NoBody interacting proteins.

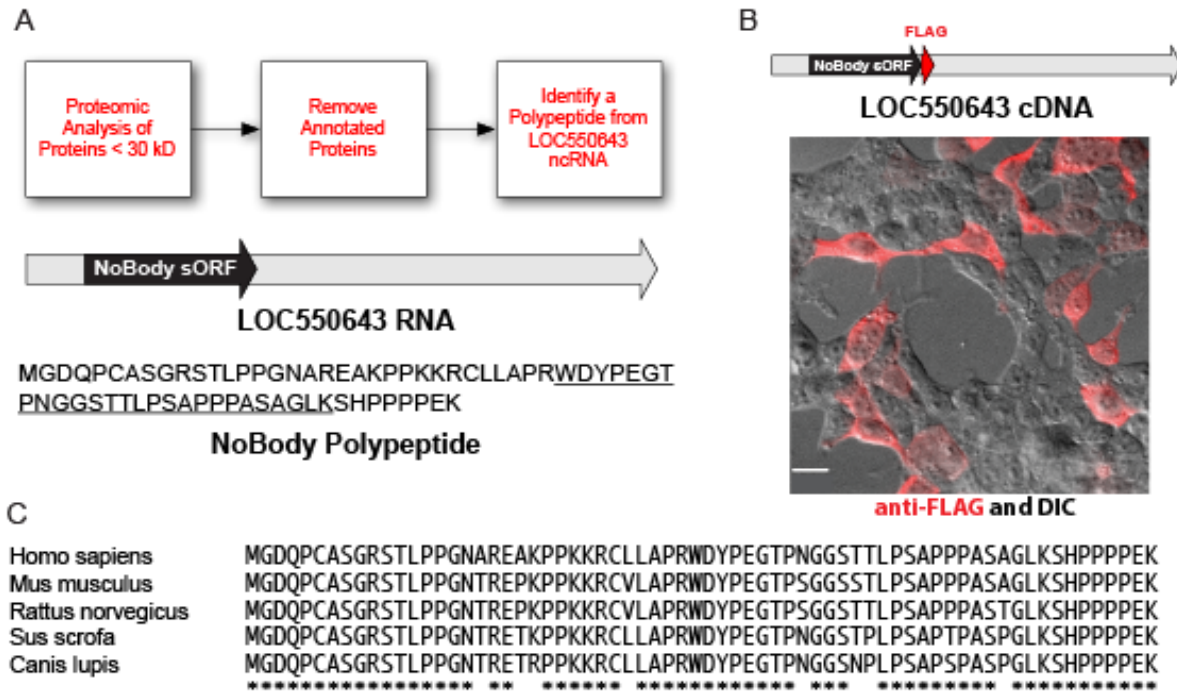
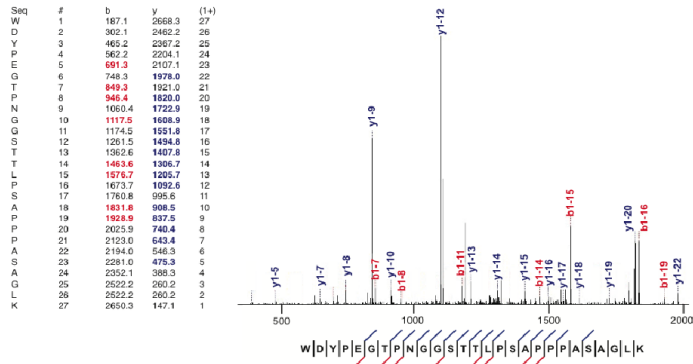


Figure 4.2. The *LOC550643* gene encodes the NoBody peptide in a short open reading frame (sORF). (A) K562 and HEK293T cellular peptides were enriched and subjected to multidimensional LC-MS proteomics. Peptide mass spectra were searched against a custom protein database obtained from the 3-frame translation of RNA-Seq data from these cell lines. Annotated peptides were removed by BLAST search to afford a list of non-annotated peptides. This workflow led to the discovery of a tryptic peptide (underlined sequence) derived from a polypeptide translated from a sORF (black) in the *LOC550643* RNA transcript (gray). The polypeptide is hereafter referred to as NoBody. (B) Transfection of an expression construct corresponding to the annotated full-length *LOC550643* cDNA sequence (gray), with an epitope tag (red) at the C-terminus of the putative short ORF (black) into HEK293T cells resulted in expression of NoBody (red anti-FLAG immunofluorescence image superimposed on differential interference contrast (DIC) image). Scale bar, 20 μ m. (C) ClustalW2 alignment of full-length NoBody polypeptide sequence from a variety of mammals. Amino acid identity is indicated by asterisks.

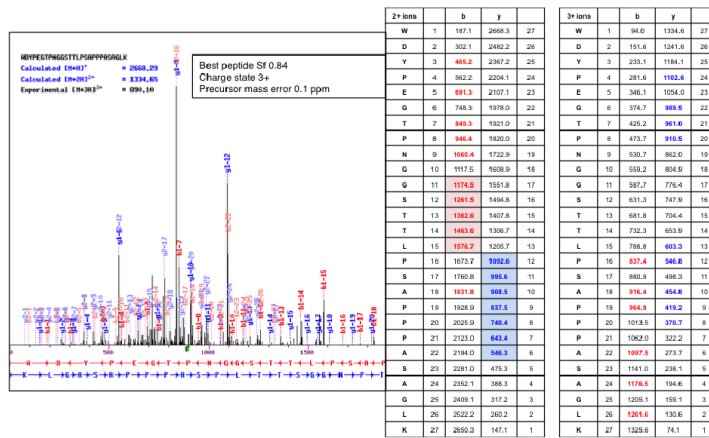
4.2.2. NoBody interacts with proteins involved in 5'-3' mRNA decay

We used functional proteomics to identify NoBody-protein interactions. Proteomic analysis of immunoprecipitates of NoBody revealed the enrichment of several proteins that comprise the cellular 5'-3' mRNA decay pathway (19, 20). These proteins include enhancer of

a



b



c

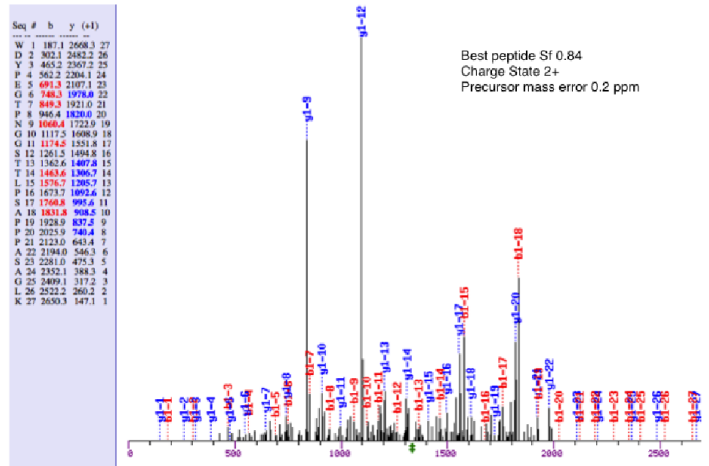


Figure 4.3. Peptidomic identification of NoBody in K562, HEK293T, and MDA-MB-231 cells. Tandem mass spectrometry (MS/MS) spectrum and detection parameters and scores are presented for the tryptic fragment of NoBody detected in (a) K562, (b) HEK293T, and (c) MDA-MB-231 cells. Fragment ion masses identified in the NoBody tryptic peptide spectrums are shown (red for b ions, blue for y ions). Black, not detected.

decapping 3 and 4 (EDC3 and EDC4), two orthologs of decapping protein 1 (Dcp1A and Dcp1B), decapping protein 2 (Dcp2), and the Lsm complex (Figure 4.4A). A Western blot with antibodies against these proteins demonstrated their enrichment by NoBody and validated the proteomics data (Figure 4.4B). These proteins are known to form a complex and function in the 5'-3' mRNA degradation pathway. The 5'-3' mRNA decay pathway is a vital process that degrades cellular RNAs and is initiated by the decapping enzyme Dcp2, which removes the 5'-7-methylguanosine cap, exposing the decapped RNA to exonuclease digestion (20).

We hypothesized that NoBody might be enriching the complex through interactions with a single protein. To identify proteins that NoBody could interact with we performed reciprocal immunoprecipitation experiments. Immunoprecipitation of EDC4 or Dcp1A indicated NoBody is primarily interacting with EDC4 (Figure 4.4C,D). A previous report suggested that EDC4 may have a “small subunit,” which may have been NoBody, but this protein was never identified (19).

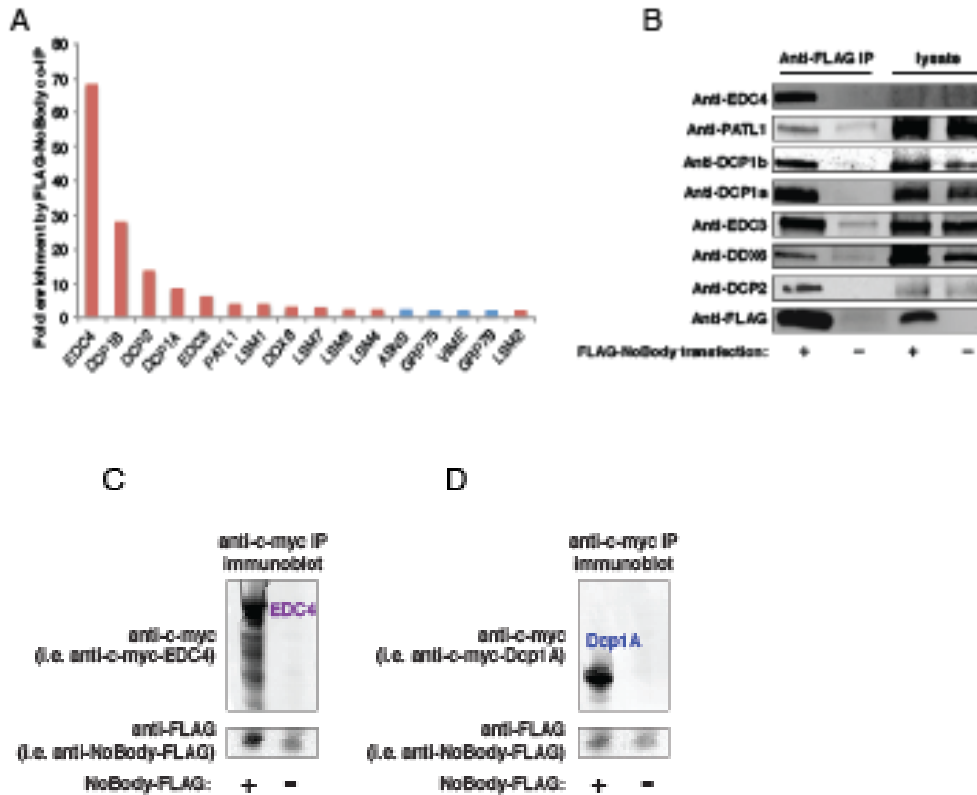


Figure 4.4. NoBody immunoprecipitation enriches a complex of proteins involved in 5'-3' mRNA decay. (A) FLAG-NoBody was immunoprecipitated from lysates of transiently transfected HEK293T cells using anti-FLAG agarose; empty vector-transfected HEK293T lysates served as a control. Protein samples were separated by SDS-PAGE, then subjected to in-gel digest and liquid chromatography-tandem mass spectrometry. Proteins were identified by Sequest search against the UniProt human protein database, and enrichment determined by the ratio of the average normalized MS1 intensity of each identified protein in the immunoprecipitation sample relative to the control. Proteins identified by >2 tryptic peptides and exhibiting an enrichment ratio >2 were retained for consideration. Red bars correspond to proteins with a *bona fide* role in 5'-3' mRNA decay. (B) Western blotting confirmation of NoBody-interaction partners identified by proteomics. The same FLAG-NoBody immunoprecipitation experiment was performed, but samples were blotted and probed with protein-specific antibodies to confirm enrichment of 5'-3' mRNA decay factors. (C,D) Reciprocal anti-c-myc immunoprecipitation of c-myc-EDC4 and c-myc-Dcp1A from HEK293T cells co-expressing FLAG-NoBody, with anti-c-myc immunoprecipitation from cells expressing NoBody-FLAG alone as a negative control, followed by anti-FLAG immunoblotting, demonstrates that FLAG-NoBody is enriched by EDC4.

4.2.3. Sequence dependence of the NoBody-EDC4 interaction

We obtained additional biochemical data to support NoBody binding to EDC4. We defined interaction sites of EDC4 and NoBody using deletion mutants of EDC4 and NoBody (Figure 4.5). Preliminary data (not shown) demonstrated that some NoBody-FLAG deletion variants did not express well, which would complicate downstream analysis. As a result, we prepared selected fusion proteins of NoBody deletion variants and enhanced green fluorescent protein (NoBody-EGFP fusions). NoBody-EGFP deletion mutants that lacked amino acids 22-41 (Δ 22-31 and Δ 32-41) did not immunoprecipitate EDC4 (Figure 4.5A), indicating that NoBody22-41 is necessary for interactions with EDC4. Immunoprecipitation of NoBody22-41-EGFP fusion proteins enriched EDC4 (Figure 4.5C,D), while WT-EGFP alone did not (Figure 4.5D), demonstrating that NoBody22-41 is sufficient for binding to EDC4.

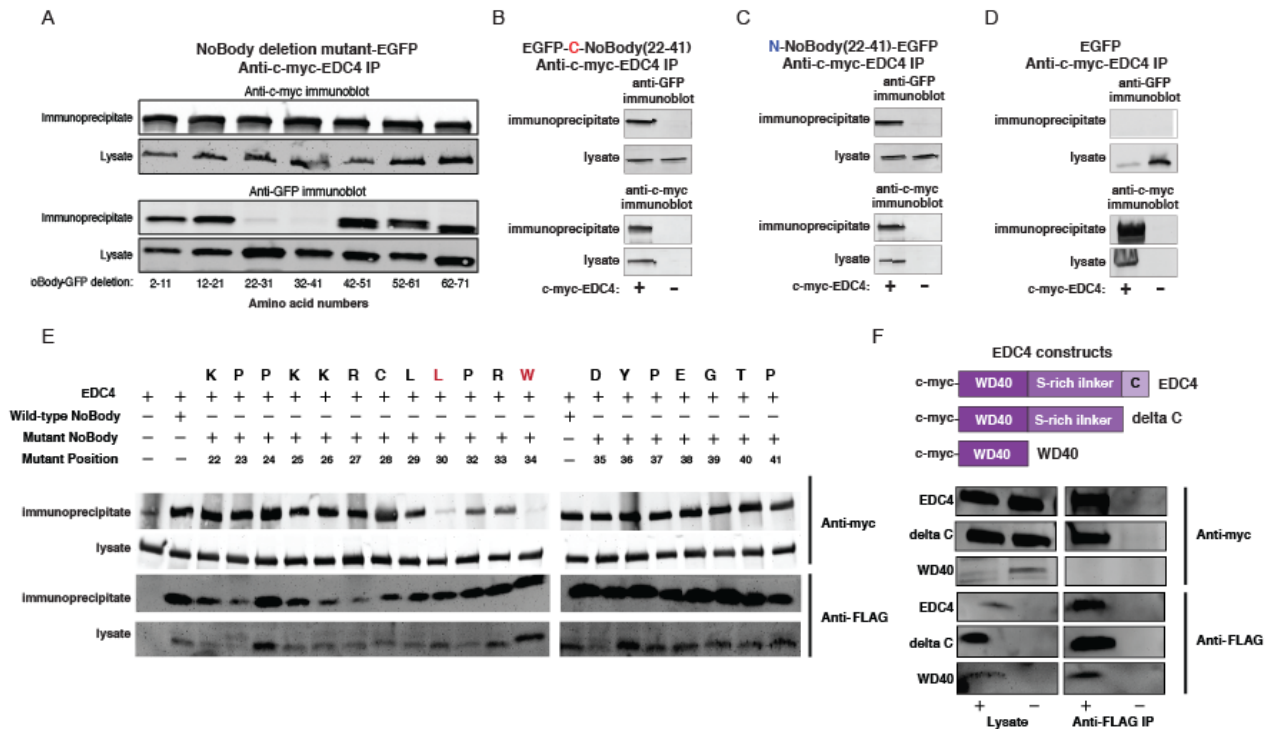


Figure 4.5. Sequence and domain dependence of the NoBody-EDC4 interaction. (A) In order to determine the region of the NoBody peptide responsible for interacting with EDC4, a series of 10-amino-acid deletions spanning the length of NoBody was prepared. These deletion constructs were fused to EGFP to improve stability when expressed in mammalian cells and to the FLAG epitope tag to facilitate immunoprecipitation. The constructs were co-transfected with c-myc-EDC4 in HEK293T cells (or c-myc-EDC4 was transfected alone as a negative control for background), then lysates were subjected to anti-FLAG immunoprecipitation. Immunoprecipitates were blotted to assess association of c-myc-EDC4, identifying a 20-amino-acid region from residue 22 to 41, that was necessary for the interaction. (B, C) The 20-amino-acid fragment of NoBody (NoBody(22-41)) was fused to the N- and C-termini of EGFP to assay its sufficiency for interaction with EDC4. The NoBody(22-41)-EGFP fusions were co-transfected into HEK293T cells with c-myc-EDC4, then lysates subjected to anti-c-myc immunoprecipitation followed by anti-GFP Western blotting to assess the interaction. A negative control (D) with non-fused EGFP demonstrates that EGFP does not interact non-specifically with EDC4. (E) Alanine scanning mutagenesis of full-length FLAG-tagged NoBody peptide between NoBody amino acid residues 22 to 41 was performed to identify residues essential for the interaction with EDC4. These constructs were co-expressed with c-myc-EDC4 in HEK293T cells and subjected to anti-FLAG immunoprecipitation, followed by anti-c-myc Western blotting to assess EDC4 interaction. Several alanine point mutations that severely inhibit the interaction were identified, especially L30A and W34A. (F) EDC4 was serially truncated to delete its C-terminal domain (deltaC), then to delete both the C-terminal domain and linker domain (WD40). These constructs were co-

Figure 4.5. (Continued) expressed with FLAG-NoBody, and anti-FLAG immunoprecipitation carried out to assess EDC4 construct enrichment, as compared to a negative control in the absence of NoBody. Whole cell lysates were probed to demonstrate equal loading. NoBody enriches full-length EDC4 and the –C-term construct, suggesting that NoBody binds to the S-rich linker domain of EDC4.

To further support a direct interaction between NoBody and EDC4, we synthesized a modified peptide comprising amino acids 22-41 of NoBody, with an N-terminal rhodamine and a C-terminal benzophenone (Rh-NoBody22-41-BPA). The benzophenone is a light-activated crosslinking group that forms a covalent link between adjacent proteins upon irradiation, and the rhodamine permits fluorescence detection of the covalent conjugate (21). In this case, if Rh-NoBody22-41-BPA binds to EDC4 then irradiation should result in a covalent bond between Rh-NoBody22-41-BPA and EDC4, showing a direct interaction. The addition of Rh-NoBody22-41-BPA to HEK293 lysates overexpressing EDC4 resulted in a light-dependent Rh-NoBody22-41-BPA labeling of EDC4 (Figure 4.6). The addition of excess biotin-NoBody22-41-BPA blocked this labeling, demonstrating that the interaction between NoBody and EDC4 is sequence-specific (Figure 4.6).

Fluorescent cross-linking peptide:
Rhodamine-KPPKKRCLLPRDYPEGTP(NoBody22-41)-Benzophenone

Non-fluorescent cross-linking peptide competitor:
Biotin-KPPKKRCLLPRDYPEGTP(NoBody22-41)-Benzophenone

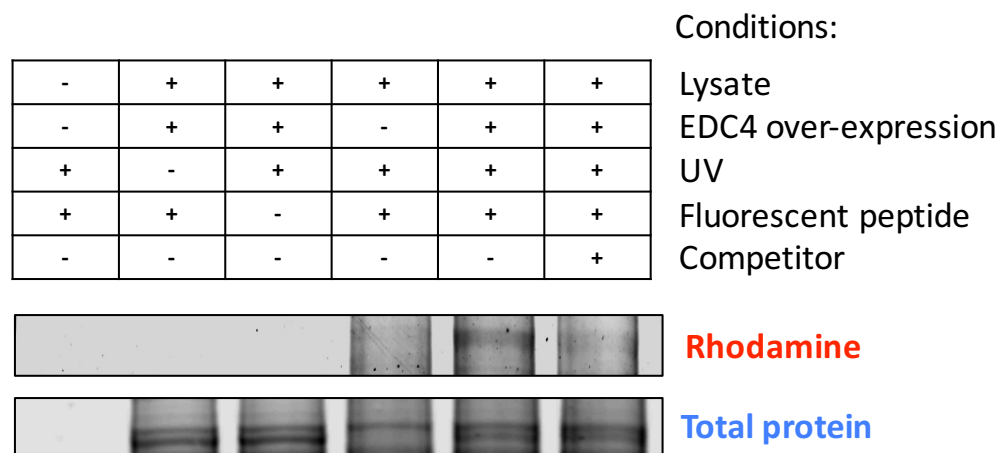


Figure 4.6. Photo-cross-linking evidence for direct physical interaction of NoBody with EDC4. A 20-amino-acid fragment of NoBody peptide required for interaction with EDC4 was prepared with a rhodamine fluorophore and benzophenone photo-cross-linker appended as a fluorescent probe of direct physical interaction between NoBody and EDC4. This peptide was introduced into lysates of HEK293T cells (either wild-type or over-expressing myc-EDC4), then subjected to UV-mediated cross-linking. Samples were then separated by SDS-PAGE and subjected to fluorescence imaging. A fluorescent band migrating at the EDC4 molecular weight (~150 kD) indicates covalent attachment and reports on direct physical interaction. Single-omission controls, as well as a competition control in the presence of non-fluorescent biotinylated benzophenone-peptide, demonstrate specificity of the cross-linking.

To identify critical residues involved in the NoBody-EDC4 interaction we performed an alanine scan between amino acids 22-41 of full-length FLAG-NoBody. Immunoprecipitation of FLAG-NoBody alanine mutants revealed that two amino acids (Leu30 and Trp34) have the strongest effect on NoBody-EDC4 binding (Figure 4.5E). In aggregate, the data establish a direct interaction between NoBody and EDC4 that is highly sequence dependent, requiring amino acids Leu30 and Trp34 for binding.

We also determined the domain on EDC4 required for NoBody binding through domain deletions of EDC4. EDC4 is a three-domain protein, with an N-terminal WD40 domain, an S-rich linker-containing domain, and a C-terminal domain that is responsible for P-body localization (22). NoBody associates with full-length EDC4 and a truncation mutant lacking the C-terminal domain but fails to associate with the EDC4 WD40 domain (Figure 4.5F). Thus, NoBody most likely binds to the S-rich linker domain of EDC4.

4.2.4. NoBody inversely regulates P-body numbers

The proteins involved in 5'-3' mRNA decay localize to ribonucleic acid protein (RNP) granules called mRNA processing bodies, or P-bodies, in all eukaryotic cells. P-bodies have been alternately hypothesized to be sites of RNA degradation by the 5'-3' mRNA decay pathway or sites for storage of translationally repressed RNAs (23). The P-body is a dynamic structure. Fluorescence recovery after photobleaching studies with P-bodies (24) indicate that some P-body proteins readily exchange between the RNA granules and the cytosol.

While the exact role of P-bodies in mRNA degradation remains unclear, it has been established that modulating the expression levels of P-body-associated proteins affects the numbers and morphologies of P-bodies. For example, EDC4 and Dcp1 promote P-body formation when overexpressed and eliminate P-bodies when silenced, and are, therefore, positive regulators of P-body formation(23). In contrast, Dcp2 and Xrn1 silencing increase P-body size and numbers (19, 23, 25-27). Because NoBody interacts with EDC4, we hypothesized that NoBody might also influence P-body numbers.

Endogenous EDC4 and Dcp1A were used as markers to image P-bodies in cells (Figure 4.7). We were unable to detect endogenous Dcp2. P-body proteins are known to mislocalize to

the cytoplasm, instead of P-bodies, if their expression levels are too high (19). To avoid mislocalization we reduced protein expression levels by using retroviral transduction instead of transient transfection. We tested retroviral transduction of GFP-Dcp2, FLAG-APEX-Dcp2, and FLAG-APEX-Rck1 (Rck1 is a kinase associated with the P-body). P-bodies in retrovirally transduced cells appeared normal, and GFP-Dcp2, FLAG-APEX-Dcp2, and FLAG-APEX-Rck1 all localized correctly to endogenous P-body foci.

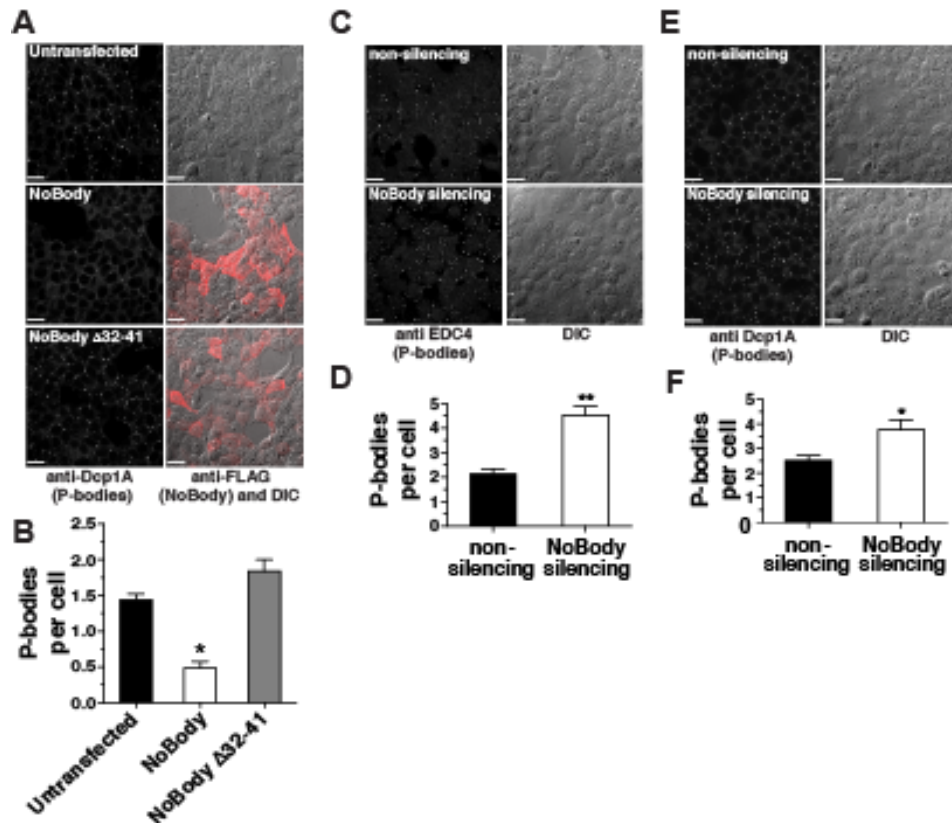


Figure 4.7. NoBody dissociates P-bodies via interaction with EDC4, and absence of NoBody increases P-body numbers. (A) Representative images of HEK293T cells, cells virally transduced with NoBody, and the non-EDC4-interacting mutant NoBody Δ 32-41. Endogenous P-bodies were visualized with anti-Dcp1A immunofluorescence, and NoBody and NoBody Δ 32-41 were detected with anti-FLAG immunofluorescence. P-bodies were counted by manual masking followed by object detection with SlideBook software. All scale bars, 20 μ m. (B) Four fields of view, including the representative images, were used to quantitate average numbers of P-bodies per cell, representing >377 cells, in each average. Means were calculated for each field of view individually, then averaged; error bars are standard deviation of

Figure 4.7. (Continued) the mean across fields of view. NoBody expression significantly decreases the average number of P-bodies per cell (*, $p < 0.0001$, ANOVA), while NoBody deletion expression is not significantly different from untransfected cells. DIC, differential interference contrast. (C-D) HEK293T cells were transfected with non- or NoBody-silencing siRNA, then fixed and endogenous P-bodies were detected using either anti-Dcp1A (C,D) or anti-EDC4 (E,F) immunofluorescence. P-bodies were counted by manual masking followed by object detection using SlideBook software. For quantitation of P-bodies per cell, >6 fields of view were analyzed, totaling >400 cells for each measurement. Means were calculated for each field of view individually, then averaged; error bars are standard error of the mean (by field of view). The increase in P-bodies per cell upon NoBody silencing is statistically significant in each case by t-test (*, $p < 0.05$, **, $p < 0.01$).

Viral transduction of NoBody into HEK293T resulted in detectable expression of the protein by cellular imaging and western blot (Figure 4.7A, Figure 4.8). As a control, we included a NoBody deletion construct lacking amino acids 22-41 (NoBody Δ 22-41) (Figure 4.7A). We then measured endogenous P-bodies using the Dcp1a marker. Unexpectedly, we found cells overexpressing NoBody had a substantial decrease in P-body numbers, with most transfected cells completely lacking P-bodies (Figure 4.7A, Figure 4.8A). Furthermore, NoBody's effect on P-body numbers correlates with NoBody binding to EDC4 because cells expressing NoBody Δ 22-41 did not have a significant change in P-body levels (Figure 4.7A). Transduction of NoBody into HeLa, BGC-823, and COS7 cells resulted in the disappearance P-bodies demonstrating that this effect is not cell line-specific.

These experiments necessarily introduce an exogenous RNA while simultaneously measuring changes in the organization of proteins involved in mRNA processing. Therefore, we needed to ensure that the decrease in P-body numbers is not due to the NoBody mRNA. The NoBody Δ 22-41 control suggests that the NoBody polypeptide, not the NoBody RNA, is the required factor to lower P-body levels, but we wanted to test this conclusion more rigorously.

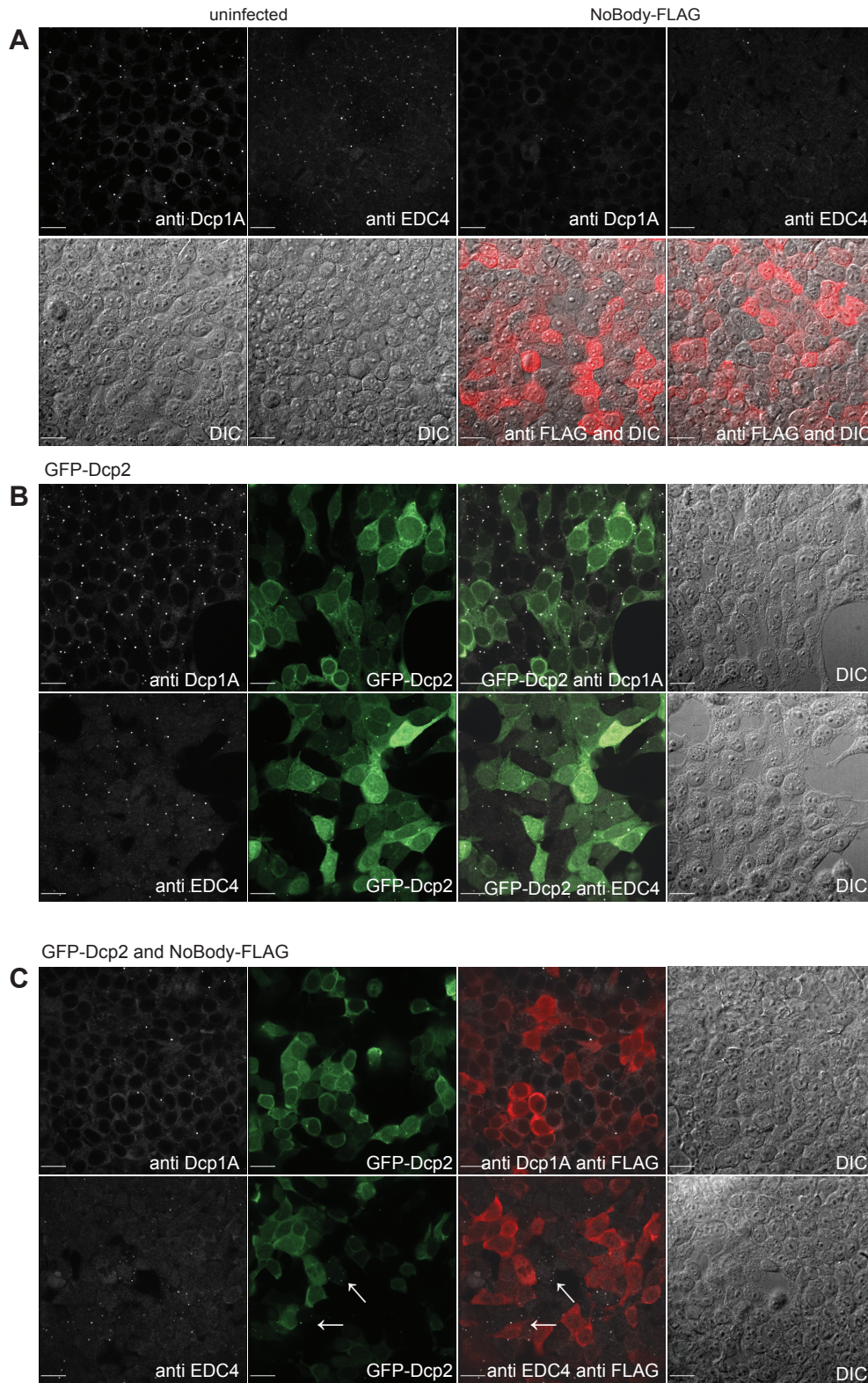
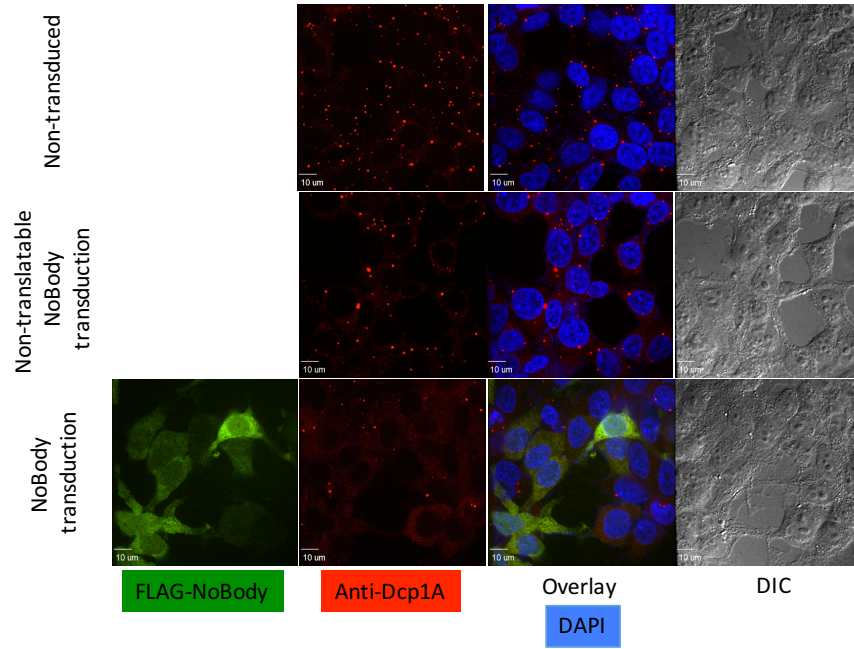


Figure 4.8. NoBody disrupts P-bodies and GFP-Dcp2 P-body localization.

Figure 4.8. (Continued) (A) HEK293T cells were retroviral transduced with NoBody-FLAG to achieve low expression levels. Comparison to uninfected cells reveals that cells expressing NoBody (as assessed by anti-FLAG immunofluorescence) exhibit few to no P-bodies (detected as puncta via EDC4 and Dcp1A immunofluorescence). (B) Retrovirally transduced GFP-Dcp2 distributes between diffuse cytoplasmic localization and puncta; the puncta are coincident with endogenous P-bodies as assessed by anti-Dcp1A and anti-EDC4 immunofluorescence. (C) GFP-Dcp2 puncta are absent from cells co-expressing NoBody (anti-FLAG immunofluorescence). Two cells expressing GFP-Dcp2 but no NoBody do exhibit GFP-Dcp2 puncta, which are coincident with endogenous P-bodies (anti EDC4 immunofluorescence). All scale bars, 20 μ m.

We compared P-body numbers in cells transfected with a NoBody expression vector to cells transfected with mutated NoBody expression vector that lacks the NoBody start codon. This mutated vector can produce RNA but no protein. While NoBody overexpression results in loss of cellular P-bodies, the expression vector lacking the NoBody start codon had no effect on P-body numbers (Figure 4.9). Thus, the NoBody polypeptide is responsible for P-body disappearance in cells.

Conversely, we found that silencing NoBody expression increases the number of P-bodies. We observed a 1.5- to 2-fold increase in P-bodies per cell when NoBody was silenced (Figure 4.7C-F). Silencing of an unrelated gene, GAPDH, did not affect P-body numbers, demonstrating that P-body disappearance is not due to a cellular response to siRNA (Figure 4.10). The reciprocal change in P-body numbers upon NoBody expression and silencing establishes NoBody as a regulator of P-body formation. To our knowledge, this a unique activity for a P-body associated protein. Together these experiments indicate that NoBody controls the distribution of the 5'-3' mRNA decay proteins between P-bodies and soluble complexes in the cytosol.



Mean P-bodies per cell

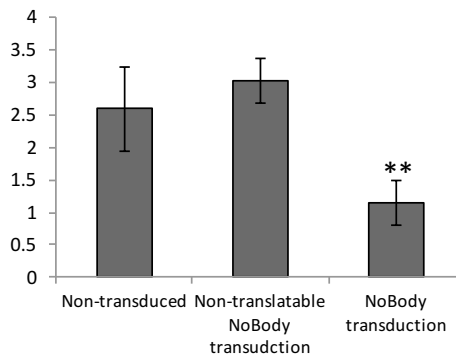


Figure 4.9. P-body dissociation requires translated NoBody peptide, and does not depend on the *LOC55043* RNA. While expression constructs for P-body quantitation consist of only the NoBody peptide coding sequence rather than the full-length *LOC550643* RNA, it remains possible that that RNA sequence could have a non-coding function that affects P-body numbers. We therefore generated a non-translatable construct with the NoBody start codon deleted. Both translatable and non-translatable (-ATG) NoBody expression constructs were incorporated into lentiviruses and transfected into HEK293T cells. Cells were stained with anti-FLAG to detect NoBody expression (green), anti-Dcp1A to quantify P-bodies (red), and DAPI to stain nuclei and count cells (blue). Mean P-bodies per cell were computed as a population average for >4 fields of view (at least 120 cells per sample) and standard deviation was calculated across fields of view. Significance was determined by t-test, $p = 2.5 \times 10^{-5}$. Untransfected cells and -ATG transfected cells show no statistically significant difference in P-body numbers.

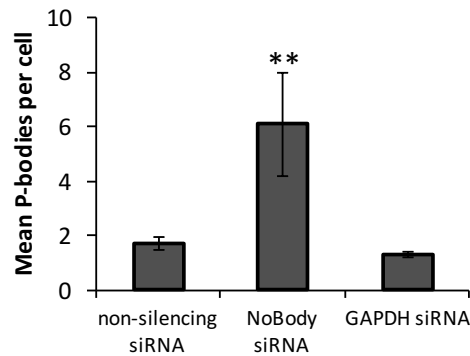
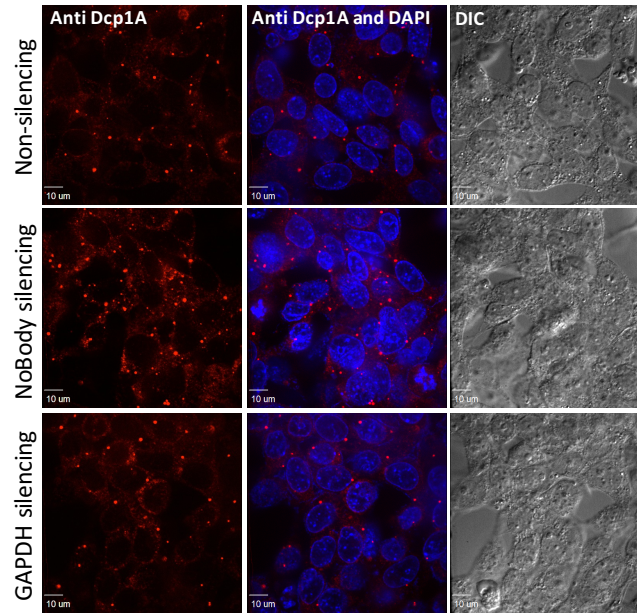


Figure 4.10. Increased P-body numbers are a specific effect of *LOC550643* silencing. In order to verify that increase P-body numbers are specific to *LOC550643* gene silencing and not a non-specific RNAi effect, we compared P-body numbers in HEK293T cells transfected with a non-silencing control siRNA, a *LOC550643*-targeted siRNA set, and an siRNA set targeted to a non-specific housekeeping gene, *GAPDH*. Cells were fixed 48 hours after transfection and stained with anti-Dcp1A (red) and DAPI (blue). Mean P-bodies per cell were computed as the ratio of P-bodies to nuclei for >4 fields of view (at least 115 cells per measurement), and significance was determined by t-test ($p=0.003$). No statistically significant difference in P-bodies per cell was observed between the non-silencing and *GAPDH*-silencing samples.

4.2.5. NoBody action and impact on P-bodies

We hypothesized that NoBody interacts with existing P-bodies prior to their disappearance, and also assumed that NoBody activity corresponds to NoBody concentrations. If so, lower NoBody levels might lead to slower elimination of P-bodies, providing an opportunity to observe a NoBody-P-body interaction. The expression of NoBody with low-titer levels of lentivirus resulted in a small population of cells (< 10% of total) with NoBody localized to the P-body (Figure 4.11A,B) and offers evidence in support of NoBody complexing with P-bodies prior to the disappearance of the P-bodies.

A key question was whether NoBody was affecting P-body numbers by controlling the endogenous levels of P-body associated proteins or whether it was affecting the formation of the P-body foci. We determined if NoBody was promoting P-body protein degradation by measuring the levels of Dcp1A in the presence and absence of NoBody. The level of endogenous Dcp1A, for example, was identical by Western blot between control and NoBody transfected or NoBody transduced cells (Figure 4.11C). Based on this data the most likely mechanism is that NoBody binds to the P-body and this binding then leads to P-body dissolution.

One intriguing question is whether these NoBody-regulated structural changes affect the activity of the 5'-3' mRNA decay pathway, which is reliant on P-body proteins. To understand the contribution of NoBody to the HEK293T cellular RNA profile, we silenced NoBody and measured changes by microarray.

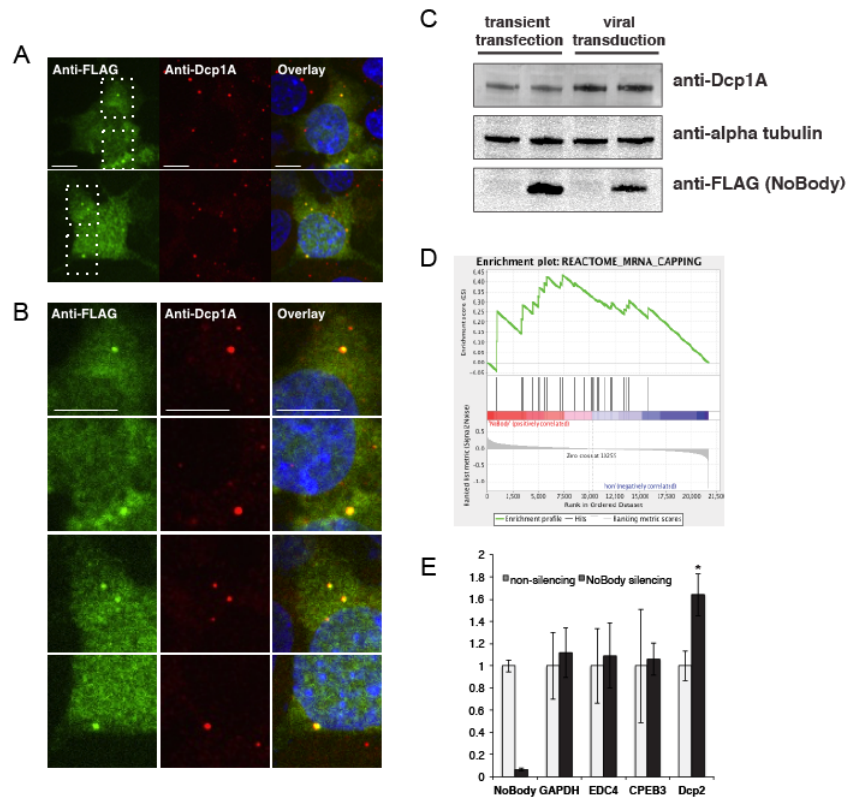


Figure 4.11. NoBody co-localizes with P-bodies in a subpopulation of cells at very low expression levels. HEK293T cells were transfected with lentivirus at low multiplicity of infection to introduce FLAG-NoBody at the lowest detectable expression level. After fixation and immunostaining for a P-body marker (Dcp1A, red) and for FLAG-NoBody (green), then counterstained with DAPI to visualize the nucleus (blue). In the merged image, NoBody/Dcp1A co-localization appears yellow. (A) In a minority of transfected cells (~10%), NoBody reproducibly forms puncta that co-localize with P-bodies. (B) Zoom-in image of boxed regions from (A) demonstrates co-localization. (C) NoBody overexpression does not affect endogenous Dcp1A protein levels. NoBody was either transiently transfected, using Lipofectamine 2000, into HEK293T cells (with mock transfection as a control) or retrovirally transduced (with untreated cells as a control), then cell were lysed and lysates probed by Western blotting. Alpha tubulin served as a loading control. (D) Correlation plots for functionally relevant gene sets that are upregulated (mRNA decapping, top) by NoBody silencing. (E) qRT-PCR validation of increased Dcp2 expression in cells silenced for NoBody expression in comparison to non-silencing siRNA-transfected cells. No actinomycin D treatment was performed in this experiment. NoBody silencing of >90% was performed in parallel, and an unrelated gene (GAPDH) and other mRNA degradation-related proteins not implicated as changing in the array data (EDC4, CPEB3) show no expression changes, validating the specificity of the Dcp2 transcriptional response.

We also used Dcp2, DcpS (an unrelated cap scavenging protein) and GAPDH silencing as controls. Gene silencing of >90% was confirmed for NoBody, Dcp2, DcpS and GAPDH using qPCR. We analyzed each set of experimental samples to determine how many transcripts have statistically significant ($p < 0.05$) changes in expression level. We observed 2,230 unregulated and 2070 down regulated transcripts upon NoBody silencing. NoBody, Dcp2, DcpS and GAPDH silenced cells had distinct profiles.

To understand the cellular role of NoBody, we utilized Gene Set Enrichment Analysis (GSEA) against the C2:CP biological pathway database. GSEA searches databases of gene sets, groups of genes with common functions, to identify biological pathways for a given perturbation (28). For example, GSEA analysis of our Dcp2 knockdown data identified a nonsense mediated decay gene set as the best match, consistent with Dcp2 function (Figure 4.12). GSEA analysis of NoBody silencing identified three matched gene sets that are associated with RNA metabolism, including mRNA capping (Figure 4.11D). These gene sets were not regulated by GAPDH and DcpS knockdown, providing evidence of a specific biological effect of NoBody silencing. Upon analyzing the genes that drove these matches we discovered that NoBody silencing was affecting Dcp2 expression (Figure 4.13). We confirmed that NoBody increases Dcp2 expression by qPCR analysis of HEK293T RNA after NoBody silencing (Figure 4.11E). Dcp2 levels were elevated in the absence of NoBody while levels of control transcripts EDC4 and GAPDH were unchanged. By this type of unbiased functional genomics analysis we find additional data in support of our biochemical and cellular studies that NoBody is involved in the mRNA decapping pathway.

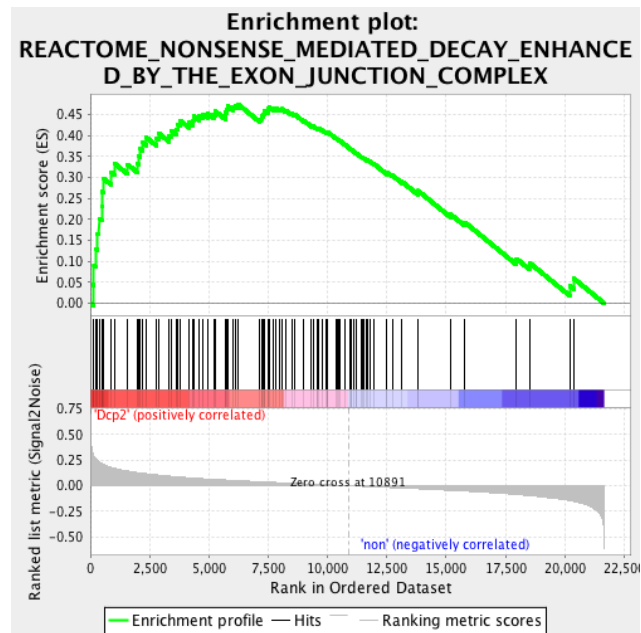


Figure 4.12. GSEA match for Dcp2 silencing. Complete list of up-regulated gene sets and associated normalized enrichment factors for Dcp2-silenced cells vs. non-silencing siRNA control. While many gene sets are enriched, the top hit corresponds to up-regulation of nonsense-mediated decay.

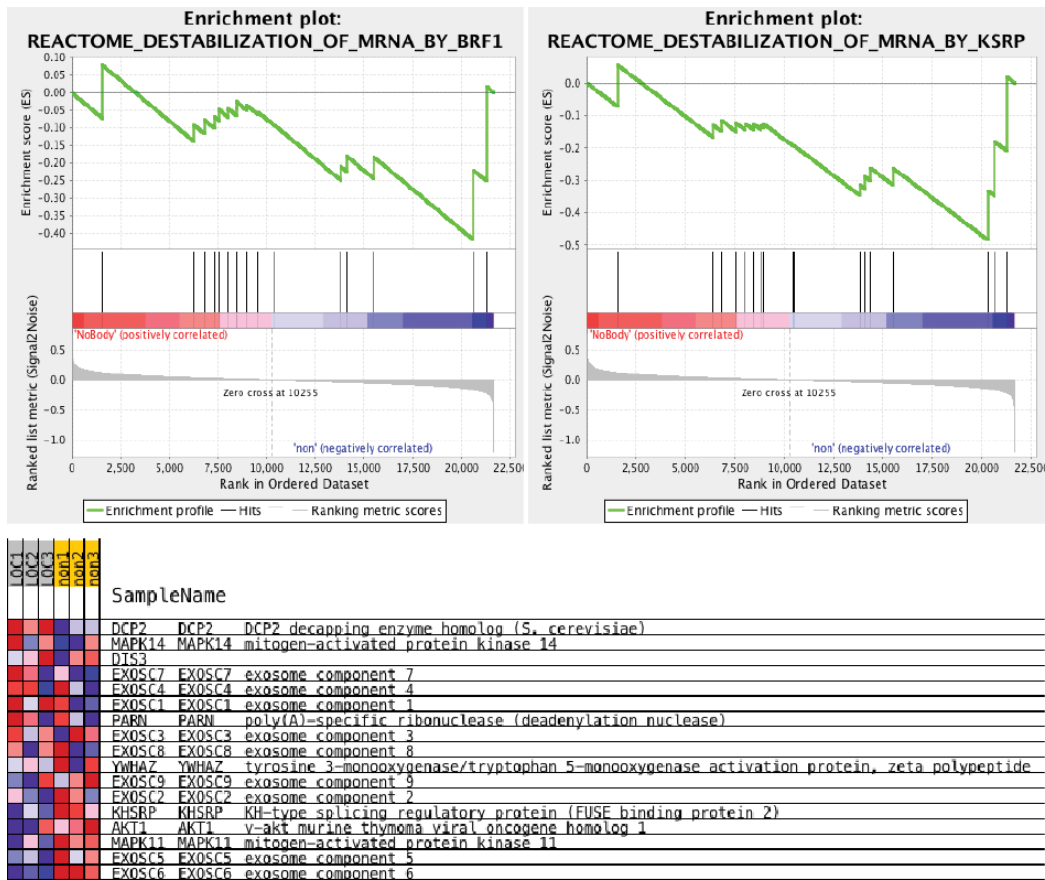


Figure 4.13. GSAE matches and top gene changes for NoBody silencing. Correlation plots for functionally relevant down-regulated gene sets (mRNA degradation by BRF-1 and KSRP, left and right, respectively). Microarray signals (black lines) are plotted against the known gene changes comprising the gene set to show positive (top) and negative (center and bottom) correlation. Heat map for mRNA degradation by KSRP gene set is shown to demonstrate the importance of the Dcp2 elevation phenotype in cells silenced for NoBody expression.

4.3. Discussion

Rapid advancements in technology have improved our ability to detect molecules in cells and tissues. RNA-Seq data has been used to create protein databases of all potential proteins in cells and tissues (10, 29). Coupled to ribosome-sequencing (Ribo-Seq) (30-34) and/or proteomics (10, 12, 29), this approach can identify new protein-coding genes, including novel

SEPs in flies, mice and humans. The total number of SEPs in flies and humans has now reached the thousands (12, 30). Aspden and colleagues developed a method named Poly-Ribo-Seq to identify over three thousand translated small open reading frames (smORFs) in *Drosophila* cells(30). For human SEPs, Vanderperre and colleagues utilized alternative ORF prediction (AltORF) and proteomics to identify 1,259 nonannotated proteins; most were SEPs (12). Work remains to be done in the area of SEP discovery, particularly in regards to the ability to profile changes in SEP levels.

Despite large numbers of SEPs, only a handful is characterized (7, 35-41). Interest in this field will only be driven if these molecules are functional. In flies, the recent description of the sarcolamban SEP with activity in heart rhythm was a terrific discovery that highlights a physiological role for SEPs (7). A mouse homolog of sarcolamban, myoregulin (40), was recently shown to control muscle contraction in mice, and a SEP from pancreatic mitochondrial DNA called MOTS-c has anti-diabetic activity in mice (41). To this growing list, we can now add NoBody, a conserved human SEP with a fundamental role in controlling P-body numbers. Our findings add to the conclusion that SEP biology is an emerging field with potential to shape our understanding of the functional proteome.

This work also contributes to the elucidation of P-body regulation. The initial discovery that decapping proteins localize to the P-body in yeast led to the proposal that P-bodies are sites of mRNA degradation (26). However, this model has been called into question because P-bodies are not necessary for NMD-mediated mRNA decay (42, 43), and loss of essential 5'-3' mRNA decay enzymes increase P-bodies (26, 27). It is, therefore, possible that P-bodies are storage granules where translationally inactive mRNAs are stored prior to degradation or release back into polysomes (23, 44). NoBody acts on the P-body to control P-body nucleation, and provides

a molecular mechanism to affect the phase transition of the decapping complex between the cytosol and the P-body.

Furthermore, the increase in P-body numbers upon NoBody silencing indicates that NoBody actively maintains P-body numbers. The functional characterization of NoBody contributes to understanding of P-body regulation, and promises to provide further insight into RNA regulation. Finally, P-bodies belong to a larger group of RNA granules, which include Cajal bodies, stress granules, germ granules, PML granules and neuronal granules (45). Experiments have recently elucidated fundamental structural requirements (multivalent interactions between proteins (46) and low complexity domains (47, 48)) that mediate the assembly of RNA and proteins into granules (47, 48). Because these physical organizing principles of ribonucleoprotein granules are general, elucidating the mechanism by which the NoBody-EDC4 interaction regulates P-body formation may provide greater insight into granule formation and function.

4.4. Materials and Methods

Microarray data are deposited in NCBI GEO under accession number GSE67632.

4.4.1. Cell culture and transfection

HEK293T and HeLa cells were culture in DMEM supplemented with 10% fetal bovine serum, penicillin, and streptomycin. Cells were maintained in a 5% CO₂ atmosphere at 37°C. Plasmid transfection was performed with Lipofectamine 2000 and Opti-MEM according to the manufacturer's instructions. siRNA specific for LOC550643 and a non-silencing siRNA were obtained from Qiagen and transfected using Dharmafect 1 according to the manufacturer's

instructions. For immunoprecipitation, cells were lysed 24 hours post-transfection; for immunofluorescence imaging, cells transfected with plasmids or siRNA were assayed 36 or 48 hours post-transfection, respectively. Cells transduced with retrovirus were assayed 48 hours post-infection.

4.4.2. Retrovirus production and transduction

Retrovirus was produced essentially as previously described(49). Briefly, HEK293T cells were transfected with construct in pFCPGW, along with pVSV-G and pdelta8.91, and growth media replaced 5 hours later. 48 hours post-transfection, lentivirus was harvested, filtered through 0.45- μm filter, aliquoted and flash-frozen. For viral titering and transduction, dilutions of lentivirus-containing media (1:250-1:5) were added to cells in growth media for 5 hours, then media was replaced with fresh complete growth media and cells were analyzed 48 hours later.

4.4.3. Western blotting

Lysates (5 μL) and IP samples (10 μL) were mixed with protein loading buffer, boiled, and separated on 4-20% Tris/glycine SDS-PAGE gels (BioRad). For analysis with multiple antibodies, replicate gels were run. Proteins were transferred to Immobilon-FL PVDF (Millipore) for 2 h at 400 mA. Immunoblots were blocked with Rockland fluorescent blocking buffer, then probed with primary antibodies at a 1:1000 dilution in the same buffer for ~2 hours at 4°C. The membrane was washed three times with TBS-T, then secondary antibodies were applied at a dilution of 1:4000 in Rockland fluorescent blocking buffer. After a final 3x TBS-T wash, infrared imaging was performed on a LICOR Odyssey instrument. Alternatively, secondary antibody-horseradish peroxidase conjugates were applied in blocking buffer at a dilution of

1:10,000 for 1 hour at room temperature, followed by development using Clarity ECL Western Blotting Substrate (BioRad) and imaging for 1-10 minutes in a ChemiDoc-It Imager with BioChemi 510 CCD camera (UVP).

4.4.4. Cloning and genetic constructs

PCR and restriction cloning were performed with standard techniques. A cDNA clone encoding a fragment of LOC550643 including the SEP-encoding sORF was obtained from Open Biosystems. The NoBody coding sequence was subcloned with a C-terminal FLAG epitope tag into pcDNA3 and used for all experiments unless specifically stated otherwise. A separate construct encoding the entirety of the NCBI cDNA sequence for LOC550643 was synthesized by Genscript, with a FLAG epitope tag at the 3' end of the NoBody coding sequence, then subcloned into pcDNA3; the full-length cDNA construct was utilized only for Fig. 1, and all subsequent experiments utilized the NoBody coding sequence only (or fusions and mutants thereof). The NoBody coding sequence was subcloned into pEGFP-N1 with HindIII and BamHI restriction sites. The AUG start site of EGFP was mutated to GGG by using QuikChange Mutagenesis Kit (Agilent Technologies). Deletion mutants of NoBody were then generated by inverse PCR using Q5 Site-Directed Mutagenesis Kit (NEB) according to the manufacturer's instructions. A cDNA clone encoding EDC4 was obtained from Open Biosystems, and c-myc epitope tag and deletions were created in this construct using standard inverse PCR. Deletion mutants of NoBody were also generated by inverse PCR. GFP-Dcp1A and GFP-Dcp2 were obtained from Addgene via Eliza Izaurralde (47). Retroviral constructs were subcloned into the pFCPGW plasmid.

4.4.5. Antibodies

Primary antibodies for Western blotting and immunofluorescence were: mouse monoclonal anti-FLAG M2 (Sigma), rabbit anti-c-myc (Sigma), rabbit anti-Dcp1A C-terminal polyclonal (Sigma), rabbit anti-EDC4 N-terminal polyclonal (Sigma), rabbit anti-PATL1 polyclonal (Abcam), rabbit anti-Dcp2 polyclonal (Novus), rabbit anti-Dcp1B polyclonal (Novus), rabbit anti-EDC3 polyclonal (Abcam), and rabbit anti-GFP N-terminal polyclonal (Sigma). Secondary antibodies for Western blotting are goat anti-mouse IR dye 800 and goat anti-rabbit IR dye 680 (LICOR), or goat anti-mouse-horseradish peroxidase and goat anti-rabbit-horseradish peroxidase conjugates (Rockland). Secondary antibodies for immunofluorescence are goat anti-mouse Alexa Fluor 568, and goat anti-rabbit Alexa Fluor 488, 568, and 647 (Life Technologies). Immunoprecipitation was performed with the following antibody beads: anti-FLAG M2 affinity gel (Sigma), anti-myc tag agarose beads (MBL International Corp.).

4.4.6. Peptidomics and LOC550643 SEP identification

The peptidomics experiments in which NoBody was detected in K562, H293T and MDA-MB-231 cells were previously reported (8, 10). NoBody was excluded from the originally reported list of SEPs because predicted proteins in the non-redundant protein database (NCBI) were filtered out.

4.4.7. Conservation analysis

A translated nucleotide BLAST (tBLASTn) search of the NCBI NR nucleotide database (50) was performed with standard parameters (max target sequences 100, expect threshold 10, word size 3) to assess NoBody conservation. In order to directly probe the zebrafish genome for

similar sequences, the NoBody sequence was searched against the zebrafish genome (14) using the BLAT algorithm (15) via the UCSC genome browser (16, 51). Both NoBody DNA and protein sequences were used as the query sequence under default search parameters. The FlyBase (13) web interface was used to probe the *Drosophila melanogaster* genome for sequences similar to the NoBody protein coding sequence via tBLASTn (52) search with either default or relaxed parameters (no low complexity filter, expect value 100).

4.4.8. Co-immunoprecipitation and proteomics

FLAG-tagged NoBody in pcDNA3 (or empty pcDNA3 vector as a negative control) were transfected into HEK293T cells using 10 µg DNA per 10 cm dish of cells. 24 hours post-transfection, cells were harvested and lysed using tris-buffered saline (TBS) with 1% Triton X-100 and Roche Complete protease inhibitor cocktail tablets. 400 µL lysis buffer was used per pellet. Cells were lysed on ice for 20 min followed by centrifugation at 14000 rpm, 4°C, 15 min. Lysate samples were saved for analysis of loading. A 50 µL aliquot of anti-FLAG agarose beads (clone M2, Sigma) was washed with 1 mL TBS-T, collected by centrifugation for 1 min at 3000 rpm, then suspended in the cell lysate supernatant. Bead suspensions were rotated at 4 °C for 1 hour, then washed 3 times with TBS-T. Elution was in 30 µL of 3X FLAG peptide (Sigma), at a final concentration of 100 µg/mL in TBS-T at 4 °C for 1 hour. Beads were removed by centrifugation and the entire supernatant was loaded on an SDS-PAGE gel.

For proteomics, immunoprecipitates were separated by SDS-PAGE with Coomassie stain, and a 70-kDa band visibly elevated in the IP sample relative to the negative control was identified. For quantitation via spectral counting, the gel was cut straight across with a clean razor to excise the same molecular weight band in each sample. Interaction candidates identified

by proteomics were subsequently confirmed by co-immunoprecipitation and Western blotting with antibodies against the endogenous proteins.

Protein-containing gel slices were digested with trypsin overnight. The resulting peptide mixtures were extracted from the gel and run directly on an Orbitrap Velos instrument (Thermo Fisher Scientific) with 90-minute liquid chromatography and tandem mass spectrometry (LC-MS/MS) using a standard TOP20 method procedure. Briefly, MS1 m/z regions for 395 –1600 m/z ions were collected at 60K resolving power and used to trigger MS/MS in the ion trap for the top 20 most abundant ions. Active dynamic exclusion of 500 ions for 90 sec was used during the LC-MS/MS method. Peptides were eluted with 300 nL/min flow rate using a NanoAcquity pump (Waters). Samples were trapped for 15 minutes with flow rate of 2 μ L/min on a trapping column 100 micron ID packed for 5cm in-house with 5 μ m Magic C18 AQ beads (Waters) and eluted with a gradient to 20 cm 75 micron ID analytical column (New Objective) packed in-house with 3 μ m Magic C18 AQ beads (Waters).

Mass spectra were analyzed using our in-house Proteome Browser System against the uniprot_human database. Carbamidomethylated cysteines were set as a fixed modification, with oxidation of methionine and N-terminal acetylation as variable modifications. A mass deviation of 20 ppm was set for a MS1 peaks and 0.6 Da was set as maximum allowed for MS/MS peaks and a maximum of two missed cleavages were allowed. Maximum false discovery rates were set to 0.01 both on peptide and protein levels. Minimum required peptide length was five amino acids.

Protein quantitation was accomplished via spectral counting, where the number of total peptides observed for each identified protein was taken as the total spectral counts and compared for the IP vs. negative control sample. All proteins elevated >10-fold relative to the negative

control and present at >20 spectral counts in the IP sample were considered candidates for confirmation by Western blotting.

4.4.9. NoBody photo-cross-linking

HEK293 cells were transiently transfected with EDC4-myc-pCMV-Sport6 using Lipofectamine 2000. Cells were harvested after 20 hours of transfection, lysed in RIPA buffer supplemented with protease inhibitor tablets. Cell debris was spun down at 20,000 g for 20 min at 4 °C. Total cell lysate (both transfected and non-transfected) was diluted to 1 mg/ml in PBS. 300 µl of the lysate was used for each reaction. Purified Rhodamine-NoBody(22-41)-BPA and Biotin-NoBody(22-41)-BPA peptides were purchased from Peptide 2.0. Both peptides were dissolved in DMSO to make 10mM stock. Rhodamine-NoBody(22-41)-BPA was used at 25 µM concentration, Biotin-NoBody(22-41)-BPA peptides was used at 100 µM to compete for the binding. Sample mixtures were incubated for 1 hour by rotating in 4°C, followed by UV crosslink (Stratalinker 1800 365nm) for 60 min; samples were set on ice during the crosslink. Reactions were quenched by directly adding 4X SDS loading dye to each sample, and boiled for 5 min. 30 µl of each sample was loaded on a 4-12% Bis-Tris gel and run in MES SDS running buffer at 200V for 20 min. The gel was imaged using Typhoon Imager 8600: Gain 550, 50 micron, and 532ex/580em. The same gel was stained in InstantBlue to show the total protein loading. The gel was imaged using Odyssey CLx, IR700 gray scale.

4.4.10. Microarray analysis and qRT-PCR

Gene silencing in HEK293T cells was afforded by transfection of anti-LOC550643, anti-Dcp2, or non-silencing siRNAs (Qiagen) with Dharmafect I (Fisher) for 48 hours. All samples

were performed in 3 biological replicates. Cells were treated with actinomycin D (5 μ M) for 4 hours to inhibit transcription and amplify mRNA degradation-related changes, according to a previous report (53), though we found that transcriptional changes dominated our results. RNA was purified using Qiagen RNeasy kits. RNA was then subjected to RNase-free Dnase(I) (New England Biolabs) digest to remove genomic DNA, followed by Qiagen column clean-up. Two-step qRT-PCR was performed using BioRad iScript cDNA Synthesis Kit and iTaq Universal SYBR Green Supermix according to the manufacturer's instructions. Primers used were Qiagen Quantitect Probes. RT-PCR was performed on a BioRad CF96 Touch Optical Reaction Module coupled to a C1000 Thermal Cycler with SYBR Green detection. For microarray analysis, polyadenylated RNA was isolated, fragmented and labeled using the Ambion WT Expression Kit and Affymetrix GeneChip WT Terminal Labeling and Controls Kit according to the manufacturer's instructions then hybridized to a GeneChip Human Gene 1.0 ST Array (Affymetrix) using an Affymetrix Hybridization Oven 640 and GeneChip Fluidics Station 450 (Affymetrix) according to the manufacturer's instructions. Chips were analyzed using a GeneChip Scanner 3000 7G (Affymetrix) according to the manufacturer's instructions. The resulting image files were processed with an in-house R program to generate lists of statistically significant ($p < 0.05$) gene changes, as well as input files for gene ontology analysis via the Gene Set Enrichment Analysis program.

4.4.11. Gene Set Enrichment Analysis (GSEA)

A downloadable Java implementation of the Broad Institute's GSEA software (28) was utilized to analyze the microarray data for biological pathway enrichment. Input expression files were generated via an in-house R program, and parsed into phenotypes with a manually

generated phenotype file. Each set of biological triplicate samples was compared against the non-silencing samples in turn. The c2.cp v5.0 database was searched with 1000 permutations for each set of samples to determine pathway enrichment. Only enriched (in the experimental samples) and de-enriched (in the control relative to samples) pathways with $p < 0.001$ were considered for analysis.

4.4.12. Immunofluorescence

HEK293T or HeLa cells were grown to 80% confluency on no. 1.5 glass coverslips in 48-well plates. Cells were transfected with plasmid DNA or siRNA according to the manufacturer's instructions. 36-48 hours post-transfection, cells were fixed with 4% formaldehyde in phosphate buffered saline, permeabilized with methanol at -20°C , and blocked with fluorescence blocking buffer (Rockland) for at least 1 hour at 4°C . Cells were stained with primary antibodies at a 1:500 to 1:1000 dilution in Rockland blocking buffer overnight at 4°C , followed by 3 phosphate-buffered saline (PBS) washes. Secondary antibodies were applied at a 1:1000 dilution in Rockland buffer for 1 hour either at room temperature or 4°C , then washed 3x with PBS. Cells were post-fixed with 4% formaldehyde in PBS and for experiments requiring cell counting, were subjected to nuclear counterstaining with Hoechst 33258 at a concentration of 0.02 mg/mL in PBS for 15 minutes, then imaged by confocal microscopy.

4.4.13. Confocal microscopy

Coverslips were inverted and imaged in PBS in MatTek imaging dishes. Confocal imaging was performed with kind permission in the laboratory of Alice Ting at MIT, on a Zeiss AxioObserver inverted confocal microscope with $40\times$ oil immersion objective, Yokogawa

spinning disk confocal head, Quad-band notch dichroic mirror (405/488/568/647), and 405 (diode), 491 (DPSS), 561 (DPSS), and 640 nm (diode) lasers (all 50 mW). Hoechst was imaged using 405 laser excitation, 445/40 emission; Alexa Fluor 488/GFP was imaged using 491 laser excitation, 528/38 emission; Alexa Fluor 568 was imaged using 561 laser excitation, 617/73 emission; and Alexa Fluor 647 was imaged using 640 laser excitation, 700/75 emission.

Differential interference contrast (DIC) images were also collected. All image collection and analysis were performed using Slidebook software (Intelligent Imaging Innovations).

4.5. References

1. C. R. Landry, X. Zhong, L. Nielly-Thibault, X. Roucou, Found in translation: functions and evolution of a recently discovered alternative proteome. *Current opinion in structural biology* **32**, 74-80 (2015).
2. S. J. Andrews, J. A. Rothnagel, Emerging evidence for functional peptides encoded by short open reading frames. *Nature reviews. Genetics* **15**, 193-204 (2014).
3. M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS biology* **5**, e106 (2007).
4. T. Kondo *et al.*, Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature cell biology* **9**, 660-665 (2007).
5. J. I. Pueyo, J. P. Couso, Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor. *Developmental biology* **355**, 183-193 (2011).
6. T. Kondo *et al.*, Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329**, 336-339 (2010).

7. E. G. Magny *et al.*, Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341**, 1116-1120 (2013).
8. J. Ma *et al.*, The Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *Journal of proteome research*, (2014).
9. A. G. Schwaid *et al.*, Chemoproteomic discovery of cysteine-containing human short open reading frames. *Journal of the American Chemical Society* **135**, 16750-16753 (2013).
10. S. A. Slavoff *et al.*, Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature chemical biology* **9**, 59-64 (2013).
11. G. Menschaert *et al.*, Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & cellular proteomics : MCP* **12**, 1780-1790 (2013).
12. B. Vanderperre *et al.*, Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS one* **8**, e70698 (2013).
13. G. dos Santos *et al.*, FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic acids research* **43**, D690-697 (2015).
14. K. Howe *et al.*, The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498-503 (2013).
15. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664 (2002).
16. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome research* **12**, 996-1006 (2002).
17. C. Cole, J. D. Barber, G. J. Barton, The Jpred 3 secondary structure prediction server. *Nucleic acids research* **36**, W197-201 (2008).

18. J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, G. J. Barton, JPred: a consensus secondary structure prediction server. *Bioinformatics* **14**, 892-893 (1998).
19. M. Fenger-Gron, C. Fillman, B. Norrild, J. Lykke-Andersen, Multiple processing body factors and the ARE binding protein TTP activate mRNA decapping. *Molecular cell* **20**, 905-915 (2005).
20. J. Collier, R. Parker, Eukaryotic mRNA decapping. *Annual review of biochemistry* **73**, 861-890 (2004).
21. A. Saghatelian, N. Jessani, A. Joseph, M. Humphrey, B. F. Cravatt, Activity-based probes for the proteomic profiling of metalloproteases. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 10000-10005 (2004).
22. M. Jinek *et al.*, The C-terminal region of Ge-1 presents conserved structural features required for P-body localization. *RNA* **14**, 1991-1998 (2008).
23. T. M. Franks, J. Lykke-Andersen, The Control of mRNA Decapping and P-Body Formation. *Molecular cell* **32**, 605-615 (2008).
24. A. Aizer *et al.*, The dynamics of mammalian P body transport, assembly, and disassembly in vivo. *Molecular biology of the cell* **19**, 4154-4166 (2008).
25. T. M. Franks, J. Lykke-Andersen, TTP and BRF proteins nucleate processing body formation to silence mRNAs with AU-rich elements. *Genes & development* **21**, 719-735 (2007).
26. U. Sheth, R. Parker, Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science* **300**, 805-808 (2003).
27. N. Cougot, S. Babajko, B. Seraphin, Cytoplasmic foci are sites of mRNA decay in human cells. *J Cell Biol* **165**, 31-40 (2004).
28. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

29. R. M. Branca *et al.*, HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods* **11**, 59-62 (2014).
30. J. L. Aspden *et al.*, Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife* **3**, e03528 (2014).
31. A. A. Bazzini *et al.*, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal* **33**, 981-993 (2014).
32. M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, E. S. Lander, Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240-251 (2013).
33. N. T. Ingolia *et al.*, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports* **8**, 1365-1379 (2014).
34. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
35. B. Guo *et al.*, Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **423**, 456-461 (2003).
36. M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS biology* **5**, e106 (2007).
37. T. Kondo *et al.*, Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329**, 336-339 (2010).
38. A. Pauli *et al.*, Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343**, 1248636 (2014).
39. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *The Journal of biological chemistry* **289**, 10950-10957 (2014).

40. D. M. Anderson *et al.*, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160**, 595-606 (2015).
41. C. Lee *et al.*, The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance. *Cell metabolism* **21**, 443-454 (2015).
42. A. Eulalio, I. Behm-Ansmant, D. Schweizer, E. Izaurralde, P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Mol Cell Biol* **27**, 3970-3981 (2007).
43. L. Stalder, O. Muhlemann, Processing bodies are not required for mammalian nonsense-mediated mRNA decay. *RNA* **15**, 1265-1273 (2009).
44. M. Brengues, D. Teixeira, R. Parker, Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. *Science* **310**, 486-489 (2005).
45. S. C. Weber, C. P. Brangwynne, Getting RNA and protein in phase. *Cell* **149**, 1188-1191 (2012).
46. P. Li *et al.*, Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336-340 (2012).
47. J. E. Braun *et al.*, A direct interaction between DCP1 and XRN1 couples mRNA decapping to 5' exonucleolytic degradation. *Nature structural & molecular biology* **19**, 1324-1331 (2012).
48. M. Kato *et al.*, Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753-767 (2012).
49. G. Tiscornia, O. Singer, I. M. Verma, Production and purification of lentiviral vectors. *Nature protocols* **1**, 241-245 (2006).
50. M. Johnson *et al.*, NCBI BLAST: a better web interface. *Nucleic acids research* **36**, W5-9 (2008).

51. D. Karolchik *et al.*, The UCSC Genome Browser database: 2014 update. *Nucleic acids research* **42**, D764-770 (2014).
52. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
53. Y. Li, M. Song, M. Kiledjian, Differential utilization of decapping enzymes in mammalian mRNA decay pathways. *RNA* **17**, 419-428 (2011).

Chapter 5

A Novel Mitochondrial SEP, SLC35A4-SEP

5.1. Introduction

Numerous translated short open reading frames (sORFs) in human cells and tissues have been discovered through gene prediction algorithm as well as proteomics experiments. However the functional capacity of these novel human sORF-encoded polypeptides (SEPs) is largely unknown. SEPs are proven to be functionally important in *Drosophila*; for example, sORF-encoded polypeptides (SEPs) serve critical biological roles, including regulating morphogenesis (tal/pri) (1-4) and heart rhythm (sarcolamban) (5). Over one thousand SEPs have now been reported in human cells and tissues by proteomics (6-10), but it is unknown whether these genes are functional. In order to better understand the importance of human SEPs, we characterized a SEP that resides on 5'UTR of SLC35A4 gene via functional proteomics and biochemical approaches. A large number of 5'UTR SEPs were detected by ribosome profiling and proteomics experiments in the recent years (8-14) and only a handful of these SEPs are characterized and found to be cis-regulator of the downstream protein translation. (15-21) However the function of the SEP itself is yet to be uncovered. Such studies will eventually lead to a broader understanding of SEP function, including unique roles for these polypeptides. This will also increase scientific knowledge regarding the functional human proteome that was previously overlooked.

Here we characterized a 103 amino acid long SEP, SLC35A4 SEP, in human cells. Imaging experiments revealed that SLC35A4 SEP is localized in mitochondria, which suggested that the SEP might be involved in cellular processes that are specific to its subcellular localization. Functional proteomics further provided evidence that the SLC35A4 SEP is possibly a part of mitochondrial protein complex. SLC35A4 SEP knockout cell line was generated using CRISPR/Cas9- mediated genome editing. By measuring the mitochondrial respiration rate of

SLC35A4 SEP wild type (WT), over expressed (OE), and knockout (KO) cell lines, we elucidated that the basal and maximal respiration rate correlates well with SLC35A4 SEP expression level. The discovery and characterization of SLC35A4 SEP reveals novel possible mechanism for mitochondrial respiration enhancement and highlights an essential function for a human SEP.

5.2. Results and Discussion

5.2.1. SLC35A4 gene expression and conservation analysis

SLC35A4 gene is currently annotated as solute carrier family 35 member A4 in UniProt KB (Q96G79) as a 324 amino acid long protein coding gene. Its annotation score is low; experimental evidence is limited at the transcript level. SLC35A4 is coded on the positive strand of chromosome 5 at q31.3 (Figure 5.1A). It consists of three exons and two introns; a single transcript spanning the three spliced SLC35A4 exons is currently annotated in RefSeq. The annotated solute carrier protein is predicted to be translated solely from exon three, and a 5'UTR SEP is translated from exon one, two and part of the exon three (Figure 5.1A), and this SEP is annotated as “alternative protein product” in UniProt as well as RefSeq human protein database. High throughput RNA sequencing of K562 (Chapter 2-4) and HEK293 cell lines have provided evidence for expression of SLC35A4 gene with high RNA read coverage (Figure 3.1B). Further by looking into existing ribosome profiling (Ribo-seq) data on HEK293 cell line(22), we noticed that the translation of SLC35A4 gene might be limited to the SEP. Ribo-seq is one way of evaluating whether a gene is translated or not by isolating ribosome protected RNA fragments then perform high throughput sequencing. In this case, we see high read coverage from Ribo-seq on SLC35A4 exon one, two and three where the SEP is translated, but very low and truncated

coverage on annotated protein coding region on exon three (Figure 5.1B). This suggests that possibly this gene is mistakenly annotated and only the SEP is translated under normal condition. Conservation analysis of the SEP sequence shows strong conservation across mammals (Figure 5.1C). PhyloCSF analysis also predicted that SLC35A4 SEP has clear evolutionary signature (Chapter 3), likely to have important biological function.

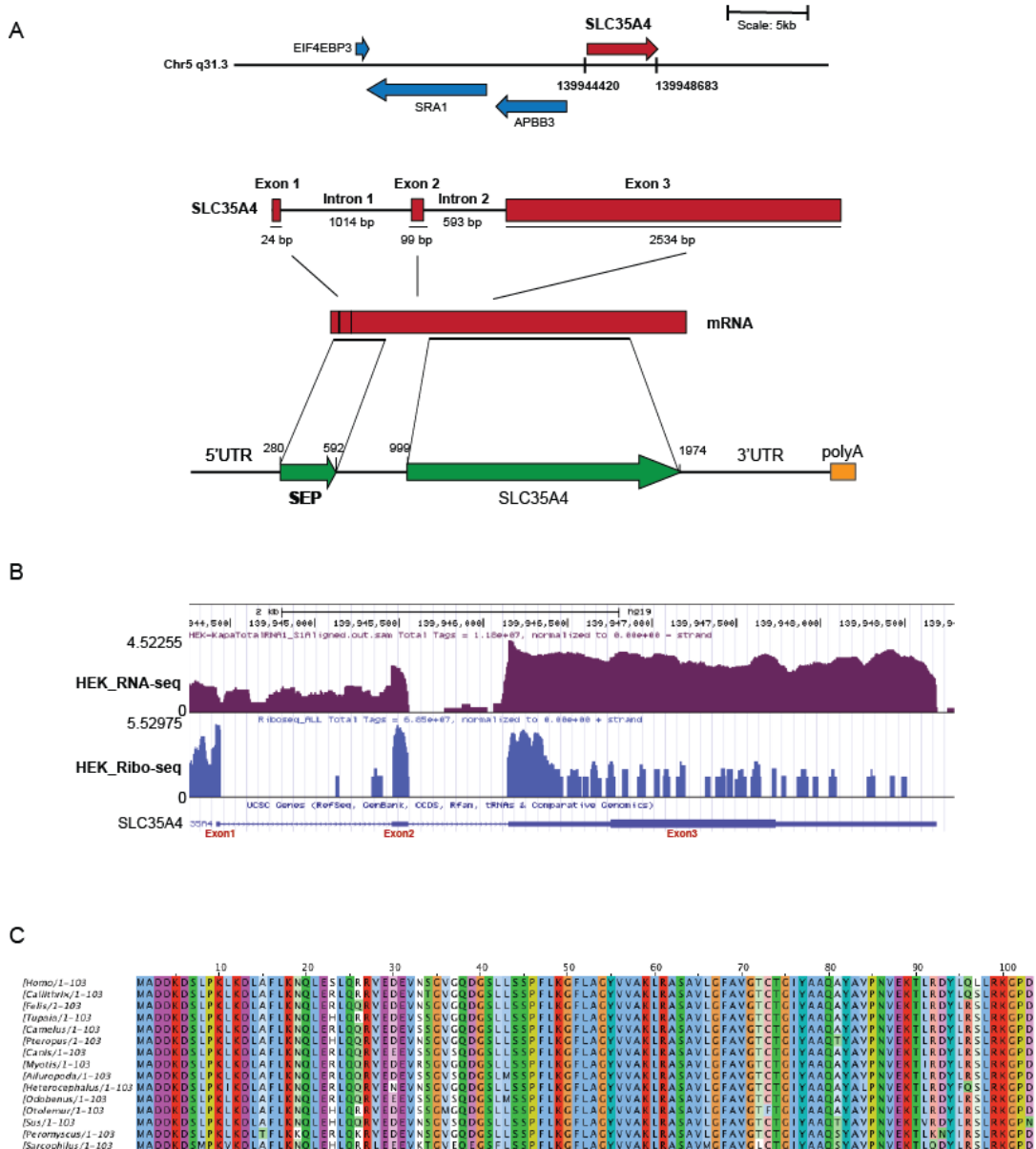


Figure 5.1. SLC35A4 gene structure.

Figure 5.1. (Continued) SLC35A4 gene structure, expression and conservation. (A) SLC35A4 (red arrow) is coded on the positive strand of chromosome 5 at q31.3. Genomic coordinates are provided below and nearby genes are indicated by gene name and blue arrows. SLC35A4 gene consists of three exons and two introns. A single transcript spanning the three SLC35A4 exons (red boxes) is currently annotated in RefSeq. The transcript is shown as a black line, sORF corresponding to the SEP and annotated SLC35A4 protein are shown as green arrows. Orange box represents poly-A tail. (B) RNA-seq and Ribo-seq coverage of SLC35A4 in HEK293 cell line. Intensity is in log-scale. The blue bar at the bottom represents SLC35A4 gene with three exons (three thicker blue lines), the thickest portion in exon three represents annotated SLC35A4 protein coding region. (C) Conservation analysis shows that SLC35A4 SEP is conserved across mammals.

5.2.2. SLC35A4 SEP detection and cellular localization

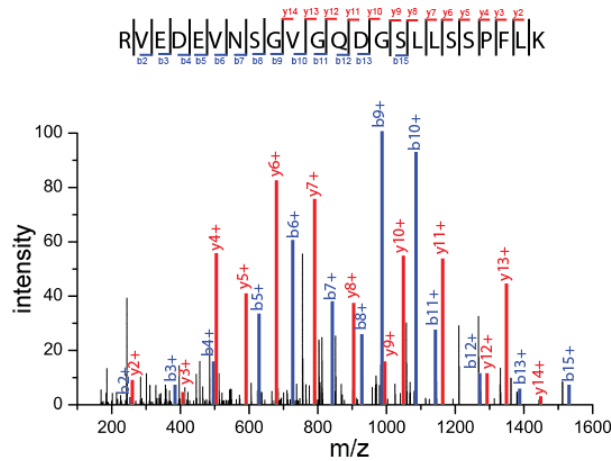
We first detected SLC35A4 SEP in K562 cells using shotgun proteomics and targeted mass spectrometry (Chapter 2). In total, four tryptic peptides were detected: NQLESLQR, RVEDEVNSGVGQDGSLLSSPFLK, GFLAGYVVAK, and TLRDYLQLLR, which gives ~50 % of the sequence coverage (Figure 5.2A). Later we have detected this SEP in additional cell lines: HEK293, A549 and HeLa (Chapter 3) and mouse T-cells and tissues (data not shown), indicating that SLC35A4 SEP is ubiquitously expressed. MS/MS spectrum of one of the most detected SLC35A4 SEP peptide (RVEDEVNSGVGQDGSLLSSPFLK) is shown in Figure 5.2A. The high-resolution spectrum and high sequence coverage of the MS/MS (consecutive b- and y-ions) represent high confidence in the SEP detection. We sub-cloned expression construct for C-terminally FLAG-tagged SLC35A4 SEP in pcDNA3.1 mammalian expression vector. Interestingly, SLC35A4-SEP localizes to mitochondria in HEK293 and HeLa (Figure 5.2B), as demonstrated by co-localization with the mitochondrial marker MitoTracker Red (Figure 5.2B). TMHMM v2.0 sequence analysis predicted a clear transmembrane motif at SLC35A4 SEP residue 62 to 84, with the N-terminal sequences on the inside and C-terminal sequences on the outside of the membrane. In order to further validate the SLC35A4 SEP endogenous expression

and localization, we generated an in-house antibody against SLC35A3 SEP. HEK293 cells were harvested and isolated into subcellular fractions: nuclear, cytosol, mitochondrial, plasma membrane (PM), and endoplasmic reticulum (ER). Western blot analysis was performed using anti-SLC35A4 SEP antibody and it provided strong evidence for mitochondrial enrichment of the SEP (Figure 5.2C). Ubiquitous expression of SLC35A4 SEP and its clear cellular localization suggest that SLC35A4 SEP is involved in mitochondrial cellular activities.

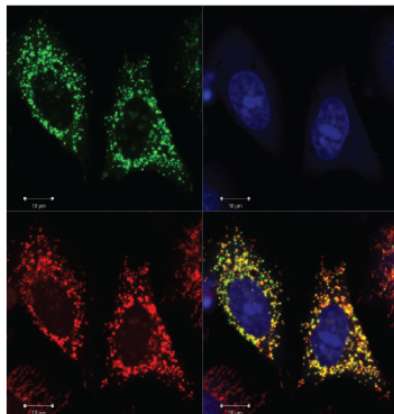
A

SLC35A4 SEP sequence:

MADDKDSLPLKLDLAF^{LN}QLES^{LQ}RRVEDEVNSGV^{GD}GS^{LL}SSPFLK^{GFLAGYV}VAKL
 RASAVLGFVAVGTCTGIYAAQAYAVPNVEK^{TLRDYLQLLR}KGPD*



B



C

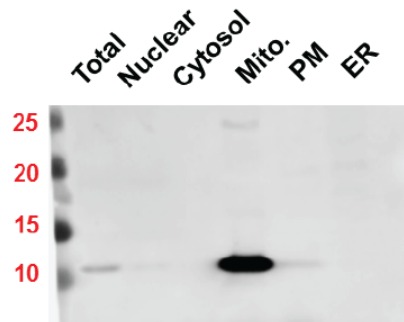


Figure 5.2. SLC35A4 SEP detection and localization.

Figure 5.2. (Continued) SLC35A4 SEP detection and localization. (A) Full sequence of SLC35A4 SEP (103 amino acid) is shown, * represents stop codon. Four tryptic peptides detected by mass spectrometry are shown in red and underlined. MS/MS spectrum of RVEDEVNSGVGQDGSLLSSPFLK is shown, blue lines represent b-ions and red lines represent y-ions detected. (B) SLC35A4 SEP-FLAG was subcloned and expressed in HeLa cells to examine its expression and localization by immunofluorescence. The SEP is detected with anti-FLAG antibody (green), nuclei are stained with Hoechst (blue). Co-staining with MitoTracker (red) indicated that SLC35A4 SEP localizes to the mitochondria (overlay). Scale bar represents 10 μ m. (C) HEK293 cells were fractionated into sub-cellular fractions: nuclear, cytosol, mitochondrial, PM and ER, and total lysate. Western blot analysis against anti-SLC35A4 SEP showed SEP enrichment in mitochondrial fraction.

5.2.3. SLC35A4 SEP enriches mitochondrial proteins involved in respiration chain

Since SEPs are short and unstructured on its own, we hypothesized that SEPs are likely to be functioning in a protein complex. Therefore we used functional proteomics to identify SLC35A4 –protein interactions. Samples were prepared in biological triplicates for statistical significance. HEK293T cells were either transfected with C-terminally FLAG epitope tagged SLC35A4 SEP in pcDNA3.1 or empty pcDNA3.1 vector as negative control. Proteomics analysis of immunoprecipitates of SLC35A4 SEP revealed the enrichment of 106 proteins with p-value less than 0.05 and at least two-fold enrichment in the SEP over expressed samples (Table 5.1). Out of which, 20% (21/106) of the enriched proteins have mitochondrial localization, which again confirms that the SEP is localized in mitochondria. We performed western blot analysis to confirm some of the most enriched proteins (TFRC, SLC16A1, Mitofilin, SLC25A3) identified by proteomics (Figure 5.3). The same immunoprecipitation experiment was repeated and the samples were blotted and probed with protein-specific antibodies to confirm enrichment of the proteins (Figure 5.3B). In all the cases tested, the enrichment of the proteins was validated by western blot analysis. Interestingly, among most of the enriched proteins in mitochondria are involved in mitochondrial respiration chain such as NADH dehydrogenase and ATP synthase

subunits (Table 5.1, in red). This provided us with hint of how SLC35A4 SEP might be functioning in cells and allowed us to develop downstream functional assay.

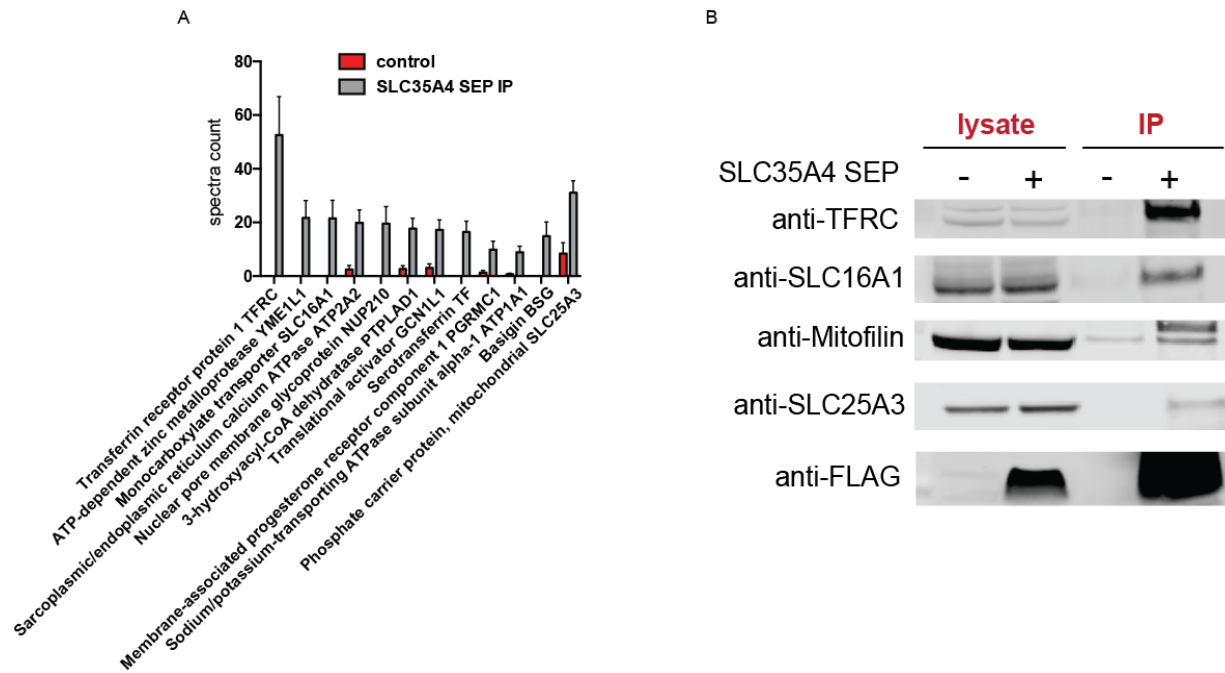


Figure 5.3. SLC35A4 SEP enriches mitochondrial proteins. (A) SLC35A4 SEP – FLAG was immunoprecipitated from lysates of transiently transfected HEK293T cells using anti-FLAG agarose gel. LC-MS/MS analysis was performed and proteins were identified by Proclix search against the UniProt human protein database appended with SEPs. Data analysis was done using IP2 ID_STAT COMPARE tool; enrichment was determined by the ratio of the average normalized spectral count of each identified protein in the immunoprecipitated samples relative to the control, with p-value less than 0.05. A partial list of the enriched proteins were plotted in bar graph. (B) Western blotting confirmation of SLC35A4 SEP interacting partners identified by proteomics.

LOCUS	Spec. count ratio: SLC/CTRL	GENE_NAME	DESCRIPTION	P-value
P02787	100000	TF	Serotransferrin OS=Homo sapiens GN=TF PE=1 SV=3	0.00061
Q6DND3	100000	HIST2H2BC	Putative histone H2B type 2-C OS=Homo sapiens GN=HIST2H2BC PE=5 SV=3	0.00924
P02786	100000	TFR3	Transferrin receptor protein 1 OS=Homo sapiens GN=TFR3 PE=1 SV=2	0.00141
P60059	100000	SEC61G	Protein transport protein Sec61 subunit gamma OS=Homo sapiens GN=SEC61G PE=2 SV=1	0.00154
Q96TA2	100000	YME1L1	ATP-dependent zinc metalloprotease YME1L1 OS=Homo sapiens GN=YME1L1 PE=1 SV=2	0.00271
Q9BV81	100000	TMEM93	Transmembrane protein 93 OS=Homo sapiens GN=TMEM93 PE=1 SV=1	0.00657
P53985	100000	SLC16A1	Monocarboxylate transporter 1 OS=Homo sapiens GN=SLC16A1 PE=1 SV=3	0.00743
Q9N6M3	100000	FITM2	Fat storage-inducing transmembrane protein 2 OS=Homo sapiens GN=FITM2 PE=2 SV=1	0.00786
P00387	100000	CY2B3	NADH-cytochrome b5 reductase 3 OS=Homo sapiens GN=CY2B3 PE=1 SV=3	0.00924
Q8TEM1	100000	NUP210	Nuclear pore membrane glycoprotein 210 OS=Homo sapiens GN=NUP210 PE=1 SV=3	0.00974
Q8WUY1	100000	C8orf55	UPF0670 protein C8orf55 OS=Homo sapiens GN=C8orf55 PE=1 SV=2	0.01102
Q14974	100000	KPNB1	Importin subunit beta-1 OS=Homo sapiens GN=KPNB1 PE=1 SV=2	0.01161
Q6DRA6	100000	HIST2H2BD	Putative histone H2B type 2-D OS=Homo sapiens GN=HIST2H2BD PE=5 SV=3	0.01174
Q15041	100000	ARL6IP1	ADP-ribosylation factor-like protein 6-interacting protein 1 OS=Homo sapiens GN=ARL6IP1 PE=1 SV=2	0.01294
P52815	100000	MRPL12	39S ribosomal protein L12, mitochondrial OS=Homo sapiens GN=MRPL12 PE=1 SV=2	0.01320
Q96UB9	100000	TMEM135	Transmembrane protein 135 OS=Homo sapiens GN=TMEM135 PE=2 SV=2	0.01373
Q7KZ99	100000	COX15	Cytochrome c oxidase assembly protein COX15 homolog OS=Homo sapiens GN=COX15 PE=1 SV=1	0.01567
O43676	100000	NDUFB3	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 3 OS=Homo sapiens GN=NDUFB3 PE=1 SV=3	0.01711
P35613	100000	BSG	Basigin OS=Homo sapiens GN=BSG PE=1 SV=2	0.01909
Q12893	100000	TMEM115	Transmembrane protein 115 OS=Homo sapiens GN=TMEM115 PE=1 SV=1	0.02216
O96005	100000	CLPTM1	Cleft lip and palate transmembrane protein 1 OS=Homo sapiens GN=CLPTM1 PE=1 SV=1	0.02396
Q75915	100000	ARL6IP5	PR1A1 family protein 5 OS=Homo sapiens GN=ARL6IP5 PE=1 SV=1	0.02681
Q9BZE1	100000	MRPL37	39S ribosomal protein L37, mitochondrial OS=Homo sapiens GN=MRPL37 PE=1 SV=2	0.02832
P08574	100000	CYC1	Cytochrome c1, heme protein, mitochondrial OS=Homo sapiens GN=CYC1 PE=1 SV=3	0.03028
P42704	100000	LRPPRC	Leucine-rich PPR motif-containing protein, mitochondrial OS=Homo sapiens GN=LRPPRC PE=1 SV=3	0.03080
Q96BR5	100000	C1orf163	Hcp beta-lactamase-like protein C1orf163 OS=Homo sapiens GN=C1orf163 PE=1 SV=2	0.03741
Q99536	100000	VAT1	Synaptic vesicle membrane protein VAT-1 homolog OS=Homo sapiens GN=VAT1 PE=1 SV=2	0.03833
O43678	100000	NDUFA2	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 2 OS=Homo sapiens GN=NDUFA2 PE=1 SV=3	0.04084
O94906	100000	PRPF6	Pre-mRNA-processing factor 6 OS=Homo sapiens GN=PRPF6 PE=1 SV=1	0.04350
Q5BJH7	100000	YIF1B	Protein YIF1B OS=Homo sapiens GN=YIF1B PE=1 SV=1	0.05953
Q9Y6A9	100000	SPCS1	Signal peptidase complex subunit 1 OS=Homo sapiens GN=SPCS1 PE=1 SV=2	0.07622
O75489	100000	NDUFS3	NADH dehydrogenase [ubiquinone] iron-sulfur protein 3, mitochondrial OS=Homo sapiens GN=NDUFS3 PE=1 SV=1	0.07939
O14920	100000	IKBK	Inhibitor of nuclear factor kappa-B kinase subunit beta OS=Homo sapiens GN=IKBK PE=1 SV=1	0.07947
Q15043	100000	SLC39A14	Zinc transporter ZIP14 OS=Homo sapiens GN=SLC39A14 PE=1 SV=3	0.08082
Q15049	100000	ATP5D	ATP synthase subunit d, mitochondrial OS=Homo sapiens GN=ATP5D PE=1 SV=2	0.08262
Q9N490	100000	PNKD	Probable hydrolase PNKD OS=Homo sapiens GN=PNKD PE=1 SV=2	0.08262
Q9H2U1	100000	DHX36	Probable ATP-dependent RNA helicase DHX36 OS=Homo sapiens GN=DHX36 PE=1 SV=2	0.08378
Q96CS3	100000	FAF2	FAS-associated factor 2 OS=Homo sapiens GN=FAF2 PE=1 SV=2	0.08434
P50416	100000	CPT1A	Carnitine O-palmitoyltransferase 1, liver isoform OS=Homo sapiens GN=CPT1A PE=1 SV=2	0.08519
Q96QU8	100000	XPO6	Exportin-6 OS=Homo sapiens GN=XPO6 PE=1 SV=1	0.08703
O75880	100000	SCO1	Protein SCO1 homolog, mitochondrial OS=Homo sapiens GN=SCO1 PE=1 SV=1	0.08841
P61803	100000	DAD1	Dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit DAD1 OS=Homo sapiens GN=DAD1 PE=1 SV=3	0.09015
Q5TG20	100000	C1orf151	UPF0327 protein C1orf151 OS=Homo sapiens GN=C1orf151 PE=1 SV=1	0.09076
P49821	100000	NDUFB1	NADH dehydrogenase [ubiquinone] flavoprotein 1, mitochondrial OS=Homo sapiens GN=NDUFB1 PE=1 SV=4	0.09155
Q9H2V7	100000	SPNS1	Protein spinstar homolog 1 OS=Homo sapiens GN=SPNS1 PE=1 SV=1	0.09184
Q00203	100000	AF3B1	AP-3 complex subunit beta-1 OS=Homo sapiens GN=AF3B1 PE=1 SV=3	0.09527
Q01650	100000	SLC7A5	Large neutral amino acid transporter small subunit 1 OS=Homo sapiens GN=SLC7A5 PE=1 SV=2	0.09549
O00461	100000	GOLIM4	Golgi integral membrane protein 4 OS=Homo sapiens GN=GOLIM4 PE=1 SV=1	0.09688
SEP71	10842.49655		SLC35A4	0.00245
P08195	37.5103523	SLC3A2	4F2 cell-surface antigen heavy chain OS=Homo sapiens GN=SLC3A2 PE=1 SV=3	0.01423
Q9H936	31.26488601	SLC25A22	Mitochondrial glutamate carrier 1 OS=Homo sapiens GN=SLC25A22 PE=1 SV=1	0.01711
Q16795	25.85364432	NDUFA9	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial OS=Homo sapiens GN=NDUFA9 PE=1 SV=2	0.04192
P0CG47	24.29789941	UBB	Polyubiquitin-B OS=Homo sapiens GN=UBB PE=1 SV=1	0.00052
P28331	23.45630106	NDUFS1	NAUH-ubiquinone oxidoreductase /5 kDa subunit, mitochondrial OS=Homo sapiens GN=NDUFS1 PE=1 SV=3	0.03371
P05023	18.8053061	ATP1A1	Sodium/potassium-transporting ATPase subunit alpha-1 OS=Homo sapiens GN=ATP1A1 PE=1 SV=1	0.00626
Q9H078	17.9563313	CLPB	Caseinolytic peptidase B protein homolog OS=Homo sapiens GN=CLPB PE=1 SV=1	0.04233
Q5VYK3	16.50136732	ECM29	Proteasome-associated protein ECM29 homolog OS=Homo sapiens GN=ECM29 PE=1 SV=2	0.08550
P35250	12.82375847	RFC2	Replication factor C subunit 2 OS=Homo sapiens GN=RFC2 PE=1 SV=3	0.05021
Q711J9	12.12632344	H2AFV	Histone H2A.V OS=Homo sapiens GN=H2AFV PE=1 SV=3	0.05058
P00403	11.57780111	MT-CO2	Cytochrome c oxidase subunit 2 OS=Homo sapiens GN=MT-CO2 PE=1 SV=1	0.02920
O00165	11.39755354	HAX1	HCLS1-associated protein X-1 OS=Homo sapiens GN=HAX1 PE=1 SV=2	0.02093
O75477	10.95643975	ERLIN1	Erlin-1 OS=Homo sapiens GN=ERLIN1 PE=1 SV=1	0.07393
O00264	10.36289898	PGRMC1	Membrane-associated progesterone receptor component 1 OS=Homo sapiens GN=PGRMC1 PE=1 SV=3	0.03564
P16615	10.11981103	ATP2A2	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2 OS=Homo sapiens GN=ATP2A2 PE=1 SV=1	0.00104
P62970	10.02093326	ATP2A3	Ubiqulin-40S ribosomal protein 33 OS=Homo sapiens GN=ATP2A3 PE=1 SV=2	0.00098
Q93084	9.127600019	ATP2A3	Sarcoplasmic/endoplasmic reticulum calcium ATPase 3 OS=Homo sapiens GN=ATP2A3 PE=1 SV=2	0.00117
Q9H9S3	8.54907962	SEC61A2	Protein transport protein Sec61 subunit alpha isoform 2 OS=Homo sapiens GN=SEC61A2 PE=2 SV=3	0.06503
Q9NV11	8.358634264	FANCI	Fanconi anemia group I protein OS=Homo sapiens GN=FANCI PE=1 SV=4	0.07812
Q9UQE7	8.051224695	SMC3	Structural maintenance of chromosomes protein 3 OS=Homo sapiens GN=SMC3 PE=1 SV=2	0.04032
P49207	7.83084541	RPL34	60S ribosomal protein L34 OS=Homo sapiens GN=RPL34 PE=1 SV=3	0.07905
Q9P036	7.641652033	PTPLAD1	3-hydroxyacyl-CoA dehydratase 3 OS=Homo sapiens GN=PTPLAD1 PE=1 SV=2	0.00339
P57089	7.443395447	TMEM33	Transmembrane protein 33 OS=Homo sapiens GN=TMEM33 PE=1 SV=2	0.01430
Q9Y5M8	7.330973448	SRPRB	Signal recognition particle receptor subunit beta OS=Homo sapiens GN=SRPRB PE=1 SV=3	0.01550
Q15392	7.307718202	DHCR24	24-dehydrocholesterol reductase OS=Homo sapiens GN=DHCR24 PE=1 SV=2	0.04974
P27824	6.794841768	CANX	Calnexin OS=Homo sapiens GN=CANX PE=1 SV=2	0.05094
Q92616	6.490887566	GCN1L1	Translational activator GCN1 OS=Homo sapiens GN=GCN1L1 PE=1 SV=6	0.00106
Q9NZ01	6.067572894	TECR	Trans-2,3-enoyl-CoA reductase OS=Homo sapiens GN=TECR PE=1 SV=1	0.01321
P39656	5.744559659	DDOST	Dolichyl-diphosphooligosaccharide-protein glycosyltransferase 48 kDa subunit OS=Homo sapiens GN=DDOST PE=1 SV=4	0.07735
P63173	5.449704227	RPL38	60S ribosomal protein L38 OS=Homo sapiens GN=RPL38 PE=1 SV=2	0.02930
Q5JTH9	5.342813063	RRP12	RRP12-like protein OS=Homo sapiens GN=RRP12 PE=1 SV=2	0.06972
P62306	5.258629638	SNRPF	Small nuclear ribonucleoprotein F OS=Homo sapiens GN=SNRPF PE=1 SV=1	0.06746
O15260	4.948152112	SURF4	Surfeit locus protein 4 OS=Homo sapiens GN=SURF4 PE=1 SV=3	0.01107
Q8TEX9	4.806369928	IPO4	Importin-4 OS=Homo sapiens GN=IPO4 PE=1 SV=2	0.04271
Q00325	4.766479403	SLC25A3	Phosphate carrier protein, mitochondrial OS=Homo sapiens GN=SLC25A3 PE=1 SV=2	0.00135
Q96566	4.749438787	CLCC1	Chloride channel CLIC-like protein 1 OS=Homo sapiens GN=CLCC1 PE=1 SV=1	0.00626
Q9Y2X3	4.726833167	NOP58	Nucleolar protein 58 OS=Homo sapiens GN=NOP58 PE=1 SV=1	0.08470
Q9NTJ3	4.448533225	SMC4	Structural maintenance of chromosomes protein 4 OS=Homo sapiens GN=SMC4 PE=1 SV=2	0.03605
P61619	4.309406391	SEC61A1	Protein transport protein Sec61 subunit alpha isoform 1 OS=Homo sapiens GN=SEC61A1 PE=1 SV=2	0.00524
P04844	4.283792085	RPN2	Dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit 2 OS=Homo sapiens GN=RPN2 PE=1 SV=3	0.02065
O75369	4.272768809	FLNB	Filamin-B OS=Homo sapiens GN=FLNB PE=1 SV=2	0.05080
P48013	4.197404278	MKI67	Antigen KI-67 OS=Homo sapiens GN=MKI67 PE=1 SV=2	0.08662
P33947	4.033158611	KDELRL2	ER lumen protein retaining receptor 2 OS=Homo sapiens GN=KDELRL2 PE=1 SV=1	0.07068
P21796	3.884132847	VDAC1	Voltage-dependent anion-selective channel protein 1 OS=Homo sapiens GN=VDAC1 PE=1 SV=2	0.00247
P08238	3.590158925	HSP90AB1	Heat shock protein HSP 90-beta OS=Homo sapiens GN=HSP90AB1 PE=1 SV=4	0.02514
P36542	3.326083636	ATP5C1	ATP synthase subunit gamma, mitochondrial OS=Homo sapiens GN=ATP5C1 PE=1 SV=1	0.00662
Q5HFF7	3.233259303	HSP90AB3P	Putative heat shock protein HSP 90-beta-3 OS=Homo sapiens GN=HSP90AB3P PE=5 SV=1	0.06031
P49755	3.125723864	TMED10	Transmembrane emp24 domain-containing protein 10 OS=Homo sapiens GN=TMED10 PE=1 SV=2	0.02309
Q9H9B4	3.098301098	SFXN1	Sideroflexin-1 OS=Homo sapiens GN=SFXN1 PE=1 SV=4	0.00123
Q32951	3.085377282	HNRNPAL2	Heterogeneous nuclear ribonucleoprotein A1-like 2 OS=Homo sapiens GN=HNRNPAL2 PE=2 SV=2	0.06239
P42166	2.798647236	TMPO	Lamina-associated polypeptide 2, isoform alpha OS=Homo sapiens GN=TMPO PE=1 SV=2	0.07031
Q15021	2.67789797	NCAPD2	Condensin complex subunit 1 OS=Homo sapiens GN=NCAPD2 PE=1 SV=3	0.06612
P06576	2.507419415	ATP5B	ATP synthase subunit beta, mitochondrial OS=Homo sapiens GN=ATP5B PE=1 SV=3	0.04744
P62266	2.493332787	RPS23	40S ribosomal protein S23 OS=Homo sapiens GN=RPS23 PE=1 SV=3	0.09010
Q02978	2.458940232	SLC25A11	Mitochondrial 2-oxoglutarate/malate carrier protein OS=Homo sapiens GN=SLC25A11 PE=1 SV=3	0.03072
Q86VP6	2.326820461	CAND1	Cullin-associated NEDD8-dissociated protein 1 OS=Homo sapiens GN=CAND1 PE=1 SV=2	0.02997
Q9NX52	2.120291104	QPCTL	Glutaminyl-peptide cyclotransferase-like protein OS=Homo sapiens GN=QPCTL PE=1 SV=2	0.01105

Table S.1.

Table 5.1. (Continued) A full list of 106 identified proteins enriched in SLC35A4 SEP-FLAG immunoprecipitated samples. Only enriched proteins with P-value < 0.1 and fold-change > 2 were retained in this list. Ratio of SLC/CTRL = 100000 represents that the protein was only detected in SLC35A4 SEP immunoprecipitated samples.

5.2.4. CRISPR/Cas9-mediated SLC35A4 SEP knockout and cellular RNA profiling

In order to elucidate the cellular function of SLC35A4 SEP, we generated a SLC35A4 SEP knockout HEK293 cell line using CRISPR/Cas9 technology. (23) Two guide RNAs (gRNAs) were designed to target the first and second exon of SLC35A4 gene (Figure 5.4A). HEK293T cells were transiently co-transfected with two gRNAs in pSpCas9 BB-2A-Puro (PX459) vector, followed by puromycin selection. Knockout (KO) efficiency was validated by western blot analysis probing against endogenous SLC35A4 SEP using in-house generated rabbit anti-SLC35A4 SEP antibody (Figure 5.4B). Samples were prepared in biological triplicates for statistical significance. As positive control, HEK293T cells were transiently transfected with FLAG tagged SLC35A4 SEP in pcDNA3.1 expression vector as over expressed (OE) samples. SLC35A4 SEP expression level was normalized by β -actin and quantification was done using Odyssey CLx built-in tool (Figure 5.4C). Complete knockout of SLC35A4 SEP expression was achieved by CRISPR/Cas9- mediated genome editing.

Next, we performed cellular RNA profiling on HEK293T wild type (WT), SLC35A4 SEP KO and OE cell lines to assess whether the SEP is involved in gene regulation. One of the most differentially up-regulated genes in the WT is MAT1A, compared to the KO cell line. This gene catalyzes a two-step reaction that involves the transfer of the adenosyl moiety of ATP to methionine to form S-adenosylmethionine and triphosphosphate, which is subsequently cleaved to PPi and Pi. S-adenosylmethionine is the source of methyl groups for most biological

methylation (RefSeq). Among the most differentially up-regulated genes in OE versus WT cell line are TGF- β and HOPX. TGF- β encodes a member of the transforming growth factor beta (TGF- β) family of cytokines, which are multifunctional peptides that regulate proliferation, differentiation, adhesion, migration, and other functions in many cell types. HOPX is an atypical homeodomain protein that does not bind DNA and is required to modulate cardiac growth and development. It acts via its interaction with SRF, thereby modulating the expression of SRF-dependent cardiac-specific genes and cardiac development (UniProt). Overall, the gene profiling experiment by RNA sequencing was not conclusive to suggest that SLC35A4 SEP is involved directly in gene regulation of a particular biological pathway. Therefore further investigation is necessary.

and WT on overall mitochondrial function were compared using a Seahorse XF96 extracellular flux analyzer to measure the oxygen consumption rate, an indicator of oxidative phosphorylation, in the presence of a series of metabolic inhibitors and uncoupling agents (Figure 5.5). Basal oxygen consumption was increased in the OE cells followed by WT and the lowest in the KO cells, which correlates well with SLC35A4 SEP expression level (Figure 5.4B, C). The rate of oligomycin- insensitive oxygen consumption, which reflects proton leakage across the inner mitochondrial membrane (24), did not show significant changes in three SLC35A4 SEP expression levels. After oligomycin, cells were injected with FCCP that permeabilizes the inner mitochondrial membrane and induces maximal, uncoupled respiration. The response to FCCP, defined as the percent increase over basal oxygen consumption, also showed highest increase in SLC35A4 SEP OE cells relative to the WT then the lowest in the KO cells (Figure 5.5). The increased cellular respiration, both basal and maximal, is in close

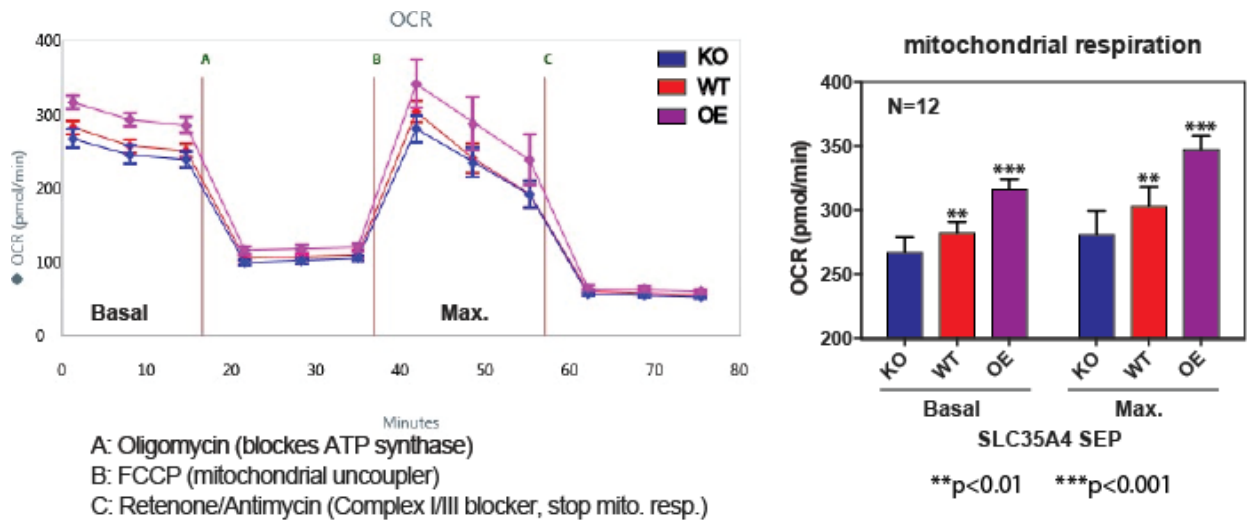


Figure 5.5. Effect of SLC35A4 SEP expression on the cellular oxygen consumption rate. Equal numbers of the transiently transfected HEK293T cells, WT and KO cells were subjected to oxygen consumption measurements in a Seahorse XF96 extracellular flux analyzer, with sequential additions of the metabolic inhibitors/activators oligomycin (A), FCCP (B), and rotenone/antimycin (C). The measurements were done in 12 wells of cells for each condition. The experiment was repeated with similar results. Oxygen

Figure 5.5. (Continued) consumption rate (OCR, an indicator of oxidative phosphorylation in pMoles/min) were measured. Three measurements were taken after each addition of mitochondrial inhibitor before injection of the next inhibitor. Basal and maximal respiration rate was plotted in bar graph representation with statistical significance.

co-relation with SLC35A4 SEP expression level, and this is not associated with respiratory chain protein abundance (Figure 5.6). In summary, the data indicates that SLC35A4 SEP over expression up-regulates mitochondrial respiration rate. We hypothesize that SLC35A4 SEP could be stabilizing respiratory chain protein complex through protein-protein interaction, which leads to enhanced mitochondrial respiration. Further investigation needs to be followed to confirm this hypothesis.

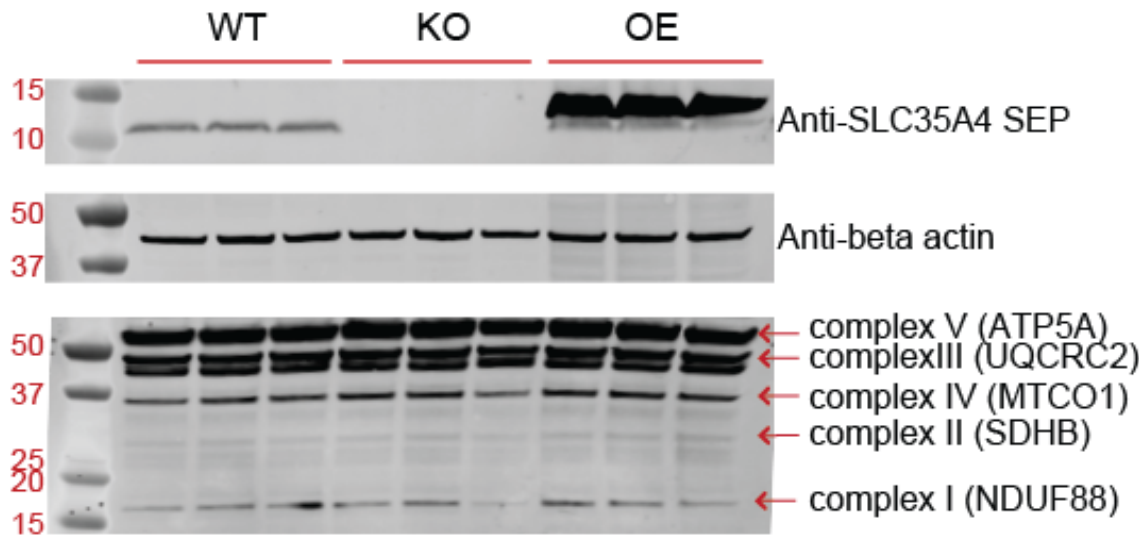


Figure 5.6. Western blot analysis of mitochondrial respiratory chain proteins. Same samples from Figure 5.4 were used to probe mitochondrial respiratory chain proteins. Total OXPHOS Rodent WB antibody cocktail (ab110413) was used to detect five proteins; each corresponds to one of the proteins in complex I-V of mitochondrial respiratory chain. No significant protein level change was detected by western blot among WT, KO and OE cell lines.

5.3. Conclusion

Despite large numbers of SEPs are detected, only a handful is characterized (1, 3, 5, 25-29). Interest in this field will only be driven if these molecules are functional. In flies, the recent description of the sarcolamban SEP with activity in heart rhythm was a terrific discovery that highlights a physiological role for SEPs(5). A mouse homolog of sarcolamban, myoregulin(28), was recently shown to control muscle contraction in mice, and a SEP from pancreatic mitochondrial DNA called MOTS-c has anti-diabetic activity in mice (29). To this growing list, we can now add SLC35A4 SEP, a conserved human SEP translated from nuclear DNA but localized in mitochondria, with a fundamental role in respiration. Our findings add to the conclusion that SEP biology is an emerging field with potential to shape our understanding of the functional proteome.

In summary, we have utilized functional proteomics and biochemical approach to characterize a 5'UTR SEP, SLC35A4 SEP in human cells. Its specific expression in mitochondria led us to hypothesize that the SEP's function might be related to its subcellular localization. Functional proteomics has demonstrated its value of elucidating SEP's protein-protein interactions. In this case, we identified 106 proteins that were enriched in SLC35A4 SEP over expressed co-immunoprecipitated samples. Out of which, 20 % was mitochondrial respiratory chain complex proteins. In order to investigate mitochondrial respiration and SLC35A4 SEP's effect on it, we used Seahorse XF96 extracellular flux analyzer to measure the oxygen consumption rate, an indicator of oxidative phosphorylation. Interestingly, mitochondrial basal and maximal respiration rate were the highest in the SLC35A4 SEP over expressing cells, followed by the wild type and the SEP knockout cells showed the lowest respiration rate. There is a close correlation between the SEP expression level and the rate of mitochondrial respiration.

This is the first evidence to date that a human SEP is possibly involved in regulation of the mitochondrial respiration. Further studies need to be conducted to confirm SLC35A4 SEP function in mitochondria and its direct binding to respiratory chain complex proteins.

5.4. Materials and Methods

5.4.1. Cell culture and transfection

HEK293T and HeLa cells were culture in DMEM supplemented with 10% fetal bovine serum. Cells were maintained in a 5% CO₂ atmosphere at 37°C. Plasmid transfection was performed with Lipofectamine 2000 and Opti-MEM according to the manufacturer's instructions. For immunofluorescence imaging, cells transfected with plasmids were assayed 24 hours post-transfection.

5.4.2. Peptidomics and SLC35A4 SEP identification

The peptidomics experiments in which SLC35A4 SEP was detected in K562, H293T, A549 and HeLa cells were previously reported (Chapter 2 and 3).

5.4.3. Immunofluorescence and confocal microscopy

HEK293T or HeLa cells were grown to 70% confluence on no. 1.5 glass coverslips in 48-well plates. Cells were transfected with SLC35A4 SEP-FLAG plasmid DNA according to the manufacturer's instructions. 24 hours post-transfection, cells were stained with 100 nM MitoTracker (Life Technologies) in growth media at 37 °C incubator for 15 min, followed by 3 PBS washes, Cells were then fixed with 4% formaldehyde in phosphate buffered saline, permeabilized with methanol at -20°C, and blocked with fluorescence blocking buffer

(Rockland) for at least 1 hour at 4°C. Cells were stained with anti-FLAG M2 (Sigma) primary antibodies at a 1:1000 dilution in Rockland blocking buffer overnight at 4 °C, followed by 3 phosphate-buffered saline (PBS) washes. Secondary antibody goat anti-mouse Alexa Fluor 488 were applied at a 1:1000 dilution in Rockland buffer for 1 hour either at room temperature or 4 °C, then washed 3 times with PBS. Cells were post-fixed with 4% formaldehyde in PBS and for experiments requiring cell counting, were subjected to nuclear counterstaining with Hoechst 33258 at a concentration of 0.02 mg/mL in PBS for 15 minutes, then imaged by confocal microscopy.

Coverslips were inverted and imaged in PBS in MatTek imaging dishes. Confocal imaging was performed at the Harvard Center for Biological Imaging on a Zeiss LSM 700 with 60 × oil immersion objective. Hoechst was imaged using 405 laser excitation, 445/40 emission; Alexa Fluor 488/GFP was imaged using 491 laser excitation, 528/38 emission; MitoTracker was imaged using 561 laser excitation, 617/73 emission. Image acquisition was performed with Zen software.

5.4.4. Co-immunoprecipitation and proteomics

FLAG-tagged SLC35A4 SEP in pcDNA3 (or empty pcDNA3 vector as a negative control) were transfected into HEK293T cells using 10 µg DNA per 10 cm dish of cells. Samples were prepared in biological triplicates for statistical significance. 24 hours post-transfection, cells were harvested and lysed using IP lysis buffer (Pierce) and Roche Complete protease inhibitor cocktail tablets. 800 µL lysis buffer was used per pellet. Cells were lysed on ice for 5 min followed by sonication in water bath for 5 min, then centrifugation at 20,000 g, 4 °C, 15 min. Lysate samples (20 µL) were saved for western blot analysis. A 100 µL aliquot of anti-FLAG

agarose beads (clone M2, Sigma) was washed with 1 mL TBS-T, collected by centrifugation for 1 min at 3000 rpm, then suspended in the cell lysate supernatant. Bead suspensions were rotated at 4 °C for 1 hour, then washed 3 times with TBS-T. Elution was in 50 µL of 3X FLAG peptide (Sigma), at a final concentration of 125 µg/mL in TBS-T at 4 °C for 1 hour. Beads were removed by centrifugation and the supernatant was collected for analysis on mass spectrometry.

Sample preparation for mass spectrometry: Co-immunoprecipitated samples were precipitated with methanol/chloroform. Air dry. To each dry pellet, add 12.5 µL of 8 M urea, and 1.5 µL of TCEP (5 mM final), and shake at 37 °C for 20 min. Then add 1.4 µL of 500 mM chloroacetamide (10 mM final), shake at 37 °C for additional 20 min in dark. Add 30.5 µL of 100 mM TEAB (Sigma), then 1 µL of 0.5 µg/µL trypsin, leave at 37 °C overnight. Samples were quenched with 5.25 µL of 90 % formic acid; spin down at maximum speed for 15 min on bench top. Transfer 50 µL of each sample to mass spec. sample vials. LC-MS/MS: Digests were analyzed by LC-MS using an Easy-nLC1000 (Proxeon) and a Q Exactive mass spectrometer (Thermo Scientific). Electrospray was performed directly from the tip of the analytical column. Buffer A was 5 % acetonitrile and 0.1 % formic acid; buffer B was 80 % acetonitrile and 0.1 % formic acid. Flow rate was 200 nl/min. Each sample was run in duplicates. The digest was loaded through the autosampler, venting to waste and desalting by use of the trap column. Peptide separation was performed in a 140min reverse phase gradient.

Data Analysis: Tandem mass spectra were extracted from raw files using RawExtract 1.9.9.2 and searched with ProLuCID against a UniProt human database appended with SEP sequence using Integrated Proteomics Pipeline – IP2 (Integrated Proteomics Applications). The search space included all fully-tryptic and half-tryptic peptide candidates. Carbamidomethylation on cysteine was considered as a static modification. Duplicate MS files were combined and searched. Data

was searched with 10-ppm precursor ion tolerance and 50-ppm fragment ion tolerance with maximum of two internal missed cleavages. Identified spectra were filtered and grouped into proteins using DTASelect. Proteins required a minimum of one peptides to be present and less than 1% FDR. Identification_STAT COMPARE tool in IP2 was used to compare control and SLC35A4 SEP over expressed dataset.

5.4.5. Western blot

Cell lysates were loaded on a Bolt 4-12 % BisTris gel, 10-well (Life Technologies) and run in MES running buffer at 200V for 20 min. Proteins were transferred to PVDF membrane using iBLOT 2 (Life Technologies) program “P0”, followed by blocking the membrane at room temperature for 1 hour. Then the membrane was blotted with primary antibody: rabbit anti-beta actin (LiCor), rabbit anti-TFRC (Cell Signaling), rabbit anti- SLC16A1 (MCT1) (Sigma), rabbit anti- mitofilin (Cell Signaling), rabbit anti- SLC25A3 (Cell Signaling), mouse anti- FLAG M2 (Sigma) at 1:1000 dilution for 1 hour at room temperature or at 4 °C overnight, and rabbit anti-SLC35A4 SEP (in-house generated) at 1:5000 dilution, rock at 4 °C overnight. For the detection of mitochondrial respiratory chain proteins, total OXPHOS rodent WB antibody cocktail (Abcam 110413) was used in 1:500 dilution at 4 °C overnight. Wash membrane three time with TBS-T, then blot with secondary antibody: goat anti-rabbit IRDye 800CW (LiCor) or goat anti-mouse IRDye 800CW (LiCor) at 1:10000 dilution, rock 1 hour at room temperature. Wash membrane three times with TBS-T then scan the membrane using LiCor Odyssey CLx at IR700 and IR800. Built-in tool in Odyssey CLx was used to quantify the intensity of the bands of interest.

5.4.6. Knockout of SLC35A4 SEP by CRISPR/Cas9-mediated genome editing

Two guide RNA (gRNA) sequences were designed to target SLC35A4 exon1 and exon 2. GGGGAAGATGGCGGATGACA targets the first exon of SLC35A4 near the SEP start codon and GCAGCGGCGTGTAGAAGACG targets the second exon. Two gRNAs were cloned into pSpCas9 BB-2A-Puro (PX459) vector and purchased from GenScript. Two gRNAs were co-transfected (1 µg each) with 8 µL of Lipofectamine 2000 (Life Technologies) in Opti-MEM to HEK293T cells growing in 6-well plate. 24 hour post-transfection, media was replaced with complete media supplemented with puromycin (1.25 µg/ml working conc.) for selection. Media containing puromycin was changed daily until non-transfected cells all died. The surviving transfected cells were propagated for future assays. Validation of efficiency of knockout was performed by western blot using in-house generated antibody against endogenous SLC35A4 SEP.

5.4.7. RNA profiling and Gene Set Enrichment Analysis

HEK293T total RNA was isolated using PureLink RNA Mini Kit (Life Technologies) according to the manufacture's instructions. On-column DNase I (NEB) treatment was performed to remove genomic DNA. Samples were prepared in biological triplicates. RNA integrity was checked using Bioanalyzer and all samples with RIN (RNA Integrity Number) greater than 8 (out of 10) were preceded with library construction. Stranded mRNA-Seq libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit according to the manufacturer's instructions (Illumina). Briefly, RNA with poly-A tail was isolated using magnetic beads conjugated to poly-T oligos. mRNA was then fragmented and reverse-transcribed into cDNA. dUTPs were incorporated, followed by second strand cDNA synthesis.

dUTP-incorporated second strand was not amplified. cDNA was then end-repaired, index adapter-ligated and PCR amplified. AMPure XP beads (Beckman Coulter) were used to purify nucleic acid after each steps of the prep. Libraries were then quantified, pooled and sequenced at single-end 50 base pair using the Illumina HiSeq 2500 platform at the Salk NGS Core. Raw sequencing data was demultiplexed and converted into FASTQ files using CASAVA (v1.8.2). Libraries were sequenced at an average depth of 15 million reads per library. Sequenced reads were quality tested using FASTQC and aligned to the human hg19 genome using the STAR aligner version 2.4.0k. Raw gene expression was quantified across all annotated exons and differential gene expression was carried out using the edgeR package v3.6.8. using representative duplicates to compute within-group dispersion. One of the three replicates did not cluster well with the other two from each sample, therefore was removed from gene expression analysis. Differentially expressed genes were defined as having an FDR < 0.05 and a log₂ fold change greater than 1. GO term and KEGG pathway enrichment analysis was carried out on differentially expressed genes using the HOMER analysis package and and the Benjamini and Yekutieli general correction for multiple testing.

5.4.8. Mitochondrial respiration assay

Seahorse XF-96 assay plate was coated with poly-lysine then washed three times with PBS prior to plating the cells. HEK293T cells were plated in a 96-well Seahorse XF-96 assay plate at 15,000 cells/well and grown in FBS- containing DMEM media for 24 hours. For over expressed (OE) samples, HEK293T cells were transfected with SLC35A4 SEP-FLAG plasmid DNA with Lipofectamine in Opti-MEM. Cartridge was calibrated in Seahorse calibrant in 37 °C incubator overnight. On the day of metabolic flux analysis, cells were changed to unbuffered

DMEM media (DMEM base medium supplemented with 25 mM glucose, 10 mM sodium pyruvate, 31 mM NaCl, 2 mM Gluta- Max, pH 7.4) and incubated at 37 °C in a non-CO₂ incubator for 1 hr. All media was adjusted to pH 7.4 on the day of assay. Three baseline measurements of OCR and ECAR were taken before sequential injection of mitochondrial inhibitors, oligomycin (1 μM), FCCP (0.5 μM) and rotenone/antimycin (1 μM each). Three measurements were taken after each addition of mitochondrial inhibitor before injection of the next inhibitor. Oxygen consumption rate (OCR, an indicator of oxidative phosphorylation) and extracellular acidification rate (ECAR, an indicator of glycolysis) were automatically calculated and recorded by the Seahorse XF-96 software (Seahorse Bioscience).

5.5. References

1. M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**, e106 (2007).
2. T. Kondo *et al.*, Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* **9**, 660-665 (2007).
3. T. Kondo *et al.*, Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329**, 336-339 (2010).
4. J. I. Pueyo, J. P. Couso, Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor. *Developmental biology* **355**, 183-193 (2011).
5. E. G. Magny *et al.*, Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341**, 1116-1120 (2013).

6. G. Menschaert *et al.*, Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & cellular proteomics : MCP* **12**, 1780-1790 (2013).
7. A. G. Schwaid *et al.*, Chemoproteomic discovery of cysteine-containing human short open reading frames. *Journal of the American Chemical Society* **135**, 16750-16753 (2013).
8. S. A. Slavoff *et al.*, Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature chemical biology* **9**, 59-64 (2013).
9. B. Vanderperre *et al.*, Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS one* **8**, e70698 (2013).
10. J. Ma *et al.*, The Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *Journal of proteome research*, (2014).
11. A. A. Bazzini *et al.*, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981-993 (2014).
12. N. T. Ingolia, Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol* **470**, 119-142 (2010).
13. N. T. Ingolia *et al.*, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**, 1365-1379 (2014).
14. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
15. S. J. Child, M. K. Miller, A. P. Geballe, Translational control by an upstream open reading frame in the HER-2/neu transcript. *The Journal of biological chemistry* **274**, 24335-24341 (1999).
16. A. G. Hinnebusch, Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome. *The Journal of biological chemistry* **272**, 21661-21664 (1997).

17. C. F. Calkhoven, C. Muller, A. Leutz, Translational control of C/EBPalpha and C/EBPbeta isoform expression. *Genes Dev* **14**, 1920-1932 (2000).
18. S. E. Calvo, D. J. Pagliarini, V. K. Mootha, Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* **106**, 7507-7512 (2009).
19. C. Akimoto *et al.*, Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochim Biophys Acta* **1830**, 2728-2738 (2013).
20. M. Iacono, F. Mignone, G. Pesole, uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* **349**, 97-105 (2005).
21. C. Jousse *et al.*, Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5'UTR. *Nucleic Acids Res* **29**, 4341-4351 (2001).
22. S. Lee *et al.*, Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **109**, E2424-2432 (2012).
23. F. A. Ran *et al.*, Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-2308 (2013).
24. M. D. Brand, D. G. Nicholls, Assessing mitochondrial dysfunction in cells. *Biochem J* **435**, 297-312 (2011).
25. B. Guo *et al.*, Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **423**, 456-461 (2003).
26. A. Pauli *et al.*, Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343**, 1248636 (2014).
27. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *The Journal of biological chemistry* **289**, 10950-10957 (2014).

28. D. M. Anderson *et al.*, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160**, 595-606 (2015).
29. C. Lee *et al.*, The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance. *Cell metabolism* **21**, 443-454 (2015).

Appendix Chapter 1

Peptidomic Discovery of Short Open Reading Frame-Encoded Peptides in Human Cells

This chapter was adapted from: Slavoff, SA.; Mitchell, AJ.*; Schwaid, AG.*; Cabili, MN; Ma, J.; Levin, JZ; Karger, AD.; Budnik, BA.; Rinn, JL; Saghatelian, A. Peptidomic Discovery of Short Open Reading Frame-Encoded Peptides in Human Cells. *Nature Chemical Biology* 2013 9(1), 59.

*Authors contributed equally

A 1.1. 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. 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). 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 (defined here as polypeptides that are synthesized on the ribosome at a length of less than 150 amino acids) between the K562 and HEK293 cell lines with a length distribution of 88-148 amino acids (19). The discordance between the small number of SEPs detected 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 novel SEPs, the largest number of human SEPs ever reported, which demonstrates that SEPs are much more abundant than 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 highlight SEPs as an interesting class of polypeptides within the human proteome.

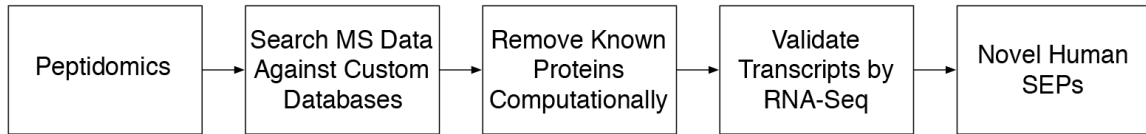
A 1.2. Results

A 1.2.1. Discovering SEPs Encoded by Annotated Transcripts

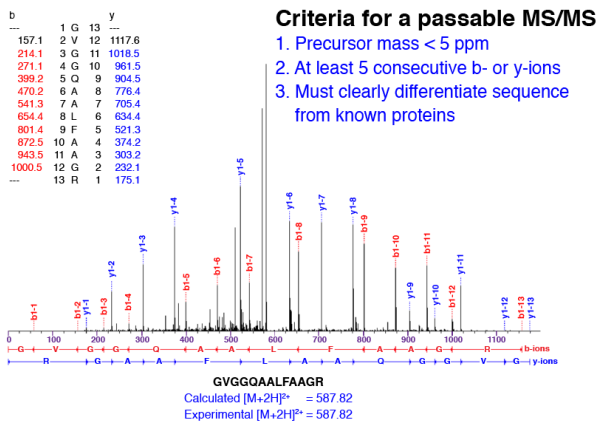
We developed a novel strategy that combines peptidomics and massively parallel RNA sequencing (RNA-seq) to discover human SEPs (Figure A1.1). Peptidomics augments the

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

A



B



C

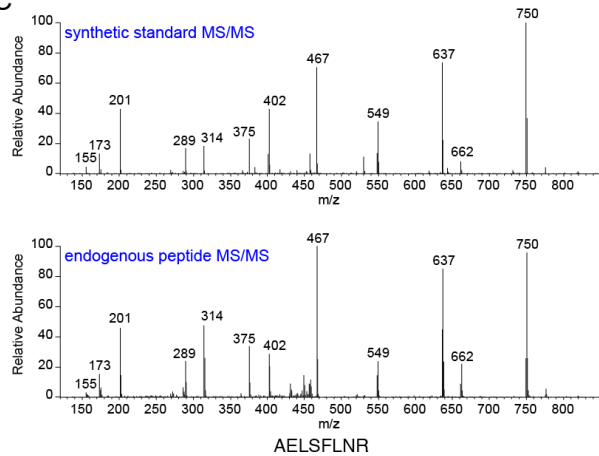


Figure A1.1. Discovering SEPs. (A) An LC-MS/MS-based peptidomics platform was used to profile K562 cells. The MS/MS data were searched against a custom protein database (RefSeq or RNA-seq) to identify polypeptides in K562 cells. Peptides shorter than 8 amino acids were discarded. Tryptic peptides that 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) MS/MS spectra for a SEP tryptic peptide. The MS/MS spectra for GVGQAALFAAGR was visually inspected to ensure sequence coverage that this peptide is unique to the sORF that encodes this SEP. (C) 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).

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 the 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) the optimal size for detection by LC-MS/MS is approximately 10-20 amino acids, indicating that SEPs detection would greatly benefit from trypsin proteolysis.

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) (Figure A1.1A). 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 eleven SEPs did have two or more

fragments (Table A1.1). This contrasts with standard proteomics studies, which, on account of the numerous tryptic fragments generated from full size polypeptides, will typically uncover two or more peptides to support the presence 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 (Figure A1.1B). 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 (Figure A1.1C).

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 (Figure A1.2 and Table A1.1). 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. Importantly, a similar analysis without trypsin failed to identify any SEPs demonstrating the importance of trypsin in generating an ideal sample for LC-MS/MS.

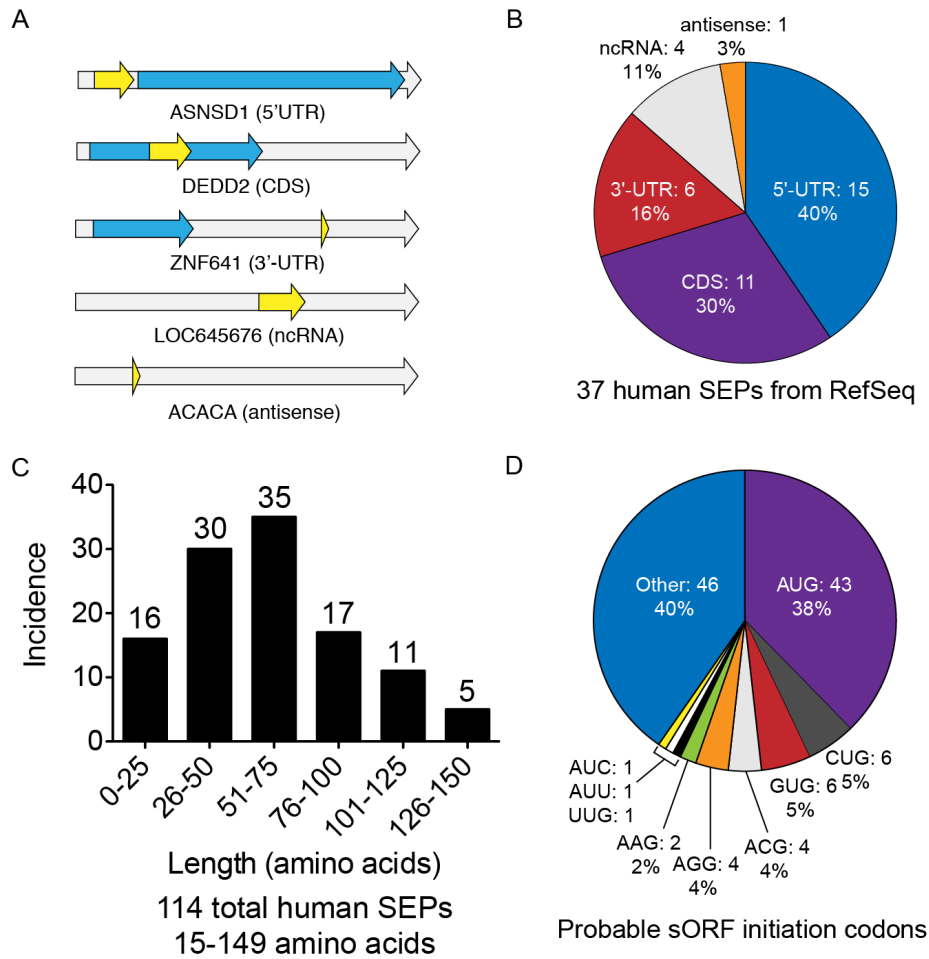


Figure A1.2. Overview of SEPs. (A) RNA maps illustrating the categories of sORFs that are translated into SEPs, including 5'UTR, CDS, 3'UTR, non-coding RNAs and antisense RNAs. The gray arrow represents the RNA, the blue arrow represents annotated protein CDS (if present), and the yellow arrow represents 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

Figure A1.2. (Continued) sequence (27). 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.)

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 (frameshifted) inside the main coding sequence (CDS), (iv) those located on non-coding RNAs (ncRNAs), and (v) those located on antisense transcripts (Figure A1.2A, 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 (Figure A1.2B). 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).

A 1.2.2. 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 (28) 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 (Figure A1.2C, Table A1.1). The average length of each 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 (Figure A1.1B). 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. Interestingly, 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 (29) (Figure A1.3).

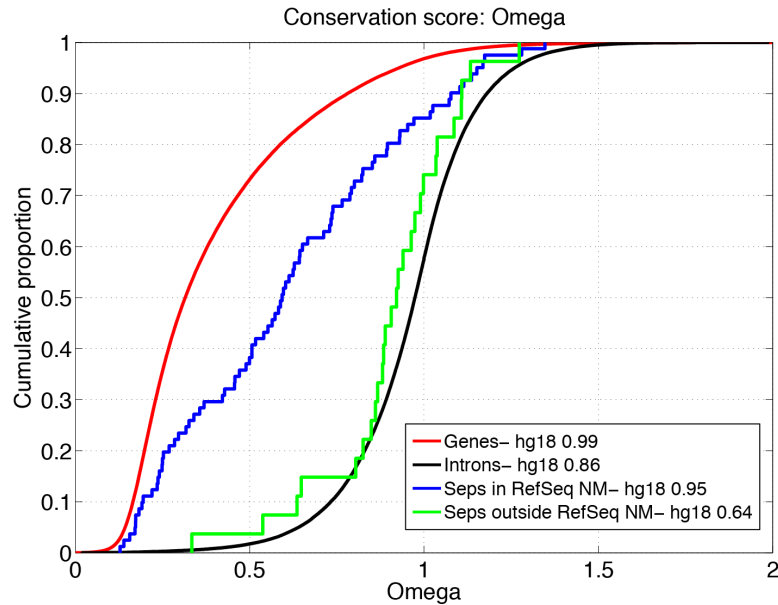


Figure A1.3. SEP-encoding sequences are under stronger evolutionary selection than the introns of known coding genes. The curves show the cumulative distribution of sequence conservation levels calculated by SiPhy(29) across 29 mammalian species (Omega) in the exons of protein coding genes (red), the 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.

A 1.2.3. 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 (27). 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 (Figure A1.2C). 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 (Figure A1.4). 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).

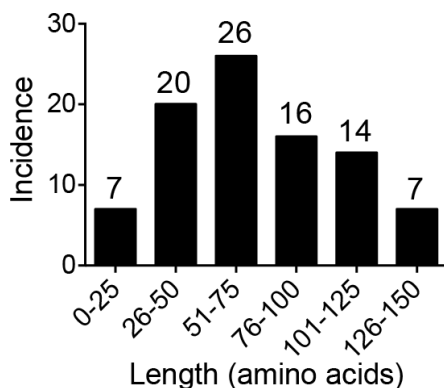


Figure A1.4. Length distribution for SEPs determined by defining sORF initiation sites the codon immediately 3' of the upstream stop codon unless an AUG was present, in which case the upstream-most AUG was defined as the start.

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 (Figure A1.2D). 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.

A 1.2.4. 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 decided to use 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 (Table A1.1). Importantly, for some of these SEPs we also find additional peptides from the SEP to provide even greater confidence in the SEP assignments.

We still needed to demonstrate that full-length SEPs are present in K562 lysates and therefore we elected to perform an isotope-dilution mass spectrometry (IDMS) experiment with chemically synthesized full-length SEPs. Specifically, we prepared two SEPs, **MLHSRKREL**R**QVLIT**N**KNQVLIT**N**KQVRLTLLTLG** and **MLRCFFPKMCFSTTIGGM**N**QRG**K**RK**, 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 (Figure A1.5). 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.

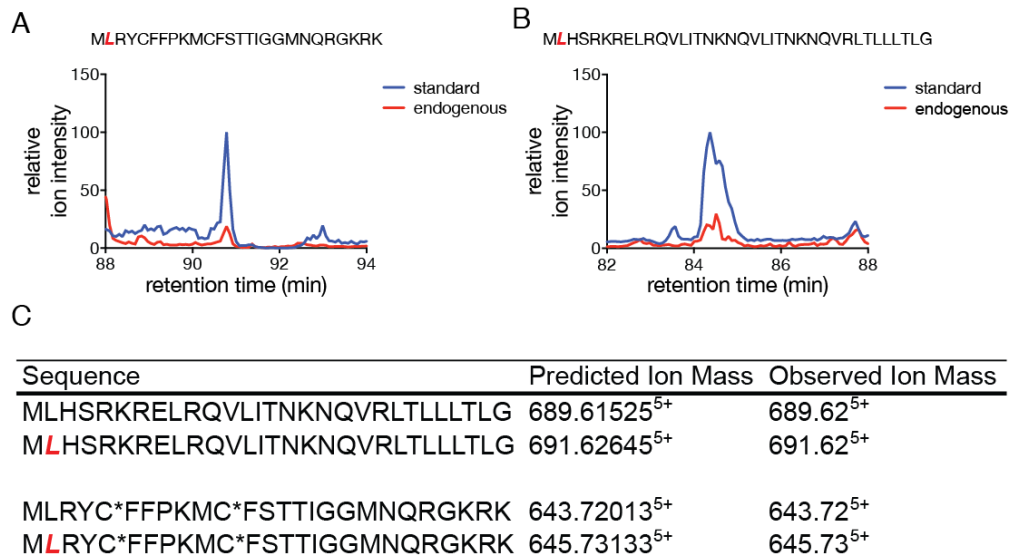


Figure A1.5. Confirmation of the presence of full-length SEPs in the K562 lysates by isotope-dilution mass spectrometry (IDMS). Full-length SEPs were synthesized by solid phase peptide synthesis and a deuterated leucine (d10-Leucine) was included to create a ‘heavy’-labeled SEP (red amino acid in sequences). Addition of these synthetic SEPs (blue lines) to K562 lysates enabled the identification of endogenous full-length SEPs (red lines) MLRYCFFPKMCFSTTIGGMNQRGKRK (A) and MLHSRKRELQRQLITNKNQVRLTLLLTLG (B). The CID for these spectra was uninterrupted 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.

A 1.2.5. 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)(Figure A1.6A). (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 (Table A1.2). 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 (Figure A1.6B).

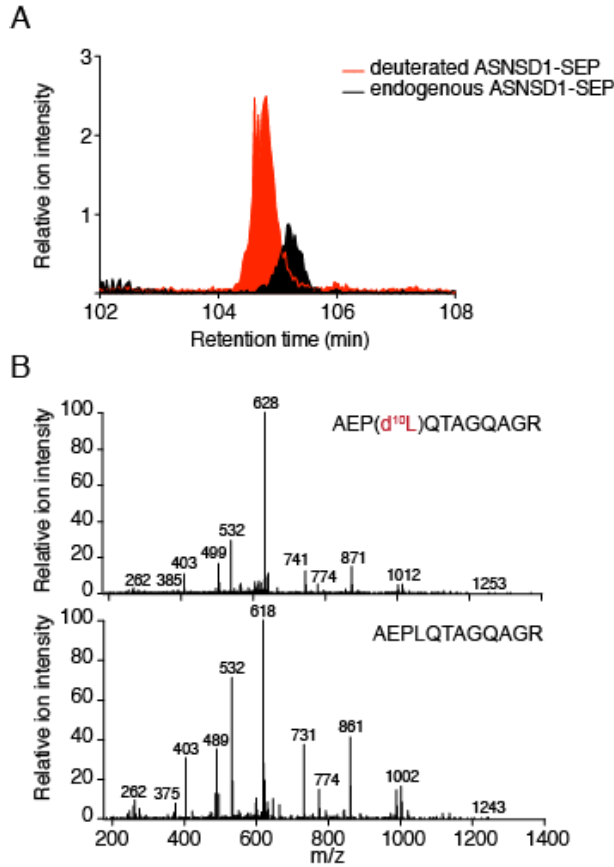


Figure A1.6. SEP quantitation. (A) SEPs were quantified by isotope dilution mass spectrometry (IDMS). We synthesized a deuterated (heavy-labeled) variant of the diagnostic SEP peptide we detected. Upon isolation of K562 cells this peptide was added and the entire mixture was prepared using our standard 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.

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

Table A1.2. Quantification of SEP trypsin peptides by IDMS

A 1.2.6. 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 (Figure A1.7). 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 (Figure A1.7A).

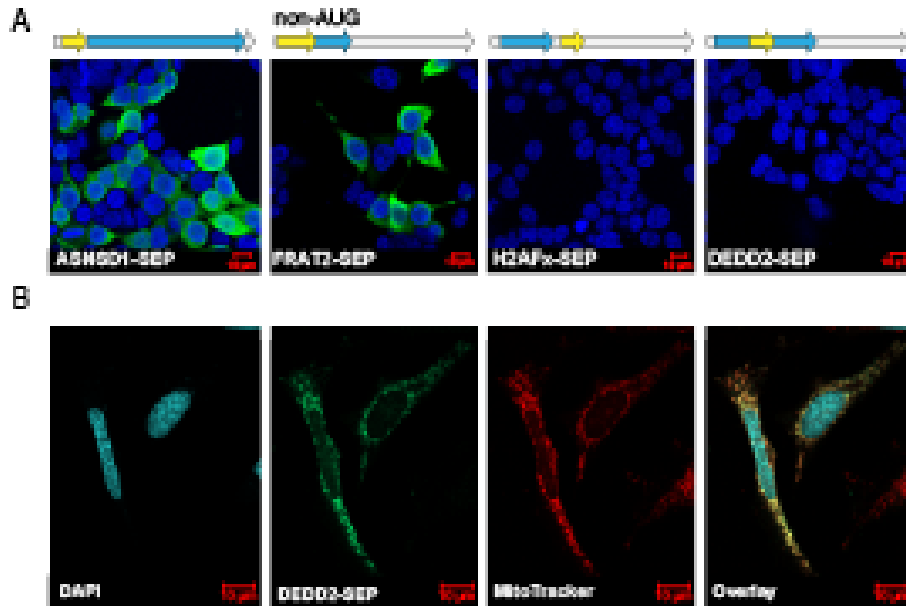


Figure A1.7. Expression of SEPs. (A) Transient transfection of HEK293T cells with constructs containing a cDNA sequence corresponding to the full-length RefSeq mRNA (i.e., including the 5'- and 3'-UTRs). We appended a C-terminal FLAG-tag on the SEP coding sequence that could be detected by immunofluorescence. 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 its 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.)

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 (Figure A1.7A). 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

(Figure A1.8). 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.

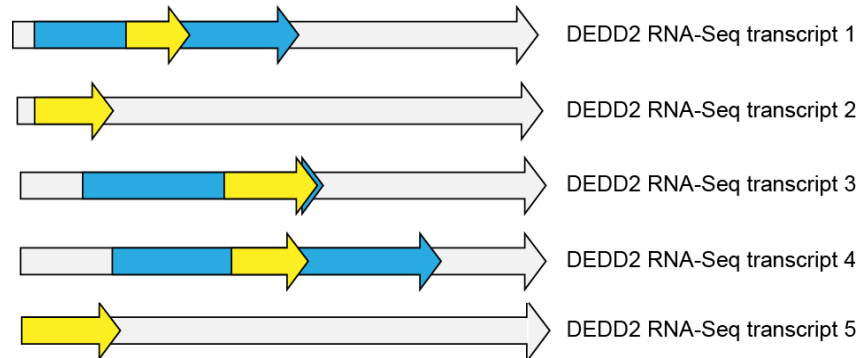


Figure A1.8. All DEDD2 RNAs detected in the K562 RNA-Seq data. DEDD2 RNA-Seq transcript 2 encodes the DEDD2-SEP. (Note: RNAs not to scale.)

A 1.2.7. 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 (Figure A1.9). 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 (Figure A1.7B). 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.

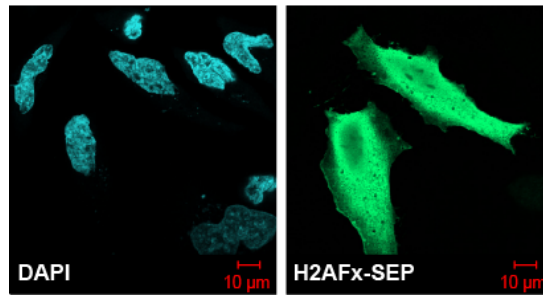


Figure A1.9. *H2AFx-SEP-FLAG* sORF expressed in HeLa cells, then detected with anti-FLAG antibody (followed by anti-mouse AlexaFluor 488, green).

A 1.2.8. 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 (Figure A1.10) was consistent

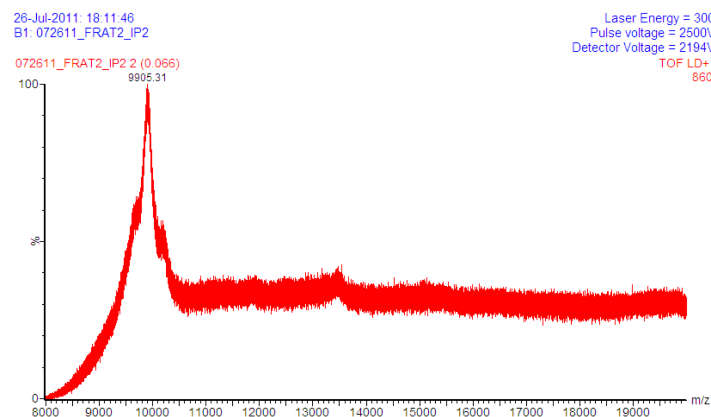


Figure A1.10. MALDI-MS of immunoprecipitated FRAT2-SEP-FLAG provides a polypeptide with a molecular weight of 9905, which identifies an ACG initiation codon with methionine as the first amino acid.

with initiation at an ACG triplet embedded within a Kozak consensus sequence (27) (Figure A1.11). Mutating the ACG to an ATG resulted in increased FRAT2-SEP translation while

FRAT2-SEP

```
      M G G H A V P E G G G R G S R R G G
gccagggACGgggggccaatgccgtgccggaggaggagggaagaggaagccggcgaggaggc

      G G G G R G G R Q L P P A A A V G D A G
ggagggggagggaagaggaggacgacagcttcctcctgctgcagcagtcgggtgacgctggg

      Q L G R G G P A G G P D R R D A A A G R
cagctcgggcgagggtggaccggctggtggcccagatcggcgagacgctgcagctggacgc

      G A G Q P G L A V R A P G G A A A G P G
ggcgaggacagccccggcctcgcctgctgcgcgccccgggggtgcccgtgcgggccccggg

      A P G C G G A D G Q G P A P G G A A A A
gccccctggctgeggcggtgccgacggacaaggccccggccccggcggtgcccgtgctgct

      A A R F G *
gccgcccgcttcgggctag
```

Figure A1.11. 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.

deletion of this ACG abolished FRAT2-SEP production, as assessed by Western blotting, thus confirming our assignment (Figure A1.12A). 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 (Figure A1.12B), 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 (Figure A1.12B). Remarkably, mutation of the ACG start codon of the FRAT2-SEP to an ATG increases FRAT2-SEP

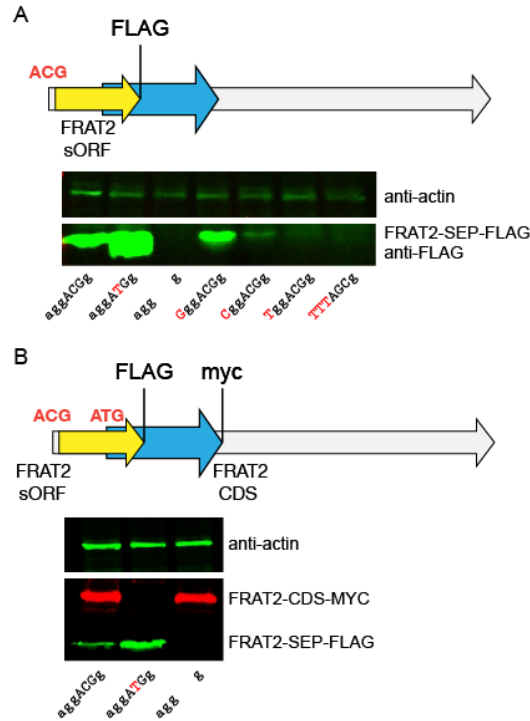


Figure A.12. Characterization of the non-AUG initiation codon of the FRAT2-SEP sORF. (A) An ACG was confirmed as the FRAT2-SEP initiation codon by site-directed mutagenesis followed by western blots of FRAT2-SEP-FLAG using an anti-FLAG antibody. Conversion of the ACG to an ATG resulted in higher expression (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 (27) 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.

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 1.2.9. 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 (Table A1.1), which represents just 0.4% of the 1866 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.

A 1.3. 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.

Coordinates of sORF	Peptides Detected by LC-MS from sORF	Probable Start Codon	Estimated SEP Length From Probable Start or Upstream Stop Codon to the Stop Codon at the 3'-end of sORF
chr9: 139256352-139264369 strand=-	AAPGALPEAAVGR (i)	ATG	96
chrX: 16859470-16888534 strand=-	AGAPAVGLLLANER	GTG	39
chr10: 99092201-99094454 strand=-	QLPPAAAVGDAGQLGR (i) APGGAAAGPGAPGCGGAGGQGPAPGGAAAAAR	ACG	103
chr9: 139557366-139565706 strand=+	ATPGLQQHQPPGPGR ATPPGGTGHELSGGAADVASVGSGR ATPPGGTGHELSGGAADVASVGSGR GMTDSEPPPHPEEK HRWPPPPGGAAPAPVR	ATG	83
chr2: 190526195-190535440 strand=+	IVVDELSNLK (i) QQQNSNIFFLADR NILDELKK EYQEIENLDK (idms/tryp) IVVDELSNLKK QQQNSNIFFLADR EYQEIENLDKTK	ATG	96
chr7: 150646657-150675423 strand=-	TAPSSATTASASCAATR	ATG	62
chr9: 123612077-123639492 strand=-	LQVGPADTQPR (i, idms/tryp) LQVGPADTQPR	ATG	88
chr14: 103800538-103809402 strand=+	STAACQTSSIATR (i)	ATG	97
chr16: 89574827-89607413 strand=+	GSSAAVGR	Other	78
chr8: 64080459-64125260 strand=+	TAAAAAAGTITRPR (i)	GTG	102
chr8: 144897399-144897840 strand=-	GVGQAALFAAGR AGDPLPLQPQGGAAAR AAQAFPPAAELAQAGPER AGDPLPLQPQGGAAAR AAQAFPPAAELAQAGPER GVGQAALFAAGR AAHPHHAQVHPAVALQPAR	GTG	88
chr10: 98288128-98346562 strand=-	AVAAAAAADPDGGR AVAAAAAADPDGGR GCESAAAEAAAAEAAAGGVGEPAGRR	ACG	91
chr4: 122737616-122745077 strand=-	GGLGAASIAADGAPR (i) GGLGAASIAADGAPR	CTG	115
chr7: 100464771-100471014 strand=+	SSTPAPPQQFLPSSI (i)	ACG	74
chr11: 65686750-65689023 strand=+	VAVEEGLPGDVAER (i) DAEQEEEVQR	ACG	107
chr10: 101992055-102005758 strand=-	EGSVHPQVE	ATG	87
chr5: 180650039-180662529 strand=+	GAIGGGGAGVGGTAGAR	ATG	143
chr19: 42713286-42721897 strand=-	VAAVAVGSQAVLQILSR	ATG	77
chr19: 12949331-12969791 strand=+	WTSSTSSPNTSGAPR	ATG	77
chr1: 150522391-150532570 strand=+	NPPLVQDTVSGK	ATG	111
chr3: 193363602-193386115 strand=+	QTAFGKWYESLLNRR	Other	63
chr19: 13059508-13067950 strand=-	AVAGAAAGAGGR	ATG	73
chr7: 100032962-100034242 strand=-	AEEQPGLGPGAAGR (i)	ATG	149
chr1: 160061156-160064154 strand=-	RAVPAQQLLQSTPTCMPWTP	ATG	54
chr22: 41740383-41756157 strand=+	NTTQESLEKGP (synthesized to confirm)	Other	32
chr4: 169908762-169911558 strand=-	EALNEFLTR	Other	22
chr11: 118964597-118966163 strand=-	AEPLQTAGQAGR (idms/tryp)	ATG	59
chr3: 124945640-125042272 strand=-	AGNLILLQ	Other	23
chr12: 48732236-48745011 strand=-	STTIGGMNQR (idms/full-length SEP)	ATG	26

Table A1.1.

chr2: 131130309-131132956	strand=+	ERPANSLIDQCSQR	ATG	54
chr6: 80194734-80199064	strand=-	VFFKNLLAFAR	Other	22
chr16: 86563805-86589025	strand=-	AELSFLNR (synthesized to confirm)	ATG	70 (non-coding, RefSeq)
chr22: 21368073-21368526	strand=-	LLPLGASPAGVGGGLAPPR LLPLGASPAGVGGGLAPPR QGPKADSDSDLETEGAR ADSDSDLETEGAR	ATG	85 (non-coding, RefSeq)
chr7: 158799724-158814542	strand=+	TWLPSCEDLTLPGGR	ATG	50 (non-coding, RefSeq)
chr1: 155532795-155708399	strand=+	FLPVDSLRLR (idms/tryp)	ATG	90 (non-coding, RefSeq)
chr17: 35441928-35444379	strand=+	SLSSYGACSR	Other	71
chr19: 3610043-3626771	strand=-	GPSGTQEMGLSR GADGGGAGSAGQIQR	ATG	102
chr22: 47048295-47073068	strand=+	APEPGAVLAPAEVVL	AGG	56 (non-coding, RNA-Seq)
chr5: 34914296-34925392	strand=+	LLVSGSPAETLPLR	ATG	128 (non-coding, RNA-Seq)
chr22: 17092426-17095991	strand=+	ALAQGLTPSQIYSA	AAG	52 (non-coding, RNA-Seq)
chr19: 54693858-54697432	strand=+	LSAPQPGDILQAPAR	GTG	89 (non-coding, RNA-Seq)
chr15: 55609385-55613829	strand=-	VYIFQPVFEQYAK	ATG	54 (non-coding, RNA-Seq)
chr12: 118649944-118650075	strand=+	NEQTELLYNK	Other	18 (non-coding, RNA-Seq)
chr2: 85132483-85133801	strand=+	ILEDFLPPSSSRPQS	Other	42 (non-coding, RNA-Seq)
chr9: 130209955-130216851	strand=-	DLPGVAPPRPSLSLSPG	ATG	65 (non-coding, RNA-Seq)
chr5: 14664778-14699800	strand=+	AAASGQPRPEMQPAEQTEIK	ATG	58 (non-coding, RNA-Seq)
chr7: 150778180-150780257	strand=-	AQHGVHSNTASPLPAGAPR	AGG	66 (non-coding, RNA-Seq)
chr22: 32014633-32026837	strand=-	KQGGFVQVSANAL SETALLALDRPLLPALR	ATG	136 (non-coding, RNA-Seq)
chr2: 200322928-200323580	strand=+	LNINQSIIVSTATQR	AGG	55 (non-coding, RNA-Seq)
chr19: 56165091-56185542	strand=+	LPGQATTQQTDFQR	Other	54 (non-coding, RNA-Seq)
chr1: 7863564-7864928	strand=-	LVSAVLAGKE	CTG	43 (non-coding, RNA-Seq)
chr7: 100169852-100183655	strand=-	PAVAATLHLPAPEGPH	ATG	49 (non-coding, RNA-Seq)
chr1: 228544743-228549628	strand=+	QELIGASLHTAR	Other	119 (non-coding, RNA-Seq)
chr10: 11925853-11937442	strand=	THLGTGQCCLPGAGGPAR	Other	100 (LINC, RNA-Seq)
chr19: 55737961-55770381	strand=-	TSDAPRPSATPPGADPLNSAGPGAR	Other	103 (non-coding, RNA-Seq)
chr12: 12966292-12982891	strand=+	VTSWDGQNPPR	ATG	50 (non-coding, RefSeq)
chr2: 231577583-231685792	strand=+	AAPGPTAAAAQASAAAR	CTG	108 (non-coding, RNA-Seq)
chrX: 5214450-5216144	strand=+	RLLIPEEK	Other	45 (non-coding, RNA-Seq)
chr1: 175913973-176153786	strand=-	SPTTDSYGIPQGCK	Other	40 (non-coding, RNA-Seq)
chr3: 88101102-88108113	strand=-	DYILSLEMFSILLWG	Other	33 (non-coding, RNA-Seq)
chr7: 66386236-66423532	strand=+	HGHSFPDPGLLLQNQGD HGHSFPDPGLLLQNQGD GGADQNNVQHQPPEGEVHQQSASPGGLHDQR	Other	122 (non-coding, RNA-Seq)
chr16: 87435666-87438903	strand=-	HDASSPLGPPR	Other	55 (non-coding, RNA-Seq)
chr9: 130128866-130129660	strand=+	CLVYVLDLITDACTIKPLFNK	Other	43 (non-coding, RNA-Seq)
chr1: 16905808-16970994	strand=-	ASPGEAGPAGGAAAGQGAPR	Other	73 (non-coding, RNA-Seq)
chr17: 62205639-62207524	strand=-	GAWGGQLATAGSGPGQR	ATG	70 (non-coding, RefSeq)
chr1: 4036227-4073316	strand=+	DTEVLINTMSK	ATT	27 (LINC, RNA-Seq)
chr1: 157243513-157253900	strand=+	VYKWLNCNVE	ATG	41 (non-coding, RNA-Seq)
chr10: 119806332-119859641	strand=+	CPFVLLMSSMILLR	Other	33 (non-coding, RNA-Seq)

Table A1.1. (Continued)

chr11: 3532972-3542051 strand=+	KPVFLLLLSIR	Other	32 (non-coding, RNA-Seq)
chr11: 82783129-82805398 strand=+	FIPTEAWYSAGR	ATG	86 (non-coding, RNA-Seq)
chr11: 65266565-65274602 strand=-	QVLITNKQ (idms/full-length SEP)	ATG	29 (non-coding, RNA-Seq)
chr16: 3054772-3058645 strand=+	IKFLLAPEENK	ATG	43 (LINC, RNA-Seq)
chr19: 23278060-23286908 strand=+	QRIPCVVILTK	Other	73 (LINC, RNA-Seq)
chr2: 107137814-107160732 strand=+	KTLPMMGMIR	Other	30 (LINC, RNA-Seq)
chr3: 107852804-107857456 strand=+	QVNEETLK	Other	143 (LINC, RNA-Seq)
chr4: 10069715-10074643 strand=-	KNLFQNTSR	Other	59 (LINC, RNA-Seq)
chr7: 96251318-96293650 strand=-	RAGYSELE	ATG	69 (LINC, RNA-Seq)
chr15: 31008518-31061502 strand=+	QMSSNILK	Other	50 (non-coding, RefSeq)
chr21: 35345400-35353552 strand=+	VAHENYMKFK	Other	59 (non-coding, RNA-Seq)
chr6: 68590370-68642035 strand=+	GIALGDIPNAR	GTG	18 (non-coding, RNA-Seq)
chr6: 141167131-141219546 strand=-	VLLDQHQR	Other	23 (non-coding, RNA-Seq)
chr15: 59060273-59063173 strand=-	YYELQRGTR	AAG	43 (non-coding, RNA-Seq)
chr17: 41373439-41383338 strand=-	GEMERGEIK	ATG	18 (non-coding, RNA-Seq)
chr19: 23441500-23457032 strand=-	CQDILEAGKR	ATC	70 (non-coding, RefSeq)
chr2: 23598100-23604170 strand=-	DLGSPMLK	ATG	52 (non-coding, RNA-Seq)
chr2: 66653867-66660602 strand=-	TASPYSRPE	ATG	58 (non-coding, RNA-Seq)
chr20: 4173737-4176599 strand=+	LTVAGQGR	ATG	66 (non-coding, RNA-Seq)
chrX: 118425492-118469573 strand=+	SPFWAGQGQSR	Other	101 (non-coding, RNA-Seq)
chrX: 1515320-1517852 strand=-	NLAGGSLIP	Other	41 (non-coding, RNA-Seq)
chr21: 35303432-35308177 strand=+	AAALQFDLR	Other	23 (non-coding, RNA-Seq)

Table A1.1. (Continued) 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/trypp), 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 shown in red. SEPs from non-coding RNAs are in the last column and the database used for their identification is also included.

A 1.4. Methods

A 1.4.1. 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.

A 1.4.2. Cell culture

Cells were grown at 37°C under an atmosphere of 5% CO₂. 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 x 10⁵ cells/mL in RPMI1640 media with 10% fetal bovine serum, penicillin and streptomycin.

A 1.4.3. Isolation and processing of polypeptides

Aliquots of 3 x 10⁷ 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 Converter). The cell lysate was then brought to 0.25% acetic acid by volume and centrifuged at 20,000 x 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 μ l 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 μ l 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 $\mu\text{g}/\mu\text{l}$ 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, 1cm³; 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.

A 1.4.4. 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 x 2.1 mm, 5 μ m, 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 μ l 0.1% formic acid (aq.) in preparation for LC-MS/MS analysis.

A 1.4.5. LC-MS/MS analysis

Samples were injected onto a NanoAcquity HPLC system (Waters) equipped with a 5 cm x 100 μ m capillary trapping column (New Objective) packed with 5 μ m C18 AQUA beads (Waters) and a PicoFrit SELF/P analytical column (15 μ m tip, 25 cm length, New Objective) packed with 3 μ m 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.

A 1.4.6. 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.

A 1.4.7. 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(28). The transcriptome data is deposited on GEO (GSE34740).

A 1.4.8. 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% H₂O, 1% acetonitrile, and 0.1% TFA; buffer B was 90% acetonitrile, 10% H₂O, 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).

A 1.4.9. SEP analysis by PAGE

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.

A 1.4.10. 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 9x10⁷ 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.

A 1.4.11. 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 (10⁷ 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.

A 1.4.12. 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.

A 1.4.13. Determinations 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 x 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 μ L of 100 μ g/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, 1cm³; 30mg, Oasis) and eluted in 50% acetonitrile. The sample was dried in a SpeedVac, and then dissolved in a final volume of 10 μ L mass spectrometry-grade water (Burdick & Jackson). This solution (1 μ L) was mixed with matrix (α -cyano-4-hydroxycinnamic acid in 50% acetonitrile, 1 μ L) 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.

A 1.4.14. 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 μ g plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. Cells were harvested by vigorous pipetting and lysed in 100 μ L Triton lysis buffer. Samples of clarified lysate (20 μ L) 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 μ m 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.

A 1.4.15. Annotation of SEPs in Table A1.1

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 Table A1.1. 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 Table A1.1 these peptides are annotated by (synthesized to confirm). In addition, we also synthesized isotopically labeled tryptic peptides (deuterated peptides) 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 Table A1.1.

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.

A 1.5. References

1. M. C. Frith *et al.*, The abundance of short proteins in the mammalian proteome. *PLoS Genet* **2**, e52 (2006).
2. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes. *Cell* **147**, 789-802 (2011).
3. F. Zhang, A. G. Hinnebusch, 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. S. E. Calvo, D. J. Pagliarini, V. K. Mootha, Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* **106**, 7507-7512 (2009).
5. J. P. Abastado, P. F. Miller, A. G. Hinnebusch, A quantitative model for translational control of the GCN4 gene of *Saccharomyces cerevisiae*. *New Biol* **3**, 511-524 (1991).
6. M. Kozak, Bifunctional messenger RNAs in eukaryotes. *Cell* **47**, 481-483 (1986).
7. A. L. Parola, B. K. Kobilka, The peptide product of a 5' leader cistron in the beta 2 adrenergic receptor mRNA inhibits receptor synthesis. *J Biol Chem* **269**, 4497-4505 (1994).
8. M. Werner, A. Feller, F. Messenguy, The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression. *Cell*, (1987).
9. C. S. Wadler, C. K. Vanderpool, 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* **104**, 20454-20459 (2007).
10. G. Jay, S. Nomura, C. W. Anderson, G. Khoury, Identification of the SV40 agnogene product: a DNA binding protein. (1981).
11. S. A. Casson *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. H. Rohrig, J. Schmidt, E. Miklashevichs, J. Schell, M. John, Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc Natl Acad Sci U S A* **99**, 1915-1920 (2002).

13. J. P. Kastenmayer *et al.*, Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* **16**, 365-373 (2006).
14. C. A. Gleason, Q. L. Liu, V. M. Williamson, 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. M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**, e106 (2007).
16. T. Kondo *et al.*, Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* **9**, 660-665 (2007).
17. Y. Hashimoto *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* **98**, 6336-6341 (2001).
18. M. R. Hemm, B. J. Paul, T. D. Schneider, G. Storz, K. E. Rudd, Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular Microbiology* **70**, 1487-1501 (2008).
19. M. Oyama *et al.*, Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol Cell Proteomics* **6**, 1000-1006 (2007).
20. A. D. Tinoco, D. M. Tagore, A. Saghatelian, Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform. *J Am Chem Soc* **132**, 3819-3830 (2010).
21. M. Svensson, K. Skold, P. Svenningsson, P. E. Andren, Peptidomics-based discovery of novel neuropeptides. *J Proteome Res* **2**, 213-219 (2003).
22. D. M. Tagore *et al.*, Peptidase substrates via global peptide profiling. *Nat Chem Biol* **5**, 23-25 (2009).
23. D. L. Swaney, C. D. Wenger, J. J. Coon, Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res* **9**, 1323-1329 (2010).
24. J. K. Eng, A. L. McCormack, J. R. Yates 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* **5**, 976-989 (1994).

25. J. R. Yates, 3rd, J. K. Eng, A. L. McCormack, D. Schieltz, Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **67**, 1426-1436 (1995).
26. H. R. Christofk, M. G. Vander Heiden, N. Wu, J. M. Asara, L. C. Cantley, Pyruvate kinase M2 is a phosphotyrosine-binding protein. *Nature* **452**, 181-186 (2008).
27. M. Kozak, Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283-292 (1986).
28. M. N. Cabili *et al.*, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-1927 (2011).
29. M. Garber *et al.*, Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).
30. M. M. Dix, G. M. Simon, B. F. Cravatt, Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell* **134**, 679-691 (2008).
31. J. C. Tran *et al.*, Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254-258 (2011).
32. R. D. Kersten *et al.*, A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat Chem Biol* **7**, 794-802 (2011).
33. H. Keshishian, T. Addona, M. Burgess, E. Kuhn, S. A. Carr, 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. L. M. de Godoy *et al.*, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251-1254 (2008).
35. B. Schwanhauser *et al.*, Global quantification of mammalian gene expression control. *Nature* **473**, 337-342 (2011).
36. M. Beck *et al.*, The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549 (2011).
37. K. D. Pruitt, T. Tatusova, D. R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65 (2007).
38. A. G. Hinnebusch, Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol Mol Biol Rev* **75**, 434-467, first page of table of contents (2011).

39. J. D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795 (2004).
40. J. E. Wedekind, G. S. Dance, M. P. Sowden, H. C. Smith, Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends in genetics : TIG* **19**, 207-216 (2003).
41. M. Guttman *et al.*, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).
42. T. R. Mercer, M. E. Dinger, J. S. Mattick, Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155-159 (2009).
43. M. Guttman *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. A. M. Khalil *et al.*, Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).
45. B. R. Fonslow *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).
46. A. J. Alpert, Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. *Anal Chem* **80**, 62-76 (2008).
47. P. Hao *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. C. Trapnell *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. J. Z. Levin *et al.*, Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709-715 (2010).
50. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

Appendix Chapter 2

Chemoproteomic Discovery of Cysteine Containing Human sORFs

This chapter was adapted from: Schwaid, A.G.*; Shannon, D.A.*; Ma, J.; Slavoff, S.A.; Levin, J.Z.; Weerapana, E.; Saghatelian, A. Chemoproteomic Discovery of Cysteine Containing human sORFs. *J Am Chem Soc.* 2014 Jan 15;136(2):820

*authors contributed equally

A 2.1. Introduction

Short open reading frame (sORFs)–encoded polypeptides (SEPs) are an emerging class of biomolecules that are comprised of peptides and small proteins from sORFs (defined here as < 150 codons) (1). The existence of these molecules is of interest because they appear to be present in a variety of different cells (1, 2) and organisms (3, 4) but are missed by traditional gene finding algorithms (5). The discovery of these molecules has already revealed a great deal about protein translation in cells (1, 2, 6, 7). Ribosome profiling (2) and mass spectrometry discovery of sORFs (1, 2, 7), for example, revealed the prevalent use of non-ATG start codons.

Genetic screens have also identified several bioactive protein-producing sORFs (4). The search for genes that prevent cell death, for instance, led to the discovery of a 75-bp sORF that inhibits apoptosis of neuronal cells. It was shown that this sORF produces a 24-amino acid peptide (4) called humanin that binds and inhibits BAX (8), revealing a new endogenous molecule with a role in cell death. The complete extent of SEPs in the human genome is unknown and therefore there may be additional bioactive peptides and small proteins awaiting discovery.

SEPs are difficult to predict with traditional gene annotation algorithms due to their small size (3). Additionally, SEPs have been shown to violate several canonical rules of protein translation. They often initiate with non-ATG start codons and some have been shown to be bicistronic (1, 2). The recent discovery of this hidden proteome by ribosome profiling (2) and mass spectrometry (1) has generated intense interest towards identifying additional SEPs.

In order to identify additional SEPs, and also to discover SEPs that have properties similar to functional proteins, making them more likely to be functional, we applied a cysteine affinity enrichment approach to identify novel cysteine containing SEPs (ccSEPs). Reactive cysteines play a variety of critical roles in protein structure and function. In particular, cysteines

are important catalytic residues in the active site of many enzymes (9). Furthermore, cysteine oxidation to sulfenic, sulfinic, and sulfonic acid in addition to S-nitrosylation are important post translational modifications (10). For example, S-nitrosylation on histone deacetylase 2 (HDAC2) was found to induce chromatin remodeling in neurons (11). Lastly, cysteines are important metal chelators and are found in the metal binding site of many metalloproteins. The incorporation of metal ions in metalloproteins is important for metalloprotein folding and also stabilizes metalloprotein secondary structure (12-14). The ability of metal binding cysteines to stabilize the secondary structure of proteins is particularly interesting in the case of SEPs. Short proteins are intrinsically more disordered so SEPs that contain metal binding cysteines are more likely to be structured and consequently more likely to be functional (15, 16). In addition to selecting for cysteines that may be amenable to further functional characterization, by using a different strategy to enrich the peptidome we anticipate the discovery of novel ccSEPs.

A 2.2. Results and Discussion

A 2.2.1. Isolation of Cysteine Containing SEPs

Our strategy began with isolating the peptidome from K-562 cells, a human leukemia cell line, by lysis of these cells followed by a molecular weight cutoff (MWCO) filter to remove proteins larger than 30 kDa (Figure A2.1)(1). We incubated the peptidome with a previously described iodoacetamide-alkyne (IA-alkyne) probe (17, 18) that reacts with the sulfhydryl side chain of cysteine to form a covalent bond to the peptide. Notably, when used at 100 μ M concentrations the IA-alkyne probe will only label reactive cysteines (18). After cysteine capture by IA-alkyne, the probe is conjugated to a biotin-labeled tobacco etch virus (TEV) recognition

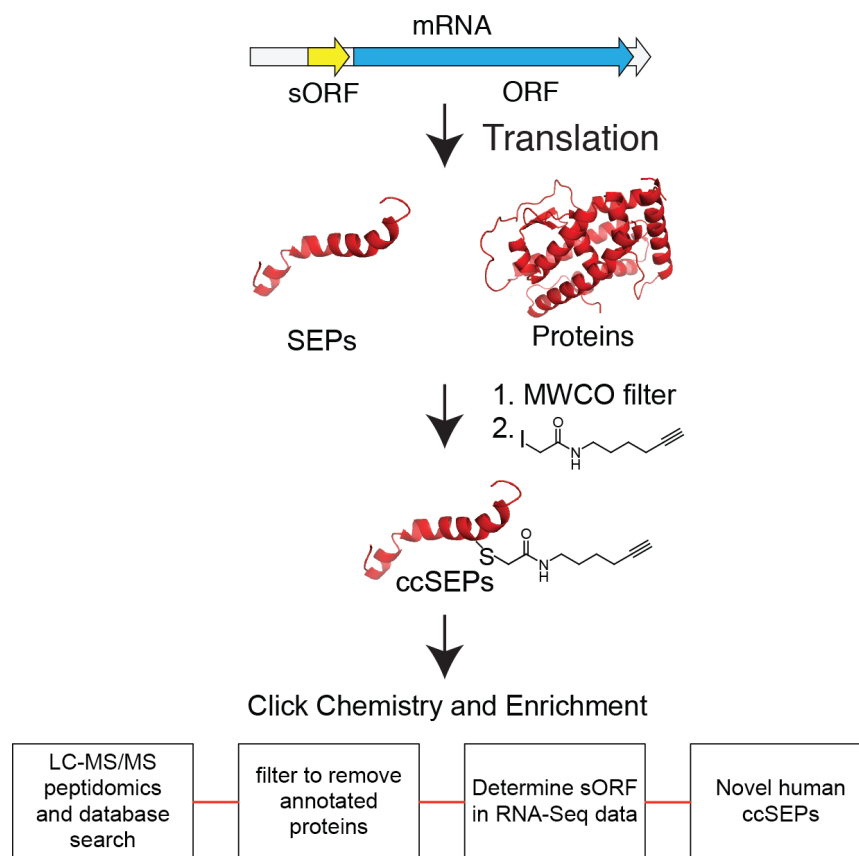


Figure A2.1. Workflow for identifying ccSEPs. The proteome and peptidome are separated by a MWCO filter and the peptidome fraction is carried forward to identify ccSEPs. Incubation of the peptidome with an iodoacetamide-alkyne (IA) probe leads to alkylation of cysteine-containing peptides including ccSEPs. Labeled peptides were then selectively enriched by conjugation to an azide-TEV-biotin tag using copper-activated click chemistry (CuACC) followed by affinity chromatography with streptavidin-coated beads. This sample is then analyzed by LC-MS/MS peptidomics and filtered to remove annotated proteins, which led to the identification of novel protein-generating sORFs that produce ccSEPs.

peptide through copper-activated click chemistry (CuACC) (17-19). Probe-labeled peptides are then separated from unlabeled peptides via streptavidin affinity chromatography to afford an enriched peptidome sample. On-bead trypsin digestion was performed, and unlabeled peptides were eluted and analyzed by offline Electrostatic hydrophilic Repulsion Liquid Chromatography (ERLIC) fractionation followed by LC-MS/MS (1, 20). The remaining bead-bound labeled

peptides were subsequently released from the beads by the addition of TEV protease, and were then analyzed by MudPIT-LC-MS/MS (21).

The data from this peptidomics analysis contains known as well as novel (i.e. non-annotated) peptides, including ccSEPs. In order to identify peptides originating from non-annotated RNAs, we used a custom database using K-562 RNA-Seq data (1, 22), which contains information on the vast majority of mRNAs in K-562 cells. Since these RNAs must be the source of any polypeptide produced we can include non-annotated genes in our peptidomics search by translating this database in three frames to generate a protein database that contains all possible peptide products.

We then matched our peptide spectra against this RNA-Seq database to reveal candidate SEPs. This approach yielded 175 hits that surpassed our preliminary cross correlation score requirements (17). After removing annotated peptides we were left with 109 candidate SEPs. Our K-562 RNA-Seq database was too large to perform a reverse database search directly. To overcome this, we constructed a forward and reversed database by appending our candidate SEPs to the Uniprot database. We used this database to filter our candidate SEP spectra using a reversed database search, and only accepted peptides with a false discovery rate < 0.05 . Subsequently, we validated that detected peptides could only originate from a single sORF (i.e. there are not two different ORFs in the RNA-Seq data that could account for the peptide). Additionally, SEPs with more than 2 missed cleavages were removed along with SEPs detected from peptides fewer than 7 amino acids in length. Furthermore spectra were visually inspected to ensure good sequence coverage and confirm that peptides detected from the TEV fraction contained an IA-modified cysteine residue. After this, we were left with 16 novel human ccSEPs (Table A2.1), with the majority having less than 6-ppm mass error.

Detected Peptide ^a	Start codon	Length (aa)	Length (Stop/ATG to stop)	transcript origin	annotated prot on transcript
C*GFFSYCSSESVCSTS	ATC	34	47	non annotated	
STSLYCHSTILC*	AAG	24	55	CDS	ALMS1
TC*DGNSNEGGGTR	AAG	19	25	non annotated	
NFPLASSPERC*FFVPK	AAG	48	49	3' UTR	UHRF1 binding protein 1, NM_017754.3
VEKLELLYIAGGNVNWYSPC*	GTG	22	24	non annotated	
YPAC*SPSPALI	CTG	29	38	non annotated	
GRGCC*RGFSAVGQGPSST	ATG	84	84	non annotated	
CPSINFQHFCHFVLCAPPIHC*	CTG	35	46	non annotated	
TC*TIPV/PAGGRPR	CTG	32	44	non annotated	
IC*DIKGLIDNV	TTG	41	45	non annotated	
TSPADAVC*PGLGRDLGSSRCCLRP	ATG	79	79	5' UTR	MRS2L
RGPGEAGMSWEEAGGLAPHLLC*CR	GTG	86	109	CDS	CDKN1A, NM_001220778
QIVLGGC*GEMV	alternate	16	16	non annotated	
GASFSEDGC*LLVG	CTG	37	41	non annotated	
GSSDIISVPC*	ATG	40	40	3'UTR	ERMP1 NM_024896.2
SSMPLIC*FLILEGLGR	ATG	29	29	3' UTR	ZBTB47, NM_145166.3

^a asterisk indicates labeled cysteine

Table A2.1. Detected peptides and the start codon and length (AUG or near cognate to stop) of their corresponding ccSEPs.

In cases where a detected peptide contained multiple cysteines, the labeled cysteine could be determined from the MS/MS data (Figure A2.2A). To verify that our labeling and enrichment is specific to the cysteine on a ccSEPs, we performed an *in vitro* assay in cell lysates. We first synthesized TCT-SEP (named for the detected peptide; Figure A2.2B) by solid phase peptide synthesis, along with a mutant of this TCT-SEP where the cysteine is replaced by a serine, TST-SEP. We incubated TCT-SEP in K-562 cell lysates and then added the IA-alkyne probe. After labeling, the lysate was mixed with a fluorescent azide in the presence of copper (II) sulfate and TCEP to promote CuACC. This fluorescently labeled lysate was then resolved on an SDS page gel to assess labeling of the TCT-SEP. Labeling of TCT-SEP was specific and robust and could be easily observed within total K-562 lysate (Figure A2.2B). The control TST-SEP was not labeled when probe-treated alone or in K-562 lysate demonstrating that labeling is occurring on the cysteine residue.

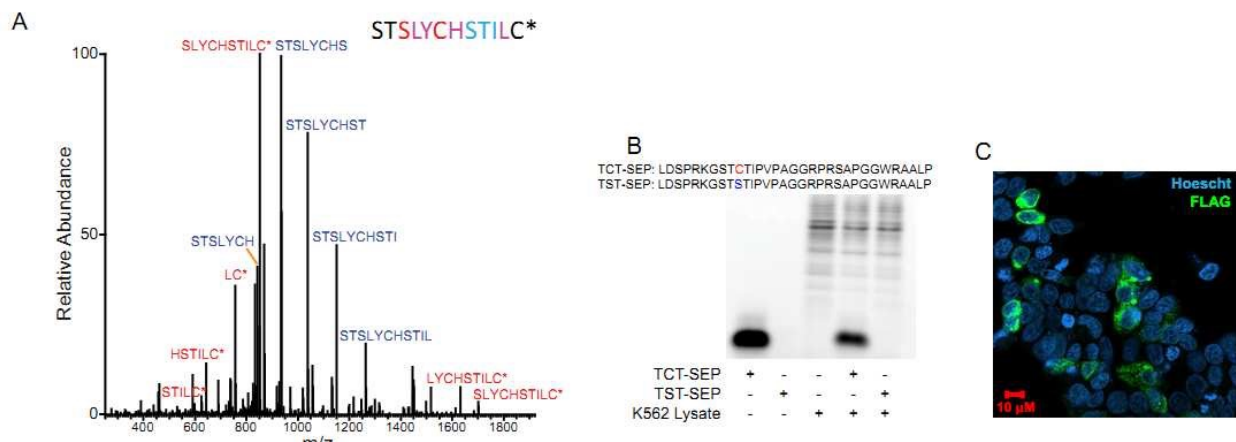


Figure A2.2. Validation of site of labeling and cellular expression of newly discovered ccSEPs. (A) In the case of ccSEPs with multiple cysteines, examination of the tandem MS spectra reveals the site of labeling. In this case, STS-ccSEP labels at the C terminal cysteine. Red indicates fragments detected by y ions, blue indicates fragments detected by b ions, and purple indicates fragments detected by both. (B) We tested labeling of one of the ccSEPs in a complex mixture by spiking the purified ccSEP into lysate and then performing a labeling reaction with rhodamine azide. If the ccSEP reacted it would fluorescently label. Mutation of the cysteine on the ccSEP to a serine abrogates labeling. (C) A C-terminal Flag tag appended to the sORF coding for TSP-ccSEP validated that this sORF does indeed produce protein. Staining of the protein product with an anti-Flag antibody confirmed expression and cellular stability of the ccSEP.

A 2.2.2. Validation of Cysteine SEP Labeling

To verify that our labeling and enrichment is specific to the cysteine on a ccSEPs, we performed an *in vitro* assay in cell lysates. We first synthesized TCT-SEP (named for the detected peptide; Figure A2.2B) by solid phase peptide synthesis, along with a mutant of this TCT-SEP where the cysteine is replaced by a serine, TST-SEP. We incubated TCT-SEP in K-562 cell lysates and then added the IA-alkyne probe. After labeling, the lysate was mixed with a fluorescent azide in the presence of copper (II) sulfate and TCEP to promote CuACC. This fluorescently labeled lysate was then resolved on an SDS page gel to assess labeling of the TCT-SEP. Labeling of TCT-SEP was specific and robust and could be easily observed within total K-

562 lysate (Figure A2.2B). The control TST-SEP was not labeled when probe-treated alone or in K-562 lysate demonstrating that labeling is occurring on the cysteine residue.

To validate the production of ccSEPs from their endogenous RNA, we transfected cells with a vector containing the sORF TSP-ccSEP, which is found on the same transcript as MRS2L. This construct contained the entire endogenous 5'UTR, which includes the sORF, and a FLAG tag was appended to the sORF to enable easy detection of protein production (Figure A2.2C). Stable ccSEP expression was then observed by immunofluorescence using an anti-FLAG antibody (green) (Figure A2.2C) and western blot (Figure A2.3). This sORF was not annotated previously, thereby highlighting the ability of this workflow to discover novel protein-coding genes. More generally, this affinity strategy successfully identified a new pool of SEPs with characteristic hallmarks of this emerging class of peptides (*1*).

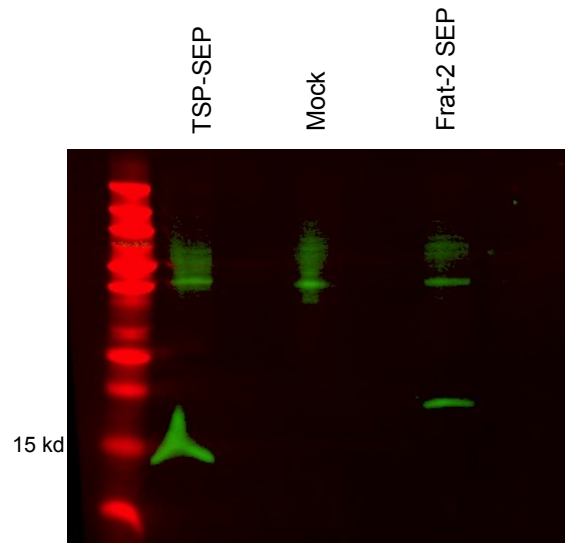


Figure A2.3. TSP SEP expression was confirmed by western blot. HeLa cells were transfected with plasmid encoding for TSP-SEP, empty vector, or Frat-2 SEP. Expression of TSP-SEP was confirmed by western blot. TSP-SEP ran several kDa higher than its anticipated molecular weight. To determine whether this represented a genuine shift in molecular weight or was an artifact, TSP-SEP was run alongside Frat-2 SEP. Frat 2-SEP, whose molecular weight has previously been determined by

Figure A2.3. (Continued) mass spectrometry, also ran several kd higher than anticipated demonstrating that the slightly slower migration of SEPs did not represent an actual change in molecular weight.

A 2.2.3. Novel ccSEPs

An overview of these newly identified ccSEPs revealed many similarities with previously identified SEPs. First, the length of their sORFs ranged between 16 and 86 codons (Figure A2.4A). SEP length was determined by measuring the number of codons between the stop codon of the sORF and the first start codon on the 5' side of this stop codon. In the case where a start codon couldn't be identified, the number of codons reflects the distance between the stop codon of the sORF and the 5' end of the transcript. Second, these SEPs had both AUG start codons or non-canonical near cognate start codons (Figure A2.4B), similar to previously discovered SEPs. Moreover, SEPs could be found in the 3'UTR, frameshifted within known genes or within the 5' UTR, in non-annotated RNAs, or in antisense transcripts (Supporting Information). As expected, we did not detect any previously observed SEPs, since our workflow was optimized towards the detection of SEPs with reactive cysteines. These identified SEPs are very small relative to the average length of a human protein, which is 335 amino acids (23). The small size of these SEPs contributes to the difficulties associated with computationally predicting the sORFs that encode them.

While specific functions for these ccSEPs await future studies, we examined these ccSEPs for sequence conservation, which is an important and well-documented signifier of biological function (24). We examined the conservation of our SEPs in several species by alignment of the translated RNA to *in silico* translated RNA and DNA databases comprising the GenBank, EMBL, DDBJ, PDB, and Refseq sequences. Of the ccSEPs we discovered, over one

third (6/16) are conserved amongst several species of primates indicating that they have been maintained throughout evolution and highlighting these ccSEPs as likely having functions. Notably, the cysteine residue that we find labeled by the IA probe is also conserved between species, including mice, despite the low overall sequence conservation across the entire SEP. This implies that this residue may be important for the SEP's biological function (Figure A2.4C). The conservation of these SEPs makes them good leads for further functional characterization, and demonstrates that this platform allows for the identification of peptides that are of significant biological interest.

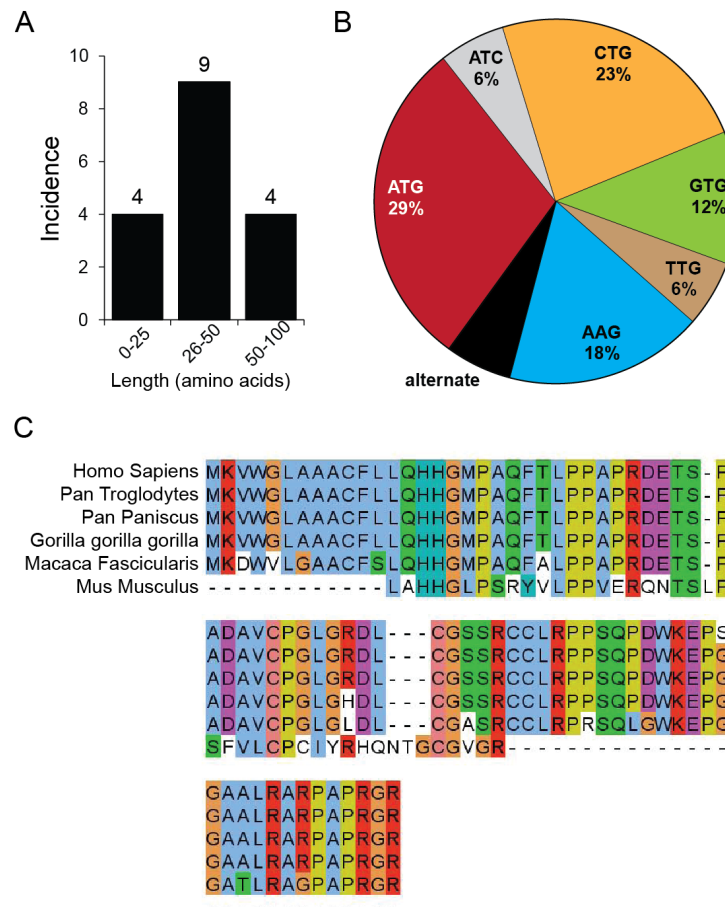


Figure A2.4. ccSEP overview. (A) Distribution of ccSEPs by their length in amino acids. SEP length was determined using the distance from an upstream in frame AUG start codon to a downstream in frame stop codon, or, when no inframe AUG was

Figure A2.4. (Continued) present, a near cognate start codon or stop codon was used instead. (B) While AUG is the predominant start codon for the production of ccSEPs, near cognate start codons (i.e. one base different from AUG) are also common. (C) TSP-SEP is strongly conserved amongst several species of primates suggesting this SEP may be functional.

A 2.3. Conclusion

In summary, we have utilized a chemoproteomics approach to identify new human ccSEPs. These results demonstrate the value of chemoproteomics to promote the discovery of additional sORFs. In this case, we identified 16 novel ccSEPs indicating the presence of even more of these molecules than had been predicted, and representing a 15% increase in the number of known SEPs. Moreover, conservation indicates that some of these ccSEPs may be functional. Furthermore, cysteine reactivity is governed by secondary structure and local environment, suggesting that enriching ccSEPs with highly reactive cysteines may identify proteins with distinct secondary structures. Additionally, certain biologically important post translational modifications, such as protein *S*-nitrosylation, occur at, and can be regulated by, redox active cysteines (25). Some of these ccSEPs are likely targeted by these oxidative modifications, which could serve to further regulate SEP function. The struggle to identify the whole range of SEPs in human cells as well as their functional role remains a key question in biology. The development of mass spectrometry methods focused on the identification of SEPs, such as chemoproteomic approaches, is a critical step towards answering these questions.

A 2.4. Methods

A 2.4.1. Cell culture

Cells were grown at 37°C under an atmosphere of 5% CO₂. K-562 cells were grown in RPMI 1640 medium with 10% FBS, penicillin and streptomycin. Cells were maintained between 1-10 x 10⁵ cells/ml. HEK293T cells were grown in DMEM with 10% FBS, penicillin and streptomycin.

A 2.4.2. Isolation of polypeptides

1x10⁹ K-562 cells were washed with ice-cold PBS 3 times. Cells were subsequently suspended in lysis buffer with 0.1M ammonium acetate, 0.5M NaCl, diprotin A (1 µg/mL), antipain (1 µg/mL), leupeptin (1 µg/mL), chymotrypsin (1 µg/mL) at pH 3.6 on ice. Cells were sonicated on ice for 20 bursts with output level 2 using 30% duty cycle (Branson Sonifier 250). Samples were then centrifuged at 3,000 x g for 10 min. Supernatant was collected and centrifuged through a 30kD molecular weight cutoff filter at 20,000 x g for one hour (Modified PES, Centrifugal Filter, VWR). Filtrate was dialyzed into PBS, and polypeptide concentration was measured using the BCA assay. Cysteine containing SEPs were then enriched from this polypeptide sample for MudPIT-LC-MS/MS analysis as described below.

A 2.4.3. MudPIT-LC-MS/MS analysis

Probe-labeling (Figure A2.5), click chemistry, and streptavidin enrichment: Polypeptide samples (n = 3) were probe labeled with IA-alkyne (100 µM) for one hour at room temperature. Probe-labeled samples were subjected to click chemistry. Biotin-TEV-azide (200 µM), TCEP (1 mM, 50X fresh stock in water), ligand (100 µM, 17X stock in DMSO:t-Butanol 1:4), and

copper(II) sulfate (1 mM, 50X stock in water) were added to the protein. Samples were allowed to react at room temperature for 1 hour. Tubes were centrifuged (10 mins, 4°C) to pellet the precipitated proteins. The pellets were suspended in cold methanol by sonication. Centrifugation was followed by a second methanol wash, after which the pellet was solubilized in PBS containing 1.2% SDS via sonication and heating (5 min, 80°C). The SDS-solubilized, probe-labeled proteome samples were diluted with PBS (5 mL) for a final SDS concentration of 0.2%. The solutions were incubated with 100 µL streptavidin-agarose beads (Thermo Scientific) at 4°C for 16 hrs. The solutions were then incubated at room temperature for 2.5 hrs. The beads were washed with 0.2% SDS/PBS (5 mL), PBS (3 x 5 mL), and water (3 x 5 mL). The beads were pelleted by centrifugation (1300 x g, 2 min) between washes.

On-bead trypsin digestion: The washed beads were suspended in 6M urea/PBS (500 µL) and 10 mM dithiothreitol (from 20X stock in water) and placed in a 65°C heat block for 15 mins. Iodoacetamide (20 mM, from 50X stock in water) was then added and the samples were placed in the dark and allowed to react at room temperature for 30 minutes. Following reduction and alkylation, the beads were pelleted by centrifugation (1300 x g, 2 min) and suspended in 150 µL of 2 M urea/PBS, 1 mM CaCl₂ (100X stock in water), and trypsin (2 µg). The digestion was allowed to proceed overnight at 37°C. The digestion was separated from the beads using a Micro Bio-Spin column (BioRad). The beads were washed with PBS (3 x 500 µL) and water (3 x 500 µL) to remove tryptic peptides and urea.

On-bead TEV digestion: The washed beads were suspended in 150 µL of TEV digest buffer with AcTEV Protease (5 µL; Invitrogen) for 12 hr at 29°C with mild agitation. The eluted peptides were separated from the beads using a Micro Bio-Spin column and the beads were

washed twice with 75 μ l water, and washes were combined with eluted samples. Formic acid (15 μ l) was added to the samples, which were stored at -20°C until mass spectrometry analysis.

Offline electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) fractionation of peptides:

ERLIC fractionation was performed offline prior to LC-MS/MS analysis using a PolyWax LP column (200 mm x 2.1 mm, 5 μ m, 300 \AA , PolyLC Inc) connected to an Agilent Technologies 1200 Series HPLC equipped with a degasser and automatic fraction collector. Runs were performed with a flow rate of 0.3 ml/min. A 70-minute gradient beginning with 0.1% acetic acid in 90% acetonitrile and ending with 0.1% formic acid in 30% acetonitrile was used, and eluent was collected in 4 fractions. Fractions were evaporated to dryness before analysis by LC-MS/MS. Samples fractionated by ERLIC were not fractionated by SCX, and were loaded directly onto a C18 column for analysis.

Liquid chromatography-mass spectrometry (LC-MS) analysis: LC-MS analysis was performed on an LTQ Orbitrap Discovery mass spectrometer (ThermoFisher) coupled to an Agilent 1200 series HPLC. Digests were pressure loaded onto a 250 μ m fused silica desalting column packed with 4 cm of Aqua C18 reverse phase resin (Phenomenex). The peptides were eluted onto a biphasic column (100 μ m fused silica with a 5 μ m tip, packed with 10 cm C18 and 3 cm Partisphere strong cation exchange resin (SCX, Whatman)). Using a gradient 5-100% Buffer B in Buffer A (Buffer A: 95% water, 5% acetonitrile, 0.1% formic acid; Buffer B: 20% water, 80% acetonitrile, 0.1% formic acid). The peptides were eluted from the SCX onto the C18 resin and into the mass spectrometer following the four salt steps outlined in Weerapana et al(17). The flow rate through the column was set to ~ 0.25 $\mu\text{L}/\text{min}$ and the spray voltage was set to 2.75

kV. The capillary temperature was set to 200°C. One full MS scan (400-1800 MW, 30000 resolution, maximum scan time of 500ms) was followed by 18 data dependent scans of the n^{th} most intense ions with dynamic exclusion enabled (LTQ scan type: Normal; maximum scan time of 100ms; 1 microscan per ion). The normalized collision energy was set to 35.

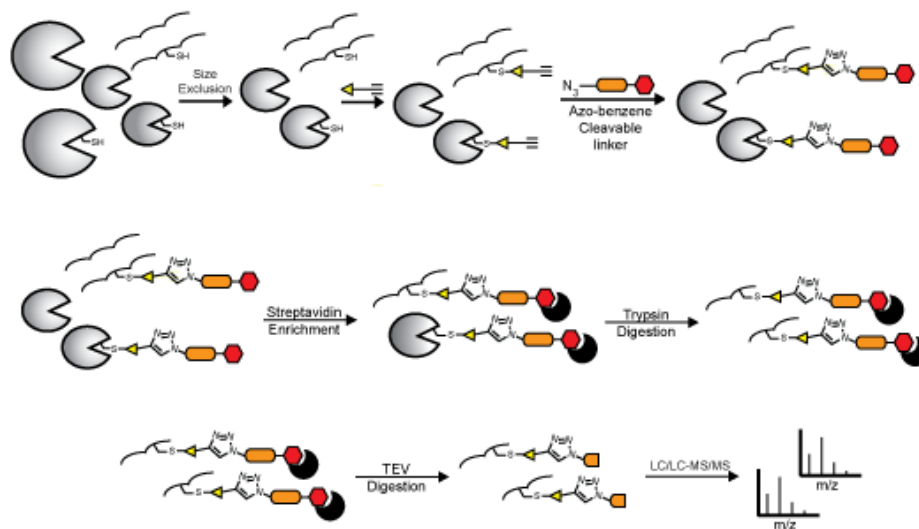


Figure A2.5. Workflow for isolation, enrichment, and orthogonal proteolysis of cysteine SEPs from K-562 lysates for LC-MS/MS identification.

A 2.4.4. MS data analysis

The generated tandem MS data was searched using the SEQUEST algorithm against the databases listed in the main text. A static modification of +57 on Cys was specified to account for iodoacetamide alkylation, and a differential modification of 464.28596 was specified on Cys, corresponding to the IA-alkyne probe conjugated to the cleaved Biotin-TEV-azide tag. The SEQUEST output files generated from the digests were filtered using DTASelect 2.0. Samples

with an XCorr score above 1.8 (+1), 2.5 (+2), 3.5 (+3) and deltaCN score above 0.08 were accepted from the search using the complete K-562 RNA-Seq database. From an average of approximately 61100 MS2 spectra per sample, resulting in an average of 96 peptide IDs per sample and a total of 175 unique peptides that passed these preliminary requirements.

Hits were then subjected to an iterative reverse database search. A reverse database was constructed by appending sORF encoded peptide sequences, which coded for unannotated detected peptides to the non-redundant human Uniprot database. This database was reversed, and detected peptides were re-searched against the forward and reversed appended human IPI database. A 5% FDR threshold was set, and actual FDR rates per sample were mostly in the range of 2.5-3.8%.

Peptide hits were then searched against the human IPI database using a string matching algorithm and matches were removed. Remaining hits were searched against the non-redundant protein database using Basic Local Alignment Search Tool (BLAST) and any peptides that matched known proteins were removed. All detected peptides have 7 or more amino acids.

Spectra of the remaining peptides were manually validated to ensure a precursor mass error of < 10 ppm. Spectra also contained at least 5 sequential b or y ions, and no more than 2 missed cleavages. In the case of peptides identified from the biotin-eluted fraction, all peptides were labeled with an iodoacetamide probe at a cysteine residue.

SEP length was calculated using the length from the first AUG or near cognate start codon upstream of the detected peptide to the first stop codon downstream of the detected

peptide. SEP length could also be calculated using the length from the first AUG or stop codon upstream of the detected peptide to the first downstream stop codon (Table A2.1).

A 2.5. References

1. S. A. Slavoff *et al.*, Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology* **9**, 59-64 (2012).
2. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147**, 789-802 (2011).
3. I. G. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso, Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family. *PLOS Biology* **5**, 1052-1062 (2007).
4. Y. Hashimoto *et al.*, A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and A β . *PNAS* **98**, 6336-6341 (2001).
5. M. C. Frith *et al.*, The Abundance of Short Proteins in the Mammalian Proteome. *Proteins* **2**, (2006).
6. S. Lee *et al.*, Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *PNAS* **109**, (2012).
7. N. Stern-ginossar *et al.*, Decoding Human Cytomegalovirus. *Science* **338**, 1088-1093 (2013).
8. B. Guo *et al.*, Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **423**, 456-460 (2003).
9. H. A. Chapman, R. J. Riese, G.-P. Shi, Emerging Roles for Cysteine Proteases in Human Biology. *Annu. Rev. Physiol.* **59**, 63-88 (1997).
10. K. G. Reddie, K. S. Carroll, Expanding the functional diversity of proteins through cysteine oxidation. *Current Opinion in Chemical Biology* **12**, 746-754 (2008).
11. A. Nott *et al.*, S-nitrosylation of HDAC2 regulates the expression of the chromatin-remodeling factor Brm during radial neuron migration. *PNAS* **110**, 3113-3118 (2013).

12. M. Ikeguchi, K. Kuwajima, S. Sugai, Ca²⁺- Induced Alteration in the Unfolding Behavior of α -Lactalbumin. *J. Biochem.* **99**, 1191-1201 (1986).
13. H. J. Coyne III *et al.*, The Characterization and Role of Zinc Binding in Yeast Cox4. *The Journal of Biological Chemistry* **282**, 8926-8934 (2007).
14. A. Morleo, F. Bonomi, S. Iamentti, V. W. Huang, D. Kurtz Jr, Iron-Nucleated Folding of a Metalloprotein in High Urea: Resolution of Metal Binding and Protein Folding Events. *Biochemistry* **49**, 6627-6634 (2010).
15. J. M. Scholtz, R. L. Baldwin, The Mechanism of α -Helix Formation By Peptides. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 95-118 (1992).
16. H. Kozlowski, W. Bal, M. Dyba, T. Kowalik-Jankowska, Specific structure-stability relations in metalloproteins. *Coordination Chemistry Reviews* **184**, 319-346 (1999).
17. E. Weerapana, A. E. Speers, B. F. Cravatt, Tandem orthogonal proteolysis-activity-based protein profiling (TOP-ABPP)--a general method for mapping sites of probe modification in proteomes. *Nature Protocols* **2**, 1414-1425 (2007).
18. E. Weerapana *et al.*, Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature* **468**, 790-795 (2010).
19. P. Wu *et al.*, Efficiency and Fidelity in a Click-Chemistry Route to Triazole Dendrimers by the Copper(I)-Catalyzed Ligations of Azides and Alkynes. *Angew. Chem. Int. Ed* **43**, 3928-3932 (2004).
20. A. Alpert, Electrostatic Repulsion Hydrophilic Interaction Chromatography for Isocratic Separation of Charged Solutes and Selective Isolation of Phosphopeptides. *Anal. Chem.* **80**, 62-76 (2008).
21. M. Washburn, D. Wolters, J. Yates, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **19**, 242-247 (2001).
22. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. World, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621-628 (2008).
23. T. Ota *et al.*, Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genetics* **36**, 40-45 (2004).
24. J. Ponjavic, C. Ponting, G. Lunter, Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research* **17**, 556-565 (2007).

25. D. T. Hess, A. Matsumoto, S. Kim, H. E. Marshall, J. S. Stamler, Protein S-Nitrosylation: Purview and Parameters. *Nature Reviews Molecular Cell Biology* **6**, 150-166 (2005).

		CGGBP1; NM_001195308.1 CGGBP1; XM_005264772.1						
range=chr15:87878674-87879358 strand==+	EESVSLMIGASQTEIER	non annotated		non-AUG	125	SOFSEESVSLMIGASQTEIEREKIVLKKDRPFNDCTISKHVTHISAIP EERENRVQTTYTVLTVENFKLITATKSQTQEDERTPNRMNTKNGT		
range=chr3:12568054-12568215 strand=-	KMGIM*VK	non annotated		non-AUG	45	FIHILFKLQTNKRKENIFKVARKGWLSLGG		
range=chr5:138275241-138283167 strand==+	GSEM*IPCLTSQLGGASTSVKLS	non annotated		non-AUG	24	RGSEMIPCLTSQLGGASTSVKLS		
range=chr17:4835352-4837123 strand==+	TRCIQYQRAFLGPTSC	GP1BA; NM_000173.5	CDS	AUG	70	MGWRISIPFSSKRTRCIQYQRAFLGPTSCLLFFSTGTPGYATVRSS IFVAGCRTMLKMSTYGSVKVWTSRP		
range=chr12:117151124-117175823 strand=-	SM*SATQQGAMAGAV	GP1BA; XM_005256611.1 C12orf49, XM_005253937.1	3' UTR	AUG	44	MGSQHLPLRVTVRCRCEQGLRLGKKSMSATQQGAMAGAVRGRLL G		
range=chr10:95057205-95057400 strand=-	WVTLEGSMLGEISQM*KQDK	non annotated		AUG	57	MDEWISKMWYLTMEYLLGLRKEILSHATTWVTVLEGSMLGEISQ MGKDKCCMIPFI		
range=chrX:15843934-15863636 strand=-	PTQFFFK	non annotated		AUG	11	MPTQFFFKREL		
range=chrX:15843934-15863636 strand=-	MPTQFFFK	non annotated		AUG	11	MPTQFFFKREL		
range=chr8:17873932-17874063 strand==+	IVYGDIRK	non annotated		non-AUG	41	LYQRREELKLQWRSNFLDKIVYGDIRKELINLAETMGSSFF		
range=chr8:17873932-17874063 strand==+	KIVYGDIR	non annotated		non-AUG	41	LYQRREELKLQWRSNFLDKIVYGDIRKELINLAETMGSSFF		
range=chr3:44462454-44493083 strand=-	ADAPAVSPESPQK	XR_133338 LOC100506301	non-coding	AUG	101	MSRQFGQGOEPLDMFFWVNEISEITYPPQKADAPAVSPESPQ KPPFPQPSVQEAQPCSPQGPAPRALAPPSPKSLKDSGRNFC PSAPTWARPKPEE		
range=chr20:45298911-45299385 strand=-	PNVQVVIK	non annotated		AUG	77	MLLYGVTLAARKRECLKTSPNVQVWIKSLFASHLLMSHWPKQVTV PSPSEWEGTTQGEYQEAWVCGGHQYIYTS		
range=chr9:88428049-88428393 strand==+	KNFFISK	non annotated		AUG	20	MPKNFFISKALAYIGARC		
range=chr6:30522829-30531396 strand==+	EAESSPATDTAAAPAAR	NM_025263 PRR3	5' UTR	AUG	93	MPGRDGGROLAPECGRRRCSPOPSGSGMAGASALLPTDPEAE TESPRASPLSPKSRCSCHRRHDAETKEAESSPATDTAAAPAARA GRDWR		
range=chr6:30522829-30531396 strand==+	RHDAETKEAESSPATDTAAAPAAR	non annotated						
range=chr6:32780349-32806575 strand=-	FSPLELAAGVVR	NM_018833 TAP2	CDS	AUG	127	MAGGDPAAAGRAVGAAKAKRAAGICGDTAAPALSHPDCLPESP GRGGLTCSQSPQRSFSPLELAAGVVRGCGAQLVTVGCSFSPSWSPG EGAGPGEQQSLDVEAAEALQAGPAPRCLLLPCPCFCFG		
range=chr16:33121842-33122879 strand==+	YLDLHER	non annotated		non-AUG	33	IRKYLDLHERQIFHSNKHNLNINCAQITTRVK		
range=chr17:48178065-48178330 strand==+	RTLSEITQGLGGGIGIRT	non annotated		non-AUG	63	RTLSEITQGLGGGIGIRTQLDLETGFLKKGEGACTLGSVSEISK FMVQVCGRAELTLQNC		
range=chr19:40327193-40330486 strand==+	KGMSHLVWPK	non annotated		AUG	47	MGFYHIGQAGLELLTSGDPPASASESVEIKGMSHLVWPKPEVLRH YT		
range=chr14:58955358-58955764 strand=-	GEGALVPAGNGIK	NM_001244189 KIAA0586	CDS	non-AUG	50	MGLSKVSVTVTLNLVSSVDESEPGICEGEGALVPAGNGIKGTGS GGGG		
range=chr13:34124857-34125083 strand=-	EEYSVGVTTMYDLKK	non annotated		non-AUG	35	DGGISGKCLIEEYSVGVTTMYDLKKEKDTLLEVL		
range=chr11:11803558-11805863 strand=-	MLFVAVK	non annotated		AUG	24	MVIAMLFVAVKWKWWSKCHQRTG		
range=chr11:11803558-11805863 strand=-	MLFVAVKK	non annotated		AUG	24	MVIAMLFVAVKWKWWSKCHQRTG		
range=chr6:118017079-118018042 strand==+	ESEIIKK	non annotated		AUG	53	MLYDCLCTSCFVDCVFSRIQKVRILSDKFNKESEIKKNOAEILELK MQLTY		
range=chr4:84156218-84156708 strand=-	KVFLLNK	NM_052834 WDR7		non-AUG	18	DWSSGDEFPQKVFLLNK		
range=chr3:33537746-33759712 strand=-	QGSVLQVALR	NM_001207044 CLASP2	CDS	AUG	61	MMKNQWMEIGHQLHQPFRFLHLKHPILPTVQGSVLQVALRLE VLLRKEVLEQLMKMIL		
range=chr19:38230342-38231032 strand=-	RNLVVVINL	NM_001172692 ZNF573	CDS	AUG	71	MNVRRNRRNLVVINLVYITGFMSLRDPMNAKSVGRTFVAINLLYI KDFILVRNPMNVQVGRTELVVIS		
range=chr8:124471523-124472268 strand==+	NLRLVQAILL	XR_111795 LOC100509315	non-coding	non-AUG	35	MEFHSPLRLECNGTISAHRNLRLVQAILLPQPPK		
range=chr14:56021802-56026332 strand=-	GHLRPSLSTKLEIILESMSE	non annotated		non-AUG	34	GHLRPSLSTKLEIILESMSELYELVLYNTGS		
range=chr5:149891915-149892156 strand==+	NLGFQRPLETLVERSWSPLHFQ	non annotated		non-AUG	34	NLGFQRPLETLVERSWSPLHFQCLVACLAPSGAQ		
range=chr18:9345769-9345975 strand==+	IYRNLVLEK	non annotated		non-AUG	24	VNFCNKASWDLKKIYRNLVLEK		
range=chr16:624062-629182 strand=-	EGGWRQVEGTGTPK	non annotated		AUG	75	MTTGERDSAKPIRATATROEDRSPGGWRQVEGTGTPKSKQGSR VLAQQEETHQPEAVPQRADPKGASASPLRRQ		
range=chr4:83831822-83832435 strand=-	ELSQYLK	non annotated		AUG	42	MDELSQYLVLPSTVTVLQVLLQLSFLLYANFLSLGSLP		
range=chr6:17701564-17701878 strand==+	SRQVDQEVRRSR	non annotated		non-AUG	24	SQHFGRSRQVDQEVRRSRTAWPRW		
range=chr3:184053746-184068183 strand==+	PGAYGLSR	NM_144635 FAM131A	3' UTR	non-AUG	80	GWGGRGPSNPPSPCLSLCFSSHRVHVQCLIESPPPGAGSC PPGAYGLSRPAPALQSGCCASFTSPSSLNPLFS		
range=chr6:10801828-10803161 strand==+	ENIPDITK	non annotated		non-AUG	32	DFVFNLSKILVENRPAFVNIPIITKPKHF		
range=chr11:160313607-160313730 strand==+	LFETKALFCGCCAALK	non annotated		non-AUG	32	FFLAFRLFETKALFCGCCAALKRWISYFFLC		
range=chr2:156867321-156871777 strand==+	LFAAPSLNLQSKFER	non annotated		AUG	28	MYLLFAAPSLNLQSKFERGRETERERID		
range=chr12:122714123-122751118 strand=-	VAEIIIER	non annotated		non-AUG	116	RVAEIIERLVSHGINLALFISWKLKENHFHCRKSFYKLYLPREISL YLPQAVISCFREWNPPCPSIFWFLGLNSLVKSPWLGLSWEQIL SCSLMCLHSPKLYTLLRA		
range=chr8:9565334-9565564 strand==+	AIVVARVVTIPK	non annotated		AUG	43	MVAIVVARVVTIPKIMHPVLSFLNFHVPVLYTFMSVYVDSLVS		
range=chr2:108294549-108364747 strand=-	GDFLNL	non annotated		non-AUG	41	AGDFLNLRIGISYQCFKSPINYYFFFLSPCLLYGILLDIS		
range=chr2:108294549-108364747 strand=-	AGDFLNL	non annotated		non-AUG	41	AGDFLNLRIGISYQCFKSPINYYFFFLSPCLLYGILLDIS		
range=chr5:139941187-139946658 strand==+	GFLAGYVAK	NM_080670 SLC35A4	5' UTR	AUG	103	MADDKDSLPLKDLAFLKNQLESQRREVEVNSGVGDGSLSS PFLKGLAGYVAKLRASAVLGFVAGTCTGYAAQAYAVPNVEKTLR DYLQLLRKGPD		

Table A3.1. (Continued) Full list of SEPs detected in Chapter 2.

range=chr5:139941187-139946658 strand=+	NQLESLQR RVDEVNSGVGDGSLSSPFLK					
range=chr5:139941187-139946658 strand=+						
range=chr8:99962637-99963699 strand=-	RLLFAGK	non annotated	non-AUG	22	GTTFSWVIRLLFAGKLNYSMS	
range=chr8:99962637-99963699 strand=-	IRLLFAGK				GTTFSWVIRLLFAGKLNYSMS	
range=chr4:43901065-43901402 strand=+	GLIENPALIR	non annotated	non-AUG	100	QRVQAEIRLAIARLKRKYLLQVNDPNRQGLIENPALIRWAYARTTN VYPNFRPTKNSLMGALCGFGLIFYIIKTERDRKEKLMQEGKLD RTVHLSY	
range=chr2:82967185-82967587 strand=+	LEETLEIAAR	non annotated	non-AUG	22	TFISPHYLEETLEIAARKCTP	
range=chr5:127419792-127524262 strand=+	HDFINLK	NR_046207 SLC12A2	non-coding	non-AUG	34	LFFHDFINLKAHRKVAPLITCMETSVLNVNSISQS
range=chr16:15198209-15199311 strand=-	QNIKGLIENLOK	NR_003610 PDXDC2P	non-coding	non-AUG	50	IWSRVVTLTLOSSENQRONIKGLENILQKEAATCVDNGLFMPLLLSL TVC
range=chr2:65855310-65855473 strand=-	SWLTPVAGK	non annotated	AUG	26	MAPLGLKDLPLSSWLTPVAGKLVMAVS	
range=chr16:11933549-11935884 strand=+	HALPLLK	non annotated	non-AUG	52	NSTNFFLLIKQRSFGGFIADKRKDGKGCGRFLSFHKQEFHALPLL KQRKE	
range=chr18:9864909-9865248 strand=+	GAGILLR	non annotated	non-AUG	44	TGAGILLRWLTHWLGLSRSSPGVPLHVLLHGLMMWHEPHSV	
range=chr8:130294845-130296415 strand=+	KQNSLIANMEK	non annotated	non-AUG	114	LIKKONSLIANMEKVLVWVMDQTSNHIPLSOSLQSKGQTLFSSTK NEGEEAAEEKFEASRVWLMRFKERSCLSHIKMQIEATRADEEGT ASDPEDPAKLIDKSGYINRFTM	
range=chr2:78557644-78561636 strand=-	LASIVANK	non annotated	AUG	36	MLELFFPNSVNIHVSHLLASIVANKMFIVSQALPYC	
range=chr2:175212881-175263481 strand=-	MKNFLAVTITGK	NM_004882 CIR1	CDS	non-AUG	110	IDWRKRKRKIEKRSFRRAEAVNTKNISPLLPHLPPPPPLLRLOK AVVRVVTIKKKYGRKERTKTSVQGITTVILKRRTLSRRESFMKNF LAVTITGKPKRSPGS
range=chr8:129965805-129966053 strand=+	QTFIGGIR	non annotated	non-AUG	25	TIRHKCQTFIGGIRTNDNFGIIDI	
range=chr5:79095189-79095905 strand=+	KDLHLSWEPEK	NM_153610 CMYA5	3' UTR	non-AUG	43	KQQPPLFSLYKEFFPKLDLHLSWEPEKNGPLSVLLVKEIL
range=chr2:146488355-46488674 strand=+	KNEFLK	non annotated	non-AUG	27	IIFKNEFLKDHVLFKSFSSYFCYC	
range=chr2:176563995-176564390 strand=-	VGLLAISPTAPK	non annotated	AUG	28	MLRKRRETVGLLAISPTAPKGFPLQL	
range=chr15:89045061-89055234 strand=+	KALFLQK	NM_001170794 BACH2	3' UTR	non-AUG	18	DQPGQHRKALFLQKIKNN
range=chr16:9153978-9159081 strand=-	AEIILK	non annotated	AUG	17	MAEIIILKAKVFDLQDF	
range=chr7:12353497-12446953 strand=-	ILRMEIFCSEKVDNLEFI	NM_001135924 VWDE	3' UTR	non-AUG	28	HQVYELQACILRMEIFCSEKVDNLEFI
range=chr5:139483515-139483659 strand=-	TPLLAYIQ	NR_102739 LOC100505636	non-coding	AUG	25	ELCCIFCGSSKTPLLAYIQDTSFSAF
range=chr4:77035820-77069583 strand=-	LDSLVLVR	non annotated	non-AUG	83	PTWSPAFATWPSILGLPRAPPVLPQPPRPRVGLDGOHLQLQV LHSAFLPQLTQALLDSLVLVRLTLDLVLVAQRELVLV	
range=chr1:52519251-52521743 strand=+	HAFNLNR	non annotated	non-AUG	54	HAFNLNRAIPSPQSNLNRPVQQLLHSPDLLSPRNLQTPGAVGE DKKKSQVA	
range=chr13:108521011-108526878 strand=+	NALILIR	non annotated	non-AUG	29	NTHPHLLDNFFNNALILIRHKHFNNG	
range=chr1:52960264-52960434 strand=+	EVEGAVSR	non annotated	non-AUG	42	TQEVGAVSRDCITALQPGKQSEVMQKQTTKIFNHTLIK	
range=chr7:12450740-12451773 strand=-	RKPLYTIGWNL	non annotated	non-AUG	64	SSGKGSNSQRDFTSHQLERLSSKRNKRVGKNAEKRPPLYTIG WNLNWSYHYKQHGSSKN	
range=chr13:108521486-108526772 strand=-	KINALLK	non annotated	non-AUG	50	SQPPLKCLIKINALLKGNILLSNCKCGVYFHTSILRKCWTSYHYK TGN	
range=chr11:64002664-64006302 strand=-	VRM*LDLILQLQ	non annotated	non-AUG	17	VRMLDLILQLQVPAVW	
range=chrX:119745582-119754975 strand=-	FQPPHHVQSSPDVK	NM_014060 MCTS1_v1.v2	3' UTR	non-AUG	32	ESCAPEPEQKGLSFQPPHHVQSSPDVKSQFWF
range=chr8:9403098-9403311 strand=-	LPFLYTVLLPK	non annotated	AUG	19	MCLPFLYTVLLPKLPSVDL	
range=chr19:40529109-40596836 strand=-	QTLFNLR	NM_001005851 ZNF780B	CDS	non-AUG	67	HSLLNIRSFILVRSNLNRSVGRPLVAQTLFNLRVILVRNPMNVR VGRLLDFTYNFLCIKNLYR
range=chr11:134094604-134117677 strand=+	VLEEGEQR	NM_052875 VPS26B	3' UTR	non-AUG	47	APFLGGLPCVLEEGERARILGWSLSELPDGTSEHVPGILSCQQG PG
range=chr9:115980741-115983650 strand=+	RLLITTSR	non annotated	non-AUG	40	QTKTRLLITTSRLIVNSKRPNRSEGRSDTPTISIRLEYS	
range=chr19:50840986-50856670 strand=+	NNLRETLAQKP	non annotated	AUG	32	MMTNTFFPRYRNSHYQPKYNNLRETLAQKP	
range=chr16:53728816-53730245 strand=+	IGTAFLNK	non annotated	AUG	16	MPSQKIGTAFLNKKE	
range=chr10:15478885-15479329 strand=+	LLLDLNKSQLGK	non annotated	non-AUG	23	KLLLDLNKSQLGKISRQKIFMAC	
range=chr16:72452815-72454466 strand=+	QVKALIK	non annotated	non-AUG	18	YTDVAVLVYFQVKALIK	
range=chr3:19778735-19778967 strand=-	IGESFEK	non annotated	non-AUG	49	RVFICIFRASFKINYKVLKIGESFEKSKITVCPPEELLHSYSKEGQ N	
range=chr8:91519522-91520168 strand=-	MGVDFLPQK	non annotated	non-AUG	26	VISTYEMGVDFLPQKINIKKIQYI	
range=chr21:35275511-35288046 strand=+	AVSNQLIPK	non annotated	non-AUG	44	TLLSDHVLGAVSNQLIPKHNLLKPNLSLQKQGLAFILRGYQKRT	
range=chr11:2391037-2391433 strand=-	LALFLPR	non annotated	AUG	41	VPSGAWGTCVAIFSRGSLWALFLPRGPCRAPSVEGENMK	
range=chr16:16170189-16236878 strand=+	QPELRLPQ	NM_004996 ABC1	3' UTR	non-AUG	49	GVSLPKVESGRQPELRLPQSDLLRITPPTSPESIFLLGENAYY FL
range=chrX:120338054-120341958 strand=-	IIAYIKK	non annotated	non-AUG	41	KFLENDDGNATDQKLWDTTKAGLRGKFAIIAYIKKKNL	
range=chr12:74853896-74854278 strand=-	WSWRTLLLLPLNI	non annotated	non-AUG	42	WSWRTLLLLPLNIRTPGSEVFEFLNLHQAQLRFSGLQKTES	

Table A3.1. (Continued) Full list of SEPs detected in Chapter 2.

range=chr21:20167140-20167299 strand=+	SIEVIIR	non annotated		AUG	23	DILLHDFLASIEVIRFFSINL
range=chr5:34183374-34185173 strand=+	HDFEVKR	non annotated		non-AUG	33	IKSINYICWIEIEKVHISHDFEVKREIRIFGKEG
range=chr5:81078543-81079471 strand=-	IPLSIVIR	non annotated		non-AUG	47	KYWATPTIKIPLSIVIRKRFIKFMSNRLTHMRQVRVFAHATPEAVSG
range=chr16:8991978-8996350 strand=+	FDINLR	non annotated		AUG	27	MNYPYRILRSDFINLRGRKQLFLY
range=chr2:69476143-69476401 strand=+	LKSTLFSGCLFVIK	NM_032208 ANTXR1	3' UTR	non-AUG	28	KNNKFLKSTLFSGCLFVIKCNVFKSSIS
range=chr22:26860929-26864708 strand=+	HLALGALR	non annotated		non-AUG	23	ITRVKIHQHLALGALRSRDLTY
range=chr7:6789180-6793491 strand=-	PTFKFMK	non annotated		non-AUG	23	ESLTKDPTFKFMKLNKDEFNYC
range=chr20:25712848-25713130 strand=-	ENLEGIATKPGK	non annotated		non-AUG	35	LSELSNGVFLHENLEGIATKPGKERTMVPMEKVHK
range=chr5:80597485-80605440 strand=-	WLIFFFFGFR	non annotated		AUG	57	MVHCSLDLPGPILLPQLAVAGITGVYHHTWLIFFFFGRHGVSLC GPGSSLIPLGK
range=chr1:161284832-161284990 strand=+	FILNPLSR	non annotated		non-AUG	32	ISCEKYPRFSFILNPLSRSHSLLSLSSHSD
range=chr7:44800637-44801496 strand=-	KTGPESVGGGTEPR	non annotated		non-AUG	48	IAAPGQACSGRWVDTQKTGPESVGGGTEPRGGQVARRSSPHY QQLVD
range=chr9:128724377-128729551 strand=+	QLLGLKKG	NR_024123 PBX3	non-coding	non-AUG	18	WFHSCQLLGLKGGKYNIH
range=chr11:3408735-3430083 strand=+	IAAGALSPLR	non annotated		non-AUG	33	LPCDLYRNIRGGEAGDITSRIAAGALSPLRCSGS
range=chr16:18089134-18089599 strand=+	ISEVILR	non annotated		non-AUG	22	PHSVTNTQKSLLSNVISEVILR
range=chr16:29982572-29984370 strand=-	ARDQYGHLIPTK	non annotated		AUG	61	MCAEIEGAEVGTARDQYGHLIPTKVASGPGQLSGARKPSPFSPR LRGSCFLSQVGGWGI
range=chr15:31218600-31220691 strand=+	LAFIFLDR	non annotated		non-AUG	65	AKIVPLHSSLDGDRVRPCLTKTKQKFRNDLAFIFLDRQCIHQDGT TGNQVLAPLAGKEHEVF
range=chr17:56066228-56086144 strand=+	GALSSELPO	non annotated		non-AUG	45	KGLFLLGQKISLTKQKALSSELPOIYWPVKRQMSKDSKFI
range=chr14:95645271-95645404 strand=+	SSLNLLMGR	non annotated		non-AUG	38	GFLFQLCLHYLHSSILNLLMGRTPPTKLTIVRVL
range=chrX:154159069-154159686 strand=+	PGDGSSEKVSYLASWR	NM_000132 F8_v1	antisense	non-AUG	119	WSCGLKCVISDRLLLLSIAPGDSSEKVSYLASWRSDKDSPCGVG LCKRSINKSLEETFCILGIVLCANQGSVFSMSFGVIALNCFCLV LGCLFEWELKLSGMLFLLSKYAEISS
range=chr13:98012703-98018647 strand=-	LVTIISR	non annotated		AUG	27	MQLAFICSSITATSEVLVTIISRAML
range=chr2:234384197-234469870 strand=-	AVNISAVR	NM_018218 USP40	CDS	non-AUG	24	KKNFRRSPAVNISAVRILGRRHGS
range=chr11:114325616-114326071 strand=+	HGDIFLK	non annotated		non-AUG	32	QSKTPSQKKKKKKQSQHTGDIFLKNSHKIK
range=chr1:161136224-161147286 strand=-	PLSYLDR	non annotated		AUG	50	MPSLSVPLSYLDRQMGSLVLIHHSYRHCSDGTQGTQRGSLSR EQLTEH
range=chr17:76001568-76001690 strand=+	VLVETHAFLVTQ	non annotated		non-AUG	17	KVLVETHAFLVTQELL
range=chr8:66567401-66567615 strand=+	FNFISKL	XM_003846486 LOC400682	3' UTR	AUG	31	MMKKWATMNTLNFNISKLVLHSDVNDVLF
range=chr3:10157357-10169022 strand=+	LTAVIMVGR	NM_018462 BRK1	3' UTR	non-AUG	61	FGEVKARPPGLKGFQTFLSVTVVYASHISLNSSTFYLTAVIMV GRGARDGFTYCTEM
range=chr16:15112097-15124265 strand=-	MCVRNSIQGWR	non annotated		AUG	17	MSMVCVRNSIQGWRHFWF
range=chr13:33080751-33086968 strand=-	PSIFLMFR	NR_026928 N4BP2L- IT2	non-coding	AUG	30	MLFSSTLPMPSIFLMFRSHLRCHFHCSTSS
range=chr12:57411725-57411927 strand=-	IFILNPR	non annotated		non-AUG	45	SRHSWKAERYTPLIKDILFNPRPPKRGKCHGTRCVAFLLSAPL
range=chr3:110639864-110640215 strand=-	HDM*VKIR	non annotated		non-AUG	20	LCDLLSKHDMVKIRRHDEG
range=chr21:34020720-34021204 strand=-	VSDM*NLNR	non annotated		AUG	42	MNFVSDMLNLRCSWSQSNMNLVLRKELWSQSSNSLQSMGIC
range=chr8:80942254-80942687 strand=+	EANEACLPFIWEWR	non annotated		AUG	30	MRGPPPRRALEANEACLPFIWEWRGEEET
range=chr8:124530101-124530292 strand=-	TYCLALCR	non annotated		non-AUG	27	NTYCLALCRKSLTPDVFNLHKKTGV
range=chr6:17759506-17762318 strand=+	LSAQSTPITR	non annotated		AUG	47	MLLIPTSPHPQHLLSAQSTPITRKLFLKGTGLLVLLWNPVLESTL
range=chr19:55649318-55653007 strand=+	AVDVLQDTR	non annotated		AUG	124	MYIDELQLQOVFLHALAVDVLQDTRAAARPRRWGQPPSTSP MQDGLNRFPGRYPAGGSNASCLTLISSKDRLPKDGSGEWGPA PFFFFLPGWSEMARSLTASASQQAIFLPQPPE
range=chr5:88014045-88018690 strand=+	GWDQVYGEENDLILF	MEF2C NM_001193350	3' UTR	AUG	49	GWDQVYGEENDLILFPCACVMSYIFRDVVFPLHKCYSFMFH GYLE
range=chr21:45105748-45115959 strand=+	APAQEAEEK	non annotated		non-AUG	134	RVPDRWSPTPKKEAETWSCARQWQPVHAGLASIAAGRPSHRP RRGGEQPHHAAPAQEAEEKGRARQPALWPAOPENSKFEKEE ENESDVKLGGAGRAGVRSWATPGSGKQWDLQFPEEAEEGRE RLCEV
range=chr1:31342371-31361123 strand=-	QGVPNR	SDC3, NM_014654.3	3' UTR	AUG	79	MWVQGHVPLPLPVMCESEGLWHLLAEQVGNRELLWLESW GLGCQSTRQARTWRAQTLSPGGQEARWGPCKDWSP
range=chr1:150343960-150344120 strand=+	RYPDLVIR	MPL, NM_005373.2	3' UTR	non-AUG	38	FFSIFRDRFSRWPWGSRYPDVLRPLPPKVLGLQV
range=chr15:49280838-49285216 strand=+	ELNTYIK	SECISBP2L	antisense	non-AUG	38	LKAKLQIITSKSYKNICWELNTYIKYVTVQNVEMKHKR
range=chr17:67992904-67996880 strand=+	SKINMLDR	non annotated		non-AUG	15	GERACSKINMLDRD
range=chr10:124768480-124817809 strand=+	IGITLKY	ACADSB, NM_001609.3	3' UTR	non-AUG	22	CKSELNIKNCVNTIDIGITLKY
range=chr7:32735224-32735731 strand=-	KMNISIQSNIVNE	non annotated		AUG	36	MEALFVPEKKTQTRPTNRMDKMNISIQSNIVNE

Table A3.1. (Continued) Full list of SEPs detected in Chapter 2.

range=chr4:74904262-74904922 strand=+	REAGELAGR	CXCL3	antisense	non-AUG	98	EQRHPOEPIAGGGGERGVGHGAQQTREAVREAGELAGRCL RPLCGSPRSANPFYAWLRLESPPERGGNSRSRSIRSPEGR RPRPRGGVGYG
range=chr2:17844495-17850489 strand=+	QIIQGAIK	SMC6	antisense	AUG	36	MLLDIQYILKFSMNPISIKQIIQGAIKFMSKIKSS
range=chr8:9423769-9423869 strand=+	ENFLDKFLKL	non annotated		non-AUG	18	GKYSYLEKENFLDKFLKL
range=chr13:70234870-70235374 strand=-	LVNCLLDLEIKENTGER	non annotated		non-AUG	55	KGAWLVNCLLDLEIKENTGERRSIECGFFTQETLQIRGTCPCQY FNVGSFYFP
range=chr3:163403983-163404349 strand=-	LEADEGYPFVEVR	non annotated		AUG	28	MRIIPALLEADEGYPFVRSWRPAWPIW
range=chr20:45298911-45299385 strand=-	PNVQVVIK	non annotated		AUG	77	MLLYGVTLAAKRECLKTSPNVQVVKSLFASHLLMHWPKQVTW PSPESVWEGTTQGCQEYQAWVCGGHQYIPYTS
range=chr1:83599137-83599370 strand=+	DLLSKVFQTLK	non annotated		non-AUG	19	KEFDLLSKVFQTLKLHICH
range=chr6:99387016-99387302 strand=-	LENLM*PAKDGLIIVERSL	non annotated		non-AUG	33	SKKVKDRKLENLMPAKDGLIIVERSLALCMLR
range=chr20:42295716-42345131 strand=+	ISDPHLVLR	MYBL2, NM_001278610.1	3' UTR	non-AUG	110	GCEADVVHTAQVLSILADNCPFKLFQPHVRYRQRQQAQPLLAG QAREGSSGPEAPKPLHDTCYVQCLEDGGLRDQGPFAHAGESP AAPGPPEAQPHISDPHLVLR
range=chr8:125467607-125470995 strand=-	FPTRAIQQVLKEN	non annotated		non-AUG	29	ASRTLQSFIFGIRCLFPTRAIQQVLKEN
range=chr2:239152640-239197306 strand=-	PTWSRCGTNKVLR	PER2 NM_022817	CDS	non-AUG	69	RPTWSRCGTNKVLRVSAFCWQRECTLVMKPLEFLKREFLOPP IHQVCSRMWVKGRSLSWATYLR
range=chr2:47600600-47601323 strand=-	DYLFKMSHT	non annotated		non-AUG	19	SSRVKDYLFKMSHTYFVT
range=chr4:86297867-86298058 strand=-	EGVREMGR	non annotated		non-AUG	36	AYQGREYRQRKGPLLFLQVQHEGVREMGRKPKHD
range=chr5:44812872-44828662 strand=-	SLFLANK	GATA6-AS1	antisense	AUG	71	MVYCFITIPRMLFRISNRILPESQPWRSYTFNVRPSSLPKYTH VFSFLANKSPLOKSKRLNSKFN
range=chr1:1424437-1425809 strand=-	VDGTVELLR	non annotated		AUG	77	MQSHQALPLLLLRQVEDHHVNAADGTVELLRQIAGQDQHEVSG QGWGRKRWGCQLRCEPENGSSAPSGRFLFGR
range=chr2:205660031-205660209 strand=+	EDGLIWNQR	non annotated		non-AUG	22	LVTFEDGLIWNQRKQELPLY
range=chr2:10244681-10248113 strand=+	M*VQLKLEVEVP	non annotated		non-AUG	26	ECKMVQLKLEVEVEPEKIKHRVFMFY
range=chr11:77319073-77325243 strand=-	AASSKEVNTDESSAAGVFHM*R	AOP11 NM_173039	3' UTR	non-AUG	29	SPSDENYKAASSKEVNTDESSAAGVFHM
range=chr3:128996926-128997117 strand=-	ENALEGCSPM*LR	non annotated		non-AUG	42	LKVGKIAWLEFKSLPLPGKPEILHQPTENALEGCSPLMR
range=chr2:231589548-231590185 strand=+	EPNGIKR	BANF1, NM_003860.3	antisense	AUG	26	EPNGIKRIESEGGEGWGRESIQESW
range=chr1:67751219-67751585 strand=+	FELEPEQDCKQ	non annotated		AUG	27	MPMEPALVAFCFDEFELEPEQDCKQ
range=chr1:150317600-150318971 strand=-	FLGGYVK	WDFY3, NM_014991.4	3' UTR	non-AUG	31	LWTFILYCRNQYLFLGGYVKDLKNYYLVCSL
range=chr22:22308818-22337181 strand=-	GLQQVQR	TOP3B, NM_003935.3	CDS	non-AUG	44	ERGAGAGPHLGPQVEGGLQQVQRGSALLRERPPRAGVRRHLQC L
range=chr1:195789567-195789749 strand=+	GNITHQILLGGMWQGE	non annotated		non-AUG	19	GNITHQILLGGMWQGEEDH
range=chr8:30535597-30585324 strand=-	IDFLNIR	GSR, NM_001195104.1	CDS	non-AUG	18	LLAENPIDFLNIRKIPN
range=chr7:22510284-22510452 strand=+	ISTIVVR	non annotated		non-AUG	29	GRIESWNRKTGEKHRVCSLISTIVRLLIS
range=chr5:10158411-10158577 strand=+	KGQEVLR	non annotated		AUG	24	MAMGKQEVLRWEVSALTGLSGGL
range=chr1:24034077-24035038 strand=-	LPIISDPLLLL	non annotated		non-AUG	26	HPTQPAQRLPISDPLLLLPSHTAWL
range=chr1:212531453-212535680 strand=-	M*LLNPPM*K	PPP2R5A, NM_006243.3	antisense	AUG	21	MLLNPPMKIFHYKTVFHTTAS
range=chr8:56636872-56637117 strand=-	MQDKLAKIFLN	non annotated		AUG	63	NCPAHGKPYLYVFGIILLIYCCGYCCTLCVRMQDKLAKIFLN GTYHLPDVTFCCKDSEL
range=chr6:4823265-4824846 strand=+	NCNILSWLPYINKE	non annotated		non-AUG	35	TPPISIESRGEKCRQQLFLTNCNILSWLPYINKE
range=chr15:41849141-41849860 strand=+	NYNKLLTLFSM	non annotated		non-AUG	29	NCTRTADGCELPRLLNKINYNKLLTLFSM
range=chr9:100689089-100700593 strand=-	SFLVNAVQLLM	HEMGN, NM_018437.4	3' UTR	non-AUG	19	THTFCGVYSFLVNAVQLLM
range=chr16:89590029-89594505 strand=-	SISWTKSFQEMLPPLVLR	SPG7 NM_003119	CDS	non-AUG	81	TSTSLSGTTWTRCTSPLASISWTKSFQEMLPPLVLRFRRLMTTA MTSKVRSRSRYIWSSRSRLLSSEPAEKQKGETNT
range=chr12:69813707-69814545 strand=+	STNSLLSTNSGVEKVG	non annotated		non-AUG	21	RSTNSLLSTNSGVEKVGCHCL
range=chr3:112539900-112542022 strand=+	TLQIINHRL	non annotated		non-AUG	37	TLQIINHRLRNSRLNTEVREKQDLKWTLLVCGIE
range=chr1:110882061-110914912 strand=+	VCLAIIISERDALK	non annotated		non-AUG	32	KSGQFSYFVFCCLAIIISERDALKVAGRQTDLL
range=chr9:95194519-95194749 strand=+	VSQDGAIALQPGRQER	non annotated		non-AUG	56	TWEAEVAVSQDGAIALQPGRQERNISKKRKKKENGNTLSLIAFK GLALCKDL
range=chr21:47685600-47693148 strand=+	VTEVLLR	non annotated		non-AUG	121	DVSLVETEONIVALELRFAGTEDTLVSQLVHLLQALQALGHVLDL GIKGGHGLLTHEVTVDPKPGALLNQGHVQGVTEVLLRDILATGE GEDHTHIPMQRGNREKDHATTSISHSI
range=chr4:145030055-145038548 strand=+	AKVLFCL	non annotated		non-AUG	20	ELRKAKVLFCLFLFYLYNY
range=chr12:94688630-94690814 strand=+	DGVFVLK	non annotated		non-AUG	57	LYRQDNWGLLWGNPPSIWGCGEAILQRSFLKGDGVFVLKSG RAWLFGKDWAF
range=chr1:53384961-53385662 strand=-	FDVHIR	non annotated		AUG	58	MTVORRRHDEITHGALTGNELGSFVHIRSCRFLPLGLMAQEHC YILPAAFELEVGW
range=chr14:59925205-59931181 strand=-	FGCLDCK	non annotated		non-AUG	33	LLELAPLTCFGCLDCKWKSILITHGAPPHPTP

Table A3.1. (Continued) Full list of SEPs detected in Chapter 2.

range=chr5:27717055-27717588 strand=-	HHLLQRLFFLSWN	non annotated	non-AUG	38	HHLLQRLFFLSWNFLGFIVANQLTINLEFVSGLRILFH
range=chr2:160547202-160547607 strand=-	KVIIIIM*ENTK	non annotated	non-AUG	53	CIGICITYTHLTKVIIIMENTKDIFTSTIRSKARISSNYAIIHHITGNISQY S
range=chr14:23741387-23746392 strand=+	LLGLESLLFLQ	non annotated	non-AUG	22	NGVGPDGRLLGLESLLFLQMTS
range=chr14:57430195-57430368 strand=-	M*PSEVTVTAIAR	non annotated	AUG	16	QMPSEVTVTAIARDDD
range=chr1:165664398-165664673 strand=+	NLRAMSEFWKEINWNIPLLSWQPPK	non annotated	non-AUG	51	CCNKVELQTSQRNLRAMSEFWKEINWNIPLLSWQPPKGFGNAPC QIFDSIF
range=chr8:126009798-126010145 strand=-	RPETLGVGAR	non annotated	non-AUG	42	EEKERTLQVLRLEIHGCAQPVLSRPETLGVGARRQRSQIVVS
range=chr6:88240174-88240816 strand=+	SEGMLEQKSQIKM*K	non annotated	non-AUG	20	KSEGMLEQKSQIKMYENIT
range=chr6:148065557-148066827 strand=+	VSVTIDEIR	non annotated	non-AUG	50	NLIKITVCACVCVSVTIDEIRLFMSVKIEAGRWWCEDSVCHSFYNTF LYV

Table A3.1. (Continued) Full list of SEPs detected in Chapter 2.

MCF10A						
Genomic coordinate	Detected peptide by shotgun proteomics	mRNA RefSeq annotation	Location	Start Type	Length	Protein Sequence
chr14:61744089-61748530 strand=-	RCKPSAKISAGGR	TMEM308, NM_031017970.2	5' UTR	Non-AUG	26	GSEIFKPPGPRCKPSAKISAGGRSAF
chr13:29283091-29289675 strand=-	SSPWATLFLR	Not Annotated		AUG	138	MPTGSSPHLLPSRLNSAHLGPCNSQPSGLMSSKGLDDPLPQQPPSPG GLCPVCSVDQGRSQGATPSKGGGWTLLESGGAVPHVMTLSEGG ARGRAEWGWAPAGHQLGRASSPWATLFLRGLTKSENPKCVRS
chr14:75745481-75748937 strand=+	LVLDFIK	FOS, XM_005267488.1 FOS, NM_005252.3	3' UTR	Non-AUG	37	SIGFIIGINLVDLDFIKLYLVQLLITITVFLAIVCSD
chr17:21729873-21731760 strand=+	GGTLDSDYNIQKESTLHLVLR	Not Annotated		Non-AUG	120	RRHPPDQORLIFAGKQLEDGGTLDSDYNIQKESTLHLVLRGGMKIFVK TLTKITLLEVEPSDTIENVKAKIQDEEGIPDQORLIFAGKLEDGRTLS DYSIQKESTLHLVLRGGC
chr6:121400627-121655644 strand=-	HLQIALR	TRC1D32, XM_005268861.1 TRC1D32, NM_152730.4	CDS	AUG	11	MRHLQIALRQL
chr2:160958233-161056589 strand=-	ASGSYWCHFMIVK	ITGB6, NM_001282355.1 ITGB6, XM_005246537.1 ITGB6, NM_000888.4 ITGB6, NM_001282388.1 ITGB6, NM_001282353.1 ITGB6, NM_001282389.1 ITGB6, NM_001282390.1 ITGB6, NM_001282354.1	CDS	Non-AUG	58	GFPWLFSSSGLSYCASGSYWCHFMIVKLPNLKQNDQKPSGKREPIHS TEDPQVLKLM
chr20:42295709-42345122 strand=+	ISDPHLVLR	MYBL2, NM_001278610.1 MYBL2, NM_002466.3	3' UTR	Non-AUG	110	GCEADDVHTAQVSILADNCPFKLFQPHVRYORRQQAQPLLAGOAR EGSSGPEAPKPLHDTCPYVQCLEDDGLRGDQGFAPFHAGESPAAGPP EAQPHISDPHLVLR
chr17:77751977-77761449 strand=+	QVGAGWGLNVKRWQ	CBX2, NM_005189.2	3' UTR	Non-AUG	25	QVGAGWGLNVKRWQQAAGVERAQOP
chr10:124768429-124817806 strand=+	IGITLKY	ACADSB, NM_001609.3	3' UTR	Non-AUG	22	CKSELNKNVCNTIDIGITLKY
chr16:56642478-56643409 strand=+	GVGVQQLLR	MT2A, NM_005953.3	3' UTR	AUG	51	MQRVQMHLLEKLLLLPGLCQVCPGLHQRGVGVQQLLRMLGQP RSQM
chr2:160802326-160919126 strand=-	NDLNKNTPVK	PLA2R1, NM_001007267.2	3' UTR	Non-AUG	46	FCLSNDFFKSIIYIGIFQNSNCHIKKHDNLNKNLNTVPKIKRQL
chr2:190526125-190535557 strand=+	NILDELKKEYQEIENLDK EYQEIENLDKTK TEMLSESK QQQNSNIFLADRTEMLSESK	ASNSD1 (NM_019048)	5' UTR	AUG	96	MPSRGTRPEDSSVLIPTDINSPTHKEDLSKSIKEQIVDELSNLKKNRV YRQQQNSNIFLADRTEMLSESKNILDELKKEYQEIENLDKTKIKK
MDA-MB-231						
Genomic coordinate	Detected peptide by shotgun proteomics	mRNA RefSeq annotation	Location	Start Type	Length	Protein Sequence
chr14:51800111-51832275 strand=+	EIKNPAR	LINC00640, NR_039358.1	non-coding	Non-AUG	56	ILSLGWKLFLEIKNPARTVPKQMAALQVRRSLGP
chr17:35990853-36002800 strand=-	SVLTLPLLR	Not Annotated		Non-AUG	50	RTYVLPQTIFLCPPHVSVTLPLLRGNCYSDFYCCILLPLLEFYMSGITH
chr20:13370036-13619583 strand=-	MKEYVK	TASP1, NM_017714.2	3' UTR	AUG	22	MDYIMKEYVKRTYNSAQSLICV
chr8:146220251-146224283 strand=+	LLSRIPWYGCTTVCLTIHLKDK	TMED10P1, NR_002807.3	non-coding	Non-AUG	37	LLSRIPWYGCTTVCLTIHLKDKRVVFSFLPLLIKLL
chr18:31158541-31327399 strand=+	SRLLPVITLLSRHVAQLLSKVK	ASXL3, XM_00528356.1 ASXL3, NM_039632.1	CDS	AUG	44	MSRLLPVITLLSRHVAQLLSKVKQTQPAISITQVTGFAGMMMG
chr20:42295709-42345122 strand=+	ISDPHLVLR	MYBL2, NM_001278610.1 MYBL2, NM_002466.3	3' UTR	Non-AUG	110	GCEADDVHTAQVSILADNCPFKLFQPHVRYORRQQAQPLLAGOAR EGSSGPEAPKPLHDTCPYVQCLEDDGLRGDQGFAPFHAGESPAAGPP EAQPHISDPHLVLR
chr1:43803475-43820135 strand=+	RYPDLVIR	MPL, NM_005373.2	3' UTR	Non-AUG	38	FFSIFRRDRFSPRWPWSRYPDVIRLPLPKVLGLQV
chr9:470294-746103 strand=+	IFSLALMEGMKQLQ	KANK1, NM_015158.3 KANK1, NM_001256877.1 KANK1, XM_005251410.1 KANK1, NM_001256876.1 KANK1, XM_005251418.1 KANK1, XM_005251417.1 KANK1, XM_005251415.1 KANK1, XM_005251419.1 KANK1, XM_005251412.1 KANK1, NM_153186.4 KANK1, XM_005251411.1 KANK1, XM_005251416.1 KANK1, XM_005251414.1	CDS	AUG	76	MVTKIQMAQKRIFSLALMEGMKQLQVMPAQMKALLPSQMTSVMSLSI LLKRRRRRMTLGEWQKGTMLQILKV
chr1:186265405-186283688 strand=+	ENMEILK	PRG4, NM_001127710.1 PRG4, NM_005807.3 PRG4, NM_001127708.1 PRG4, NM_001127709.1	3' UTR	Non-AUG	22	HKYNLKVILENMEILKFTFTS
chr10:76871393-76941881 strand=+	RAIDIER	SAMD8, NM_001174156.1 SAMD8, NM_144660.2 SAMD8, NM_005269540.1	3' UTR	Non-AUG	18	LCRAIDIERPNNRLQWGL
chr11:65686728-65689048 strand=+	DAEQEEVQR	DRAP1, NM_006442.3	5' UTR	Non-AUG	138	GPASGVLTTGGARTPRETAGAARAAPGRGTATGRRRAGPAGRLRAA ALDPTRRERPRDAEQEEVQRAVPAGAQEDHADGRDRWEGGGGA CHILPGARALPRVAEEGLPSDPAEREDHDHPEAVHRAAAV
chr20:306239-310867 strand=+	RSFSGNPAVEEGTTQEN	SOX12, NM_006943.3	3' UTR	Non-AUG	19	EVRSFSGNPAVEEGTTQEN
chr1:228395861-228566575 strand=+	GGAARPARATDGAPRAGGGGQGGAGHP PGQSPILR	OBSCN, NM_001098623.2 OBSCN, XM_005273287.1 OBSCN, XM_005273288.1 OBSCN, XM_005273289.1 OBSCN, XM_005273290.1 OBSCN, XM_005273291.1 OBSCN, XM_005273292.1 OBSCN, XM_005273293.1 OBSCN, XM_005273294.1 OBSCN, XM_005273295.1 OBSCN, XM_005273296.1 OBSCN, XM_005273297.1 OBSCN, XM_005273298.1 OBSCN, NM_001271223.2 OBSCN, XM_005273299.1 OBSCN, XM_005273300.1 OBSCN, XM_005273301.1 OBSCN, XM_005273302.1 OBSCN, XM_005273303.1 OBSCN, XM_005273304.1 OBSCN, XM_005273305.1 OBSCN, XM_005273306.1 OBSCN, XM_005273307.1 OBSCN, XM_005273308.1	CDS	Non-AUG	83	ARCPGPGQAAPGPAAPAEGRVHCAGGAARPARATDGAPRAGGGGR QGGAGHPGQSPILRCDPPAALWHLPWPQPLPQT
chr15:65488337-65503840 strand=-	LDPGQDARAHISSETLVQSRHRAVGGGR	CILP, NM_003613.3	CDS	Non-AUG	29	LDPGQDARAHISSETLVQSRHRAVGGGR

Table A3.1. (Continued) Full list of SEPs detected in Chapter 2.

chr17:3827163-3867758 strand=-	EAASTVRTAQTAGR	ATP2A3, XM_005256658.1 ATP2A3, NM_174953.2 ATP2A3, NM_174954.2 ATP2A3, NM_174956.2 ATP2A3, NM_174955.2 ATP2A3, XM_005256657.1 ATP2A3, NM_005173.3 ATP2A3, NM_174957.2 ATP2A3, XM_005256656.1 ATP2A3, NM_174958.2	5' UTR	Non-AUG	37	PPGGEAASTVRTAQTAGRAAWRRRRICSRPPTCCATSR
Tumor						
Genomic coordinate	Detected peptide by shotgun proteomics	mRNA RefSeq annotation	Location	Start Type	Length	Protein Sequence
range=chr1:78409737-78425885 strand=+	KILEM*ELIQ	Not Annotated		Non-AUG	17	EIHKLLKILEMELIQE
range=chr1:78409737-78425885 strand=+	ILEM*ELIQ					
range=chr19:522521-522833 strand=-	AIVELVK	Not Annotated		Non-AUG	32	FIGSLITYPKPSSVLGNWETGVVSAIVELVKI
range=chr12:115108455-115109381 strand=-	CNYLITLVNRWNNAWK	TBX3, NM_005996.3 TBX3, NM_016569.3	3' UTR	Non-AUG	24	TKYLCNYLITLVNRWNNAWKIL
range=chr2:99862519-99863895 strand=+	FQYM*EINTAG	Not Annotated		Non-AUG	37	FQYMEINTAGQSLERMVENEKLLVRLVRIVFAHF
range=chrX:100353186-100424873 strand=+	NYLDKQMLILVNNDFKAA	Not Annotated		Non-AUG	58	NVQCSKNETCKEIGTCDSYTGKTRNLYDKDQMLILVNNDFKAAISIFKE IKATMFN
range=chr17:40970559-40972008 strand=+	QPQNYM*ALVLITRLDK	Not Annotated		Non-AUG	92	SSSPKQPQNYMALVITLITRLDKPQLRKKVWLSKQSLCMWTKKGPITQT NLKSHLIFFLWGTAFPNVYTLVWLNLFSTLSSQGQRTCEPKS
range=chr11:14665294-14895861 strand=+	KSTLLNNHIK	PDE3B, XM_005252972.1 PDE3B, NM_000922.3 PDE3B, XM_005252971.1	3' UTR	Non-AUG	43	IISYKSTLLNNHIKIARKWIFTMFKLASVWICTLLVKYI
range=chr14:27991709-27992439 strand=+	KYNLQMIQ	Not Annotated		Non-AUG	41	KYNLQMIQCHFIDSFLLFTLYSIPLCGPNMIYSTLLCMSL AAGGQAACGLAQSRAAGGGWTOGHTGCGSEQAAGRGGERRKGV LDGTEPARWAPGTSCGRRRAGRVGAGHRARSGRSPWRELPALPR RPRVEREPGRERAGLVGRVQGARPGPWSWSSCAGRHTESLPARG 138P
range=chr4:1803098-1810596 strand=-	RAGRVGAGH	Not Annotated		Non-AUG	138	
range=chr2:74257245-74259066 strand=+	KKETGLLVQ	Not Annotated		AUG	41	MGLSCIQVEVCLVAGKQSRKKTGLLVQAEWRGLDWRGGF
range=chr17:37682110-37691204 strand=-	QDVWFK	Not Annotated		AUG	29	MVCLVRFPSARTSWPPHTVQDVWFKMHH
range=chr17:37710622-37710961 strand=+	QGPILDVK	Not Annotated		AUG	52	MHRLHFSDRGNRVTFEVTGPERGQKVDQENHSLGLQSQGPILD 52VKSNS
range=chr7:757156-757313 strand=+	KVAGSPEHK	Not Annotated		Non-AUG	27	WWCTPVVPAIQEAKVAGSPEHKSSRLQ
range=chr3:14520433-14530859 strand=-	TLLASRLPT	Not Annotated		Non-AUG	73	TLLASRLPTMHAHKLHNMWEGTMAQRGQVTCRLRSKLTITIFGL 73APPTSPGPGKPLPQGTAPQSGRGG
range=chr17:35684526-35684974 strand=-	LAFIGQR	Not Annotated		Non-AUG	15	LAGLAFIGQRTKQD
range=chr16:8742491-8744671 strand=+	VATVSNIL	Not Annotated		Non-AUG	22	QTRVATVSNILCISKLEERT
range=chr6:104975101-104975201 strand=-	KDLSFAIQINPWPQD	Not Annotated		Non-AUG	16	KDLSFAIQINPWPQDI
range=chr6:31606802-31611451 strand=+	VALSSLPR	Not Annotated		Non-AUG	21	PGVALSSLPRCQRGQMVAPG
range=chr19:23921996-23938778 strand=-	TLKCKMQDEN	ZNF681, XM_005259770.1 ZNF681, NM_138286.2 ZNF681, XM_005259766.1 ZNF681, XM_005259767.1 ZNF681, XM_005259788.1 ZNF681, XM_005259769.1	3' UTR	AUG	39	MRDLLLLGVHYLRPFLVKELRKLCKMQDENVSGEALSG
range=chr9:2039912-2046006 strand=+	LVTDITK	Not Annotated		AUG	24	MICYTSQLVLYHLVTDITKHWFYF
range=chr16:67963770-67967286 strand=-	EGGQSGQKK	Not Annotated		Non-AUG	75	AWFSSAPPGVSGPGPALIQPWEQLSSQQPREGARLAGRDEGGQSG 75KQKQDMPGQWSREQACRAGRQGEVGVWSHK
range=chr17:68010245-68038237 strand=+	ENLLPRNI	Not Annotated		Non-AUG	52	TYFINLSILIFYYSFIREFVEHLFYVIDTRNEIVKENLLPRNIHSGIKV
range=chr6:143771955-143810369 strand=+	KTLLLMCMK	PEX3, XM_005267181.1 PEX3, NM_003630.2	3' UTR	Non-AUG	45	KTLLLMCMKLVPLSNWRNDFSFKNYSIGIHLFKIHWVNHLYLE
range=chr20:42295716-42345131 strand=+	ISDPHLVLR	MYBL2, NM_001278610.1 MYBL2, NM_002466.3	3' UTR	Non-AUG	110	GCEADDVHTAQVLSILADNCPFKLFQHPVRYQRRLAQPLLAGQAR EGSSGPEAPKPLHDTCPYVQCLEDDGLRGDQGFAPFHAGESPAAPGPP 110EAQPHISDPHLVLR
range=chr17:20056612-20057839 strand=+	QSENVK	Not Annotated		AUG	22	MQQRQERQSENVKMKCAGLD

Table A3.1. (Continued) Full list of SEPs detected in Chapter 2.

