View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DSpace@MIT

Journal of proteome research

Article

pubs.acs.org/jpr

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

Jiao Ma,[†] Carl C. Ward,[†] Irwin Jungreis,[§,∥] Sarah A. Slavoff,[†] Adam G. Schwaid,[†] John Neveu,[‡] Bogdan A. Budnik,[‡] Manolis Kellis,[§,∥] and Alan Saghatelian*[,†]

[†]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, United States

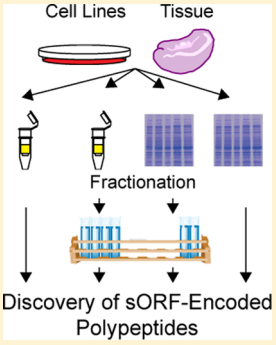
[‡]MSPRL, Center for Systems Biology, Harvard University, 52 Oxford Street, Cambridge, Massachusetts 02138, United States

[§]MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, United States

[∥]The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02139, United States

Supporting Information

ABSTRACT: The existence of nonannotated protein-coding human short open reading frames (sORFs) has been revealed through the direct detection of their sORF-encoded polypeptide (SEP) products. The discovery of novel SEPs increases the size of the genome and the proteome and provides insights into the molecular biology of mammalian cells, such as the prevalent usage of non-AUG start codons. Through modifications of the existing SEP-discovery workflow, we discover an additional 195 SEPs in K562 cells and extend this methodology to identify novel human SEPs in additional cell lines and human tissue for a final tally of 237 new SEPs. These results continue to expand the human genome and proteome and demonstrate that SEPs are a ubiquitous class of nonannotated polypeptides that require further investigation.



KEYWORDS: *short open reading frames (sORFs), sORF-encoded peptides (SEPs), peptidomics*

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 nonannotated transcripts has demonstrated the existence of functional noncoding RNAs, such as long intergenic noncoding RNAs (LINCs).[7,8] The identification of additional RNAs also raises the possibility that there may also exist additional nonannotated 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] Frith 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 of 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 nonannotated 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 frequently used,[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 reveals the existence of nonannotated 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

Received: December 21, 2013
Published: February 3, 2014

 © 2014 American Chemical Society dx.doi.org/10.1021/pr401280w | J. Proteome Res. 2014, 13, 1757−1765

product. Because of transcript amplification and the number of reads per sequencing, experiment ribosome profiling is more sensitive and will identify the largest number of sORFs, but the bias of MS toward 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 fewer than 150 amino acids in length).[18] To improve on these results, Slavoff and coworkers utilized next-generation RNA sequencing (RNA-Seq) to identify all possible protein-coding mRNA transcripts, including nonannotated 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 the 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 nonannotated cellular polypeptides. Recent work from others has also supported this conclusion, with Vanderperre and colleagues having characterized 1259 nonannotated 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.

MATERIALS AND METHODS

Cell Culture

K562 cells were grown in RPMI1640 medium supplemented with 10% FBS, penicillin and streptomycin at a density of (1 to 10) × 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% CO_2.

Tissue Sample

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

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 μL) 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 the samples were cooled on ice, the cells were sonicated on ice for 20 bursts at output level 2 with a 30% duty cycle (Branson Sonifier 250; Ultrasonic Convertor). 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*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 an 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: 2−5, 5−10, and 10−15 kDa. 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.

Peptidome Isolation from Tissue

Frozen human breast tumor sample (∼200 mg) was immersed in boiling water (200 μL) for 10 min. 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 μL of ice-cold acetic acid (0.5% v/v). The aqueous fraction and the homogenate were combined and centrifuged at 20 000*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 (<15 kDa) were analyzed by LC−MS/MS, as described later.

ERLIC Fractionation[20,21]

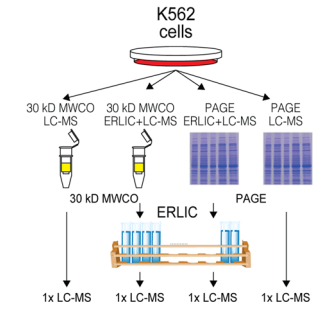
After trypsin digest the samples were dried in a speed vac and suspended in ERLIC buffer A (90% acetonitrile 0.1% acetic acid). Samples were then fractionated using an HPLC (Agilent 1200 HPLC) equipped with an ERLIC column (PolyWAX LP Column, 200 × 2.1 mm, 5 μm, 300 Å (PolyLC)). Samples were separated 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.

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 and 1600 m/z ions were collected at 60K resolving power for the MS1, and these data were 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 s was used throughout the LC−MS/MS method. Samples were trapped for 15 min with flow rate of 2 μL/min on a trapping column 100 μm ID packed for 5

dx.doi.org/10.1021/pr401280w | J. Proteome Res. 2014, 13, 1757−1765

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 1. Workflows tested in the discovery of novel human SEPs. (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 30 kDa 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) Number of total SEP and novel SEPs identified in K562 cells using each of the four different SEP discovery workflows.

cm in-house with 5 μm Magic C18 AQ beads (Waters) and eluted onto 20 cm × 75 μm 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 min (A: 0.1% formic acid in water; B: 0.1% formic acid in acetonitrile).

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 were 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 ppm, 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 <3.5 ppm. After analysis, the data were 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 that passed the search criteria was then searched against the Uniprot human (SwissProt) protein database using a string-searching algorithm. Peptides found to be identical and to overlap with part of annotated proteins were eliminated from the list. The remaining peptides were then searched one more time against nonredundant 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 noncoding. 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., nonannotated RNAs).

RNA-Seq Library Preparation and Transcriptome Assembly

Total RNA (3000 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]

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 predicted transitions for each peptide, and we used 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 min gradient and targeted MRM tandem mass spectrometry using the aforementioned Skyline method. Samples were trapped for 15 min with flow rate of 2 μL/min on a trapping column 100 μm ID packed for 5 cm in-house with 5 μm Magic C18 AQ beads (Waters) and eluted with a gradient to 20 cm × 75 μm 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 min (A: 0.1% formic acid in water; B: 0.1% formic acid in acetonitrile).

RESULTS AND DISCUSSION

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

dx.doi.org/10.1021/pr401280w | J. Proteome Res. 2014, 13, 1757−1765

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

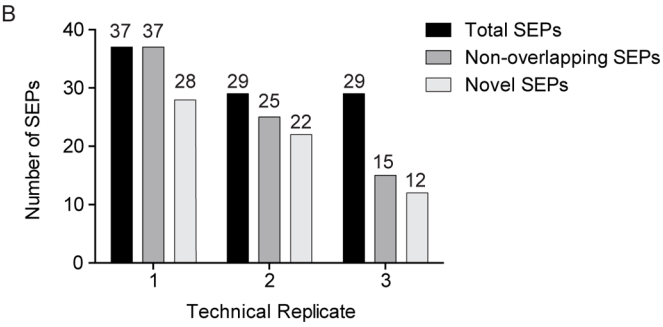


Figure 2. Biological and technical replicates lead to the discovery of novel SEPs. (A) Number of SEPs detected in four biological replicates of K562 cells. Each of these samples was analyzed using the PAGE+LC−MS SEP discovery workflow. For each replicate, the detected SEPs include the total number of SEPs identified as well as the novel SEPs that were characterized for the first time. (B) Three technical replicates of biological replicate #4 from panel A were performed using the PAGE+LC−MS workflow with K562 peptidome. The total number of SEPs detected in each run (black), nonoverlapping 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).

interaction chromatography (ERLIC),[20,21] followed by LC−MS/MS (Figure 1A). 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 (2013), Vanderperre and colleagues have identified 1259 nonannotated polypeptides.[19] Below, total SEPs refer to the total number of SEPs discovered and novel SEPs refer to any SEPs from these groups that were not identified in the Slavoff et al. or Vanderperre et al. manuscripts.

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 and 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 ≤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 nonannotated 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 previously identified (Figure 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 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 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 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).

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 data-dependent data acquisition does not completely sample all peptides in a sample and therefore does not provide information on all ions.[17] On the basis of models of this process, they determined that for yeast-cell-soluble lysate 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 detect only 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 was 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 1). An average of 22 SEPs were detected per run with a range between 11 and 37 SEPs in each sample (Figure 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 1), 6 from MWCO + LC−MS (Figure 1), and 61 from four PAGE+LC−MS biological replicates (Figure 2)). The lack of overlap between samples is consistent with our previous observations

dx.doi.org/10.1021/pr401280w | J. Proteome Res. 2014, 13, 1757−1765

and supports the idea that SEP detection is variable, as predicted from proteomics studies.[17]

Next, we tested the impact of performing technical replicates. We analyzed biological replicate #4 (Figure 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 2B). Of the 29 SEPs identified in the second run, 25 were not detected in the first run (i.e., nonoverlapping SEPs), and of the 29 detected in the third run, 15 were not detected in the first or second runs (Figure 2B). The number of novel SEPs identified per run decreased from 28 to 12 as more runs were performed, but there was still a substantial number 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 or technical replicates to increase the number of SEPs discovered. Additionally, we also performed five more technical replicates (using biological replicate #3 from Figure 2A) and detected 48 SEPs (with 32 of these being novel SEPs) (Supporting Information, Figure 1). At this point, we had identified a total of 195 novel SEPs (i.e., not identified in Slavoff et al. or Vanderperre et al.) in K562 cells through a combination of different workflows and biological and technical replicates.

Three to four biological/technical replicates matched 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 was detected among biological or technical replicates repeatedly, such as ASNSD1-SEP and CIR1-SEP. ASNSD1-SEP is the most frequently SEP and therefore is likely to have high cellular concentration and stability. ASNSD1-SEP also shows an unmistakable evolutionary signature of protein coding regions (Supporting Information, Figure 2), as measured across 29 eutherian mammals by PhyloCSF.[28] In total, 195 novel SEPs represent a >200% increase from the previous study and also the largest number of SEPs ever reported from a single cell line.

Using Targeted LC−MS/MS to Rapidly Validate Novel SEPs

In the majority of cases, a single peptide is used to identify an 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 coded 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 of 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 MRM experiments,[27] was used to define the MRM transitions for peptides derived from 105 SEPs. These SEPs included the 81 from the PAGE+ERLIC +LC−MS and 7 from MCWO+LC−MS (Figure 1) as well as 17 SEPs from the biological replicates #1 and #2, which were identified by PAGE+LC−MS (Figure 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%) (Supporting Information, Table S1). Skyline analysis of PRR3-SEP (Figure 3), for example,

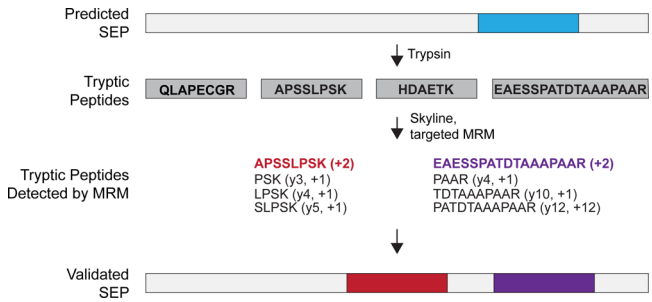


Figure 3. Validating SEPs with targeted mass spectrometry. Analysis of PRR3-SEP by Skyline and subsequent MRM targeted LC−MS identifies additional peptides from this SEP. The tryptic peptide (blue box) that was detected in the original shotgun proteomics experiment led to the initial identification of the PRR3-SEP. To identify additional peptides from PRR3-SEP, we used Skyline 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.

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 (Supporting Information, Figure 3 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.

Using molecular biology and peptide synthesis we had previously validated 17 out of 86 novel SEPs (20%) by expression or coelution over several weeks.[12] Here we validated 36 out of a 105 SEPs (34%) by identifying a second peptide in approximately 1 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.

Overview of the 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

dx.doi.org/10.1021/pr401280w | J. Proteome Res. 2014, 13, 1757−1765

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 toward shorter lengths. The SEPs range between 8 and 134 amino acids long, with the majority (>90%) of new SEPs being <100 amino acids long (Figure 4A).

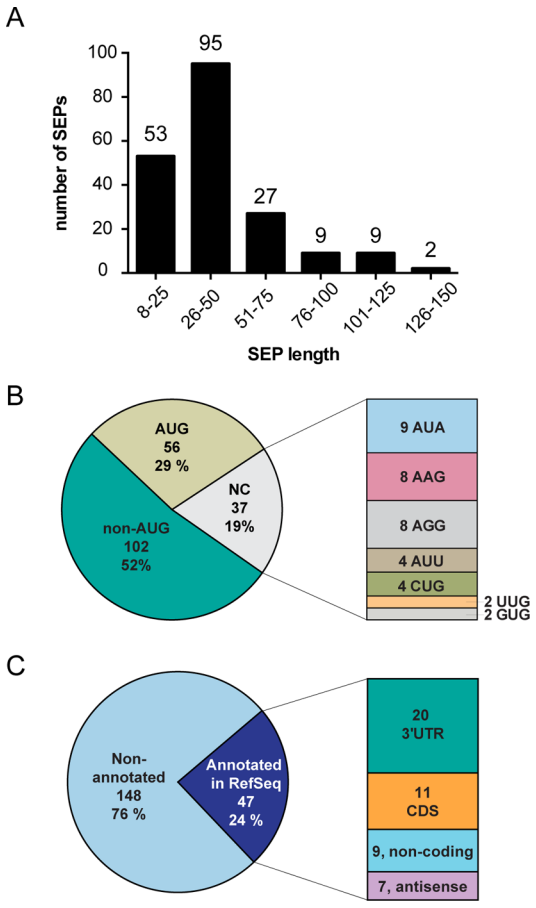


Figure 4. Overview of 195 novel SEPs identified in K562 cells. (A) Length of each SEP was determined using a defined set of criteria (see Methods), and the length distribution reveals that the majority (>90%) of SEPs discovered are between 8 and 100 amino acids. (B) SEPs utilize AUG, near cognate codons (i.e., one base away from AUG), and unknown codons to initiate translations. (C) SEPs are primarily derived from nonannoated 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 and CDS and on noncoding RNAs such as antisense RNAs and noncoding RNAs.

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 was 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, then no start codon was assigned. Many of these SEPs (~70%) do not appear to initiate with an AUG codon (Figure 4B).

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 SEPs we discovered, 47 in total, are derived from RefSeq RNAs (Figure 4C) A breakdown of the distribution of these SEPs on the RNAs reveals that a majority are found on the 3′-UTR. We counted sORFs in the 3′-UTR only if there was 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 were not alternative exons.

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] Because 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.), noncoding 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 noncoding 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 nonannotated 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 (Supporting Information, Figure 4) as we do for the 195 SEPs (Figure 3), 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.

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]

dx.doi.org/10.1021/pr401280w | J. Proteome Res. 2014, 13, 1757−1765

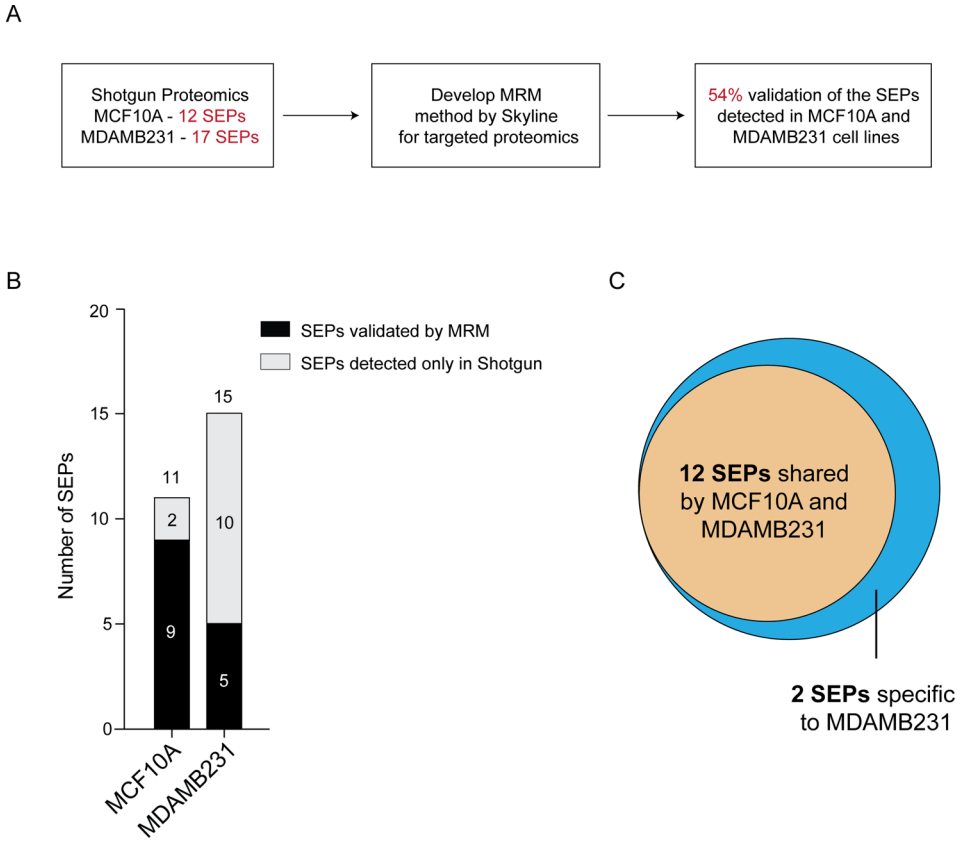


Figure 5. SEP derived from MCF10A and MDAMB231 cell lines. (A) Steps in the discovery and validation of SEPs from these cell lines. (B) Total of nine and five SEPs were validated using MRM in the MCF10A and MDAMB231 cell lines, respectively. (C) These 14 validated SEPs were targeted in MCF10A and MDAMB231, while 12 SEPs found in both cell lines, two SEPs, TASP1-SEP, and CAMD8-SEP, were specific to the MDAMB231 cell line.

We obtained RNA-Seq data for these cell lines, assembled these 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 5A and Supporting Information, Figure 5). Analysis of these SEPs by Skyline followed by a targeted MRM LC−MS/MS experiment validated 14 of these SEPs (out of 29) − 9 in the MCF10A cell line and 5 in the MDAMB231 cell line (Figure 5B).

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 5C). 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.

SEPs Are Present 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 3 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 were annotated or nonannotated. The SEP length for these tdSEPs varied between 15 and 138 amino acids, the percentage of AUG usage was 24%, and most were derived from nonannotated RNAs (80%), which is consistent with data obtained from cell lines (i.e., K562, MCF10A, and MDAMB231) (Figure 6). 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.

■ 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 (Table 1), 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

dx.doi.org/10.1021/pr401280w | J. Proteome Res. 2014, 13, 1757−1765

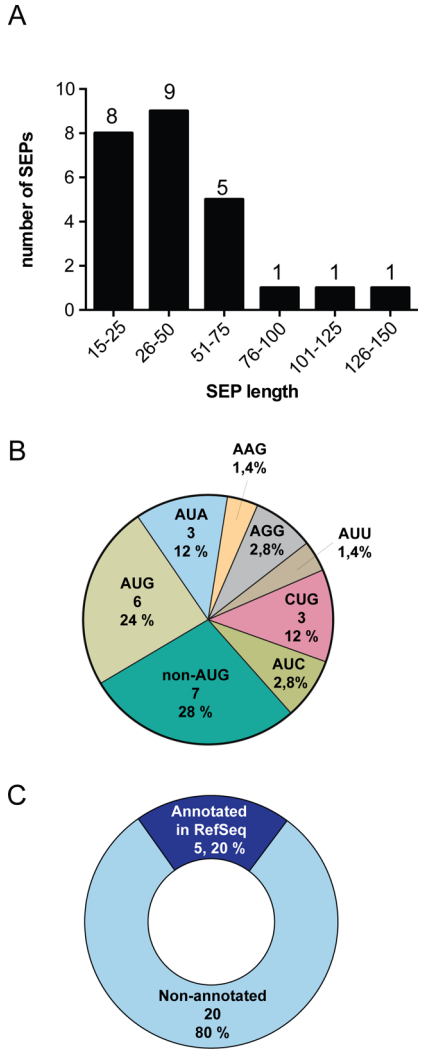


Figure 6. Discovery of 25 tumor derived SEPs (tdSEPs). (A) Length distribution, (B) initiation codon usage, and (C) RNA source of tdSEPs were similar to the distributions seen for SEPs derived from cell lines.

Table 1. Total Number of SEPs Discovered from K562, MCF10A, MDAMB231, and Tumor Samples

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

differentially profiling SEPs in disease models to identify SEPs that might cause disease or can serve as a biomarker.

■ ASSOCIATED CONTENT

Supporting Information

Data for the six additional technical replicates, homology of ASNSD1 across 29 eutherian mammals, overview of 36 targeted MRM LC−MS SEPs, overview of MCF10A and MDAMB231 SEPs, and table of the peptides identified during targeted MRM LC−MS experiment. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: saghatelian@chemistry.harvard.edu.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.A.S. is supported by a National Research Service Award postdoctoral fellowship (1F32GM099408-01). A.S. is supported by National Institute of General Medical Sciences grant R01GM102491. M.K. is supported by NIH U41 HG007234 and NSF CAREER 0644282. We wish to acknowledge the MGH Pathology Research Services assistance in procuring research tissue for this study.

■ ABBREVIATIONS

SEP, sORF-encoded polypeptide; kDa, kilodalton; MWCO, molecular weight cutoff; ERLIC, electrostatic repulsion hydrophilic interaction chromatography; LC−MS/MS, liquid chromatography-tandem mass spectrometry; MS, mass spectrometry; PAGE, polyacrylamide gel electrophoresis; CDS, coding sequence; UTR, untranslated region; MRM, multiple-reaction monitoring

■ REFERENCES

(1) Bertone, P.; Stolc, V.; Royce, T. E.; Rozowsky, J. S.; Urban, A. E.; Zhu, X.; Rinn, J. L.; Tongprasit, W.; Samanta, M.; Weissman, S. Global identification of human transcribed sequences with genome tiling arrays. *Science* **2004**, *306*, 2242.

(2) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57.

(3) Johnson, J. M.; Edwards, S.; Shoemaker, D.; Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **2005**, *21*, 93.

(4) Kapranov, P.; Cheng, J.; Dike, S.; Nix, D. A.; Duttagupta, R.; Willingham, A. T.; Stadler, P. F.; Hertel, J.; Hackermüller, J.; Hofacker, I. L. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **2007**, *316*, 1484.

(5) Nagalakshmi, U.; Wang, Z.; Waern, K.; Shou, C.; Raha, D.; Gerstein, M.; Snyder, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **2008**, *320*, 1344.

(6) Trapnell, C.; Williams, B. A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M. J.; Salzberg, S. L.; Wold, B. J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511.

(7) Khalil, A. M.; Guttman, M.; Huarte, M.; Garber, M.; Raj, A.; Morales, D. R.; Thomas, K.; Presser, A.; Bernstein, B. E.; van Oudenaarden, A. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci.* **2009**, *106*, 11667.

(8) Mitchell Guttman, I. A.; Garber, M.; French, C.; Lin, M. F.; Feldser, D.; Huarte, M.; Zuk, O.; Carey, B. W.; Cassady, J. P.; Cabili, M. N. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **2009**, *458*, 223.

(9) Gish, W.; States, D. J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **1993**, *3*, 266.

(10) Kochetov, A. V. AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. *Bioinformatics* **2005**, *21*, 837.

(11) Ingolia, N. T.; Lareau, L. F.; Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **2011**, *147*, 789.
(12) Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M. N.; Ma, J.; Levin, J. Z.; Karger, A. D.; Budnik, B. A.; Rinn, J. L.; Saghatelian, A. Peptidomic discovery of short open reading frame−encoded peptides in human cells. *Nat. Chem. Biol.* **2013**, *9*, 59−64.
(13) Frith, M. C.; Forrest, A. R.; Nourbakhsh, E.; Pang, K. C.; Kai, C.; Kawai, J.; Carninci, P.; Hayashizaki, Y.; Bailey, T. L.; Grimmond, S. M. The abundance of short proteins in the mammalian proteome. *PLoS Genet.* **2006**, *2*, e52.
(14) Guénet, J. L. The mouse genome. *Genome Res.* **2005**, *15*, 1729.
(15) Ingolia, N. T.; Brar, G. A.; Rouskin, S.; McGeachy, A. M.; Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **2012**, *7*, 1534.
(16) 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.* **2013**, *135*, 16750−16753.
(17) Liu, H.; Sadygov, R. G.; Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **2004**, *76*, 4193.
(18) Oyama, M.; Kozuka-Hata, H.; Suzuki, Y.; Semba, K.; Yamamoto, T.; Sugano, S. Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol. Cell. Proteomics* **2007**, *6*, 1000.
(19) Vanderperre, B.; Lucier, J.-F.; Bissonnette, C.; Motard, J.; Tremblay, G.; Vanderperre, S.; Wisztorski, M.; Salzet, M.; Boisvert, F.-M.; Roucou, X. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS One* **2013**, *8*, e70698.
(20) Hao, P.; Zhang, H.; Sze, S. K. Application of electrostatic repulsion hydrophilic interaction chromatography to the characterization of proteome, glycoproteome, and phosphoproteome using nano LC-MS/MS. *Methods Mol. Biol.* **2011**, *790*, 305.
(21) Hao, P.; Qian, J.; Dutta, B.; Cheow, E. S.; Sim, K. H.; Meng, W.; Adav, S. S.; Alpert, A.; Sze, S. K. Enhanced separation and characterization of deamidated peptides with RP-ERLIC-based multidimensional chromatography coupled with tandem mass spectrometry. *J. Proteome Res.* **2012**, *11*, 1804.
(22) Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal. Chem.* **2000**, *72*, 757.
(23) MacCoss, M. J.; Wu, C. C.; Yates, J. R. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **2002**, *74*, 5593.
(24) Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D. R.; Pimentel, H.; Salzberg, S. L.; Rinn, J. L.; Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562.
(25) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403.
(26) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389.
(27) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26*, 966.
(28) Lin, M. F.; Jungreis, I.; Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **2011**, *27*, i275.
(29) Chittum, H. S.; Lane, W. S.; Carlson, B. A.; Roller, P. P.; Lung, F.-D. T.; Lee, B. J.; Hatfield, D. L. Rabbit β-globin is extended beyond its UGA stop codon by multiple suppressions and translational reading gaps. *Biochemistry* **1998**, *37*, 10866.
(30) Tork, S.; Hatin, I.; Rousset, J. P.; Fabret, C. The major 5′ determinant in stop codon read-through involves two adjacent adenines. *Nucleic Acids Res.* **2004**, *32*, 415.
(31) Salzberg, S. L. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *CABIOS, Comput. Appl. Biosci.* **1997**, *13*, 365.
(32) Gunning, P.; Leavitt, J.; Muscat, G.; Ng, S.-Y.; Kedes, L. A human beta-actin expression vector system directs high-level accumulation of antisense transcripts. *Proc. Natl. Acad. Sci.* **1987**, *84*, 4831.
(33) Katayama, S.; Tomaru, Y.; Kasukawa, T.; Waki, K.; Nakanishi, M.; Nakamura, M.; Nishida, H.; Yap, C.; Suzuki, M.; Kawai, J. Antisense transcription in the mammalian transcriptome. *Science* **2005**, *309*, 1564.
(34) Jessani, N.; Liu, Y.; Humphrey, M.; Cravatt, B. F. Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. *Proc. Natl. Acad. Sci.* **2002**, *99*, 10335.
(35) Albini, A.; Iwamoto, Y.; Kleinman, H.; Martin, G.; Aaronson, S.; Kozlowski, J.; McEwan, R. A rapid in vitro assay for quantitating the invasive potential of tumor cells. *Cancer Res.* **1987**, *47*, 3239.