analytical chemistry

Article

pubs.acs.org/ac

Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides

Jiao Ma,[†,‡] Jolene K. Diedrich,[‡,§] Irwin Jungreis,[∥,⊥] Cynthia Donaldson,[‡] Joan Vaughan,[‡] Manolis Kellis,[∥,⊥] John R. Yates, III,[‡,§] and Alan Saghatelian*[,‡]

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

[‡]Salk Institute for Biological Studies, Clayton Foundation Laboratories for Peptide Biology, 10010 North Torrey Pines Road, La Jolla, California 92037, United States

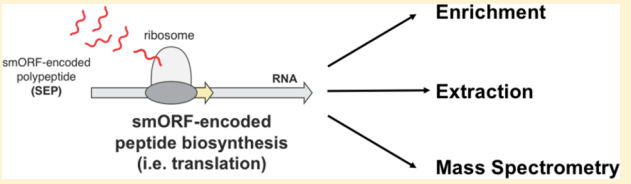
[§]Department of Chemical Physiology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[∥]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: Computational, genomic, and proteomic approaches have been used to discover nonannotated protein-coding small open reading frames (smORFs). Some novel smORFs have crucial biological roles in cells and organisms, which motivates the search for additional smORFs. Proteomic smORF discovery methods are advantageous because they detect smORF-encoded polypeptides (SEPs) to validate smORF translation and SEP stability. Because SEPs are shorter and less abundant than average proteins, SEP detection using proteomics faces unique challenges. Here, we optimize several steps in the SEP discovery workflow to improve SEP isolation and identification. These changes have led to the detection of several new human SEPs (novel human genes), improved confidence in the SEP assignments, and enabled quantification of SEPs under different cellular conditions. These improvements will allow faster detection and characterization of new SEPs and smORFs.



An expression screen for genes that prevent neuronal cell death revealed a novel class of human bioactive peptides.[1] In this screen, a neuronal cell line was engineered to express the Alzheimer's disease protein V642I-APP. Transfection of these engineered cells with a cDNA library identified neuroprotective genes that prevented cell death. One of the protective genes was identified as a 16S rRNA, which was shown to contain a previously unknown 75-bp protein-coding short open reading frame (smORF). smORFs are defined as protein-coding sORF of less than 100 amino acids. The 16S ribosomal smORF produces a 24-amino acid peptide called humanin, which prevents cell death by inhibiting pro-apoptotic BCL-2 proteins.[2,3]

Humanin differs from traditional bioactive peptides, peptide hormones, and neuropeptides, in two ways. First, peptide hormones and neuropeptides are generated from proteolysis of longer proteins called prohormones.[4−8] By contrast, humanin is translated from a smORF as a peptide and does not require further proteolysis for activation. Second, peptide hormones and neuropeptides bind through cell surface receptors, receptor tyrosine kinases (RTKs) and G protein-coupled receptors (GPCRs), while humanin binds an intracellular protein. These differences indicate that humanin is part of a distinct class of bioactive peptides.

Additional work has revealed that genomes harbor many nonannotated smORFs, and some of these smORFs are biologically active.[9−11] In flies, for example, deletion of the tal/pri gene, which encodes several smORFs, results in loss of segmentation of the embryo, and a truncated limb and a missing tarsus in the adult fly.[12,13] Functional smORFs have also been identified in bacteria,[14−16] plants,[17] and other eukaryotes.[17−24]

The biological activity of these novel genes has led to emerging strategies for smORF discovery. smORFs have been discovered by computational,[9,18,19,25] genomic (Ribo-Seq),[18,26,27] and proteomic methods.[28,29] While computational and genomics methods infer protein-coding genes, proteomics provides direct evidence for smORF translation and demonstrates that the resulting smORF-encoded polypeptides (SEPs) are stable enough to be detected. We use a cutoff of 150 amino acids for SEPs because we found a substantial fraction of

Received: January 15, 2016
Accepted: March 16, 2016

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

nonannotated protein-coding ORFs between 100 and 150 amino acids (about 10% of our total).[21]

Proteomic discovery of SEPs and smORFs requires the combination of proteomics and genomics (i.e., RNA-Seq), referred to as proteogenomics.[28,29] Novel SEP discovery begins by enriching the proteome for low molecular weight peptides and small proteins (<30 kDas (kDa)). This fraction is proteolytically digested and analyzed by liquid chromatography-tandem mass spectrometry (LC−MS/MS) proteomics.[21,23,24,28] The resulting LC−MS/MS data set is then interrogated using a protein database from the three-frame translation of the RNA-Seq data[21−23] (Figure 1). Removal of known proteins identifies nonannotated SEPs and smORFs. To identify known (i.e., annotated) SEPs, the human UNIPROT database is used in this workflow instead (Figure 1).

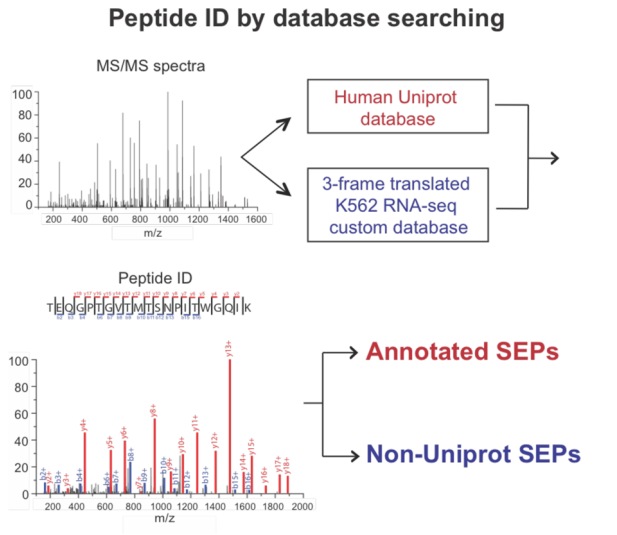


Figure 1. Overview of the SEP discovery workflow. To identify known and novel SEPs MS/MS spectra are searched against the Human UNIPROT database (known SEPs) and a 3-frame translated RNA-Seq custom database (novel SEPs). Peptides that uniquely match to a UNIPROT protein entry that are less than 150 amino acids in length are annotated as known SEPs. Peptides that match to an entry in the RNA-Seq 3-frame translated database that are less than 150 amino acids in length and do not overlap with any UNIPROT proteins are novel SEPs (i.e., nonannotated, non-UNIPROT).

The small size of SEPs compared to proteins make smORF/SEP discovery using proteomics challenging. We typically have to identify a smORF/SEP from a single tryptic peptide because they are shorter than normal proteins. We previously improved proteome fractionation methods to identify more SEPs.[21] Here, we examine the impact of different isolation, enrichment, and mass spectrometry approaches to improve the workflow further. These efforts led to a more confident identification of SEPs and the discovery of 37 nonannotated human SEPs (i.e., 37 novel human genes).

MATERIALS AND METHODS

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% CO_2 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 nonadherent cells (K562) by centrifugation. HEPES-buffered saline (pH 7.5) was used to wash the cells to remove residual media and FBS.

SEP Enrichment Methods. 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 Convertor). For the acid precipitation, the addition of acetic acid (to a final concentration of 0.25% by volume) was followed by centrifugation at 14 000*g* for 20 min at 4 °C. This step precipitates larger proteins to reduce the complexity of the supernatant and enriches lower molecular weight proteins that are then analyzed by LC−MS/MS proteomics for SEPs. For the 30-kDa MWCO, the addition of acetic acid (to a final concentration of 0.25% by volume) was followed by centrifugation at 14 000*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 centrifuged at 25 000*g* for 30 min and supernatants removed and filtered through 5 μM 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 prepared with one column volume methanol and then equilibrated with two-column volumes triethylammonium formate (TEAF) buffer, pH 3.0 before the sample was applied. The cartridges were then washed with two column volumes TEAF and the SEP enriched fraction eluted by the addition of acetonitrile:TEAF pH 3.0 (3:1) and lyophilized using a Savant Speed-Vac concentrator. BCA protein assay (Thermo Scientific) was used to measure protein concentration of each sample after extraction and enrichment.

SEP Extraction Methods. Four different methods were compared for extraction of SEPs from 4×10^7 total cells: (1) 50 mM HCl, 0.1% β-mercaptoethanol (β-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*g* for 30 min, and supernatants filtered through 5 μM 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 prepared with one column volume methanol and equilibrated with two-column volumes triethylammonium formate (TEAF) buffer, pH 3.0 before the sample was applied. The cartridges were then washed with two column volumes TEAF and the SEP enriched fraction eluted by the addition of acetonitrile:TEAF pH 3.0 (3:1) and lyophilized using a Savant Speed-Vac concentrator. BCA protein assay (Thermo Scientific) was used to measure protein concentration of each sample after extraction and enrichment.

Digestion and Sample Preparation for LC−MS/MS. An aliquot of 100 μg of enriched samples was 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 °C in 2 M urea/100 mM TEAB, pH 8.5, with trypsin

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

(Promega). Digestion was stopped with formic acid, 5% final concentration.

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 75 μm packed with PepMap C18 2um particles was used. Electrospray was performed directly from the tip of the analytical column. Buffers 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-h reverse phase gradient. The gradient increases from 5 to 22% B over 280 min, 22−32% B over 60 min, 32−90% B over 10 min, followed by a hold at 90% B for 10 min. The column was re-equilibrated with buffer A before injection.

The Q Exactive was operated in a data-dependent mode. Full MS1 scans were collected with a 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 were collected with default values for AGC target of 1e6 and 5e5 and maximum injection times of 60 and 120 ms for MS and MS/MS scans, respectively. Data were 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 120 and 500 ms. All other parameters remained unchanged.

Orbitrap Fusion Tribrid LC−MS/MS Analysis. C8 SPE enriched samples were analyzed on an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific). The digest was injected directly onto a 50 cm, 75um ID column packed with BEH 1.7um C18 resin (Waters). Samples were separated at a flow rate of 200 nL/min on an nLC 1000 (Thermo Scientific). 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. The column was re-equilibrated with 20 μL of buffer A before the injection of sample. Peptides were eluted directly from the tip of the column and nanosprayed directly into the mass spectrometer by application of 2.5 kV 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 120 K resolution with a mass range of 400 to 1500 m/z and an AGC target of 4e5 and maximum fill time of 50 ms. The cycle time was set to 3 s. Within this 3 s window 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 35 ms or HCD and detection in the Orbitrap with an AGC target of 5e5 and max fill time of 250 ms. 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 s.

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[30] using Integrated Proteomics Pipeline—IP2 (Integrated Proteomics Applications). 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 (RNA-Seq database), and the UNIPROT Human database. The transcriptome data are 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 determine annotated and nonannotated SEPs, data files from technical replicates were combined and searched by ProLuCID. For HCD, data were searched with 50-ppm precursor ion tolerance then filtered to 10-ppm, and 50-ppm fragment ion tolerance with a maximum of two internal missed cleavages using either the custom database or UNIPROT Human database. For CID, data was searched with 500-ppm precursor ion tolerance then filtered to 10-ppm, and 50-ppm fragment ion tolerance with a maximum of two internal missed cleavages using either the custom database or UNIPROT Human database. Identified spectra were filtered and grouped into proteins using DTASelect.[31,32] Proteins and SEPs required, at least, one peptide to be identified with a setting of less than 1% FDR for all searches. Unique peptides identified by searching the UNIPROT database that belonged to smORFs of fewer than a 150 codons were kept and were referred to as "annotated SEPs".

To identify nonannotated 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 a 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. We visually inspect the MS2 spectra for all of the smORF/SEP peptides to validate the assignment. In particular, we required that any critical amino acid residues that uniquely distinguish the peptide was detected in the MS2 data.

The next step is to determine whether the nonannotated peptides are from smORFs or not. The nonannotated peptides are searched against NCBI Human Reference Sequence Database (RefSeq) using tBLASTn, which identifies an RNA that could have produced the SEP. After identifying an RNA and sequence that encodes the peptide, we annotate the downstream in-frame stop codon, and then try to identify the upstream in-frame start codon.

We assign start codons to any in-frame ATG. If there is no in-frame ATG, we look for an in-frame near-cognate codon (i.e., ACG, AAG, CUG, etc.) in a Kozak sequence[33] to assign as the start codon. Lastly, if an in-frame ATG or near-cognate start codon cannot be found, we identify the upstream in-frame stop codon, and if the distance between the upstream and downstream in-frame stop codons is less than 150 codons, we annotated the gene as a smORF. If the peptides did not match to any RNA sequences with the RefSeq RNA database, then it means that they were derived from RNAs that were present in the RNA-Seq data but not in the RefSeq database. For these peptides, we repeat these steps for assigning the smORF using RNAs from the RNA-Seq database.

Arsenite Treatment Experiments. HEK293 cells were grown to ∼70% confluence and then treated with 10 μM sodium arsenite for 24 h. Cellular proteins were extracted using the lysis buffer followed by centrifugation 20 000g for 20 min at

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

4 °C to remove any insoluble particulates. The concentrations were determined using a Bradford assay and 100 μg 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. XICs were extracted with Skyline and peak identity was confirmed by correlating retention time to the identified spectra from the database search results. The AUC (area under the curve) for the peptide ions was used to determine the relative quantity of each peptide between control and arsenite-treated samples. The extraction of the isotopic peaks for each peptide and comparison to the theoretical isotopic distribution at a resolution of 60k validated the selected peptide ion we used for quantitation.

Raising SLC35A4-SEP Antibody. Antisera against SLC35A4 was raised in rabbits against a synthetic peptide fragment encoding Cys[34]SLC35A4(2−34) coupled to maleimide activated keyhole limpet hemocyanin (ThermoFisher, Waltham MA). The peptide, ADDKDSLPKLKDLAFLKNQLESLQRRVEDEVNC, 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 amount 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 3 weeks and were bled 1 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.

Western Blot Analysis. Control and sodium arsenite-treated HEK293 cells were extracted by lysis buffer. Protein concentration was measured using Bradford assay (BioRad). Thirty 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 200 V for 20 min. Proteins were transferred to PVDF membrane and then blocked at room temperature for 1 h using LiCor Blocking Buffer. The membrane was then blotted with primary antibody; rabbit anti β-actin (LiCor) 1:1000 for 1 h 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. Washed membrane three time with TBS-T, then blotted with secondary antibody:goat antirabbit IRDye 800CW (LiCor) at 1:10 000 dilution, rocked 1 h at room temperature. Washed membrane three times with TBS-T then scanned the membrane using LiCor Odyssey CLx at IR700 and IR800. The built-in tool in Odyssey CLx was used to quantify the intensity of the bands of interest.

RESULTS AND DISCUSSION

Enrichment Optimization. Identifying all the SEPs in cells and tissues is required to characterize smORF biology. In a complex mixture such as total cell lysate, detecting small and low abundant proteins is challenging, as detection is naturally biased toward the detection of more abundant proteins.[34] Therefore, SEP detection will likely benefit from an enrichment step, but we have yet to test this assumption. Here, we compare different enrichment methods for their ability to identify the greatest number of known and unknown SEPs from cells.

We began these experiments using K562 cells, which we chose because the first SEPs were discovered using this cell line.[22] The total proteome is prepared by boiling K562 cells to inactivate all proteolytic activity and then lysing the cells by sonication. We used three methods to enrich the <30 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 2A). The results are clear. The 30-kDa MWCO resulted in poor recovery compared to the acid precipitation and SPE.

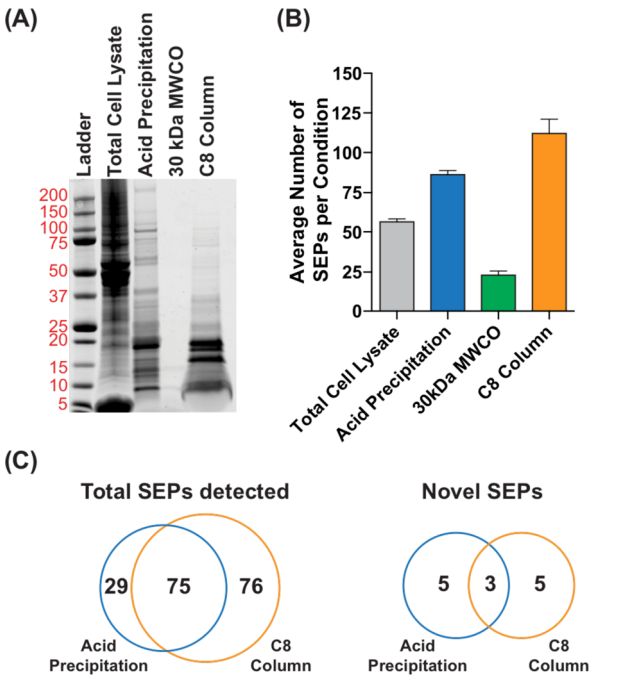


Figure 2. Comparison of different methods for SEP enrichment using K562 cells. (A) Cell lysates were prepared by boiling in water followed by sonication. SEPs were enriched from this lysate by acid precipitation, a 30-kDa MWCO filter, or C8 SPE (i.e., C8 column). The results from these enrichments were analyzed by SDS-PAGE (30 μg total protein per lane, Coomassie stain). (B) Analysis of these samples by proteomics identified the average number of SEPs in each sample. (C) Venn diagrams of the total SEPs (known and novel) and novel SEPs in the acid precipitation and C8 column samples detected by proteomics.

Analysis of total lysate by SDS-PAGE reveals that a majority of the proteome is larger than 30 kDa. Acetic acid precipitation aggregates larger proteins leaving behind lower molecular weight proteins in solution. SDS-PAGE of the solution after acetic acid precipitation led to the majority of the signal coming from proteins less than 30 kDa (Figure 2). Previously, we had relied on MWCO filtration to enrich the lower molecular weight proteome, but this method results in significantly less protein by SDS-PAGE, which hurts our ability to detect SEPs (Figure 2). The solid phase extraction method using selective carbon groups (C8) bonded to silica-based sorbents was

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

originally developed to enrich plasma and tissue extracts for peptide hormones by removing larger molecular weight proteins before measurement by radioimmunoassay.[35,36] Applying this method to enrich the lower molecular weight proteins gave excellent results by SDS-PAGE (Figure 2).

We then determined whether the results we measured by SDS-PAGE correlated with the number of known and unknown SEPs that we could detect using proteomics. Enriched and nonenriched proteome samples were reduced, alkylated, and trypsin digested followed by LC−MS/MS analysis. Samples were analyzed using a 6-h 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 1), the RNA-Seq database and the UNIPROT database.

Analysis of the LC−MS/MS data sets using the human UNIPROT database revealed 70, 96, 35, and 143 known SEPs from the nonenriched, acetic acid precipitated, MWCO and SPE enriched samples, respectively. We analyzed the LC-MS/MS data sets using a custom database made from the three-frame translation of RNA-Seq data from K562 cells, which contains all potential translated proteins in K562 cells. A search of our proteomics data against the RNA-Seq database enabled us to identify several nonannotated SEPs.

From the nonenriched, acetic acid precipitated, MWCO and SPE enriched samples, we identified 4, 8, 1, and 8 non-annotated SEPs, respectively. The average number of SEPs detected, annotated or novel, correlate with the protein recovery we observed by SDS-PAGE (Figure 2B, and Figure S2 of the Supporting Information). The data indicate that the acetic acid precipitation and C8 SPE methods are better than the 30 kDa MWCO filter we have used in the past. Many SEPs were only identified using the acetic acid precipitation or C8 SPE method (Figure 2C). This was consistent in several other cell lines that we tested. (Figure S1). Also, all the methods provide SEP of similar lengths and hydrophobicity (Figure S3 and Figure S4). Therefore, we recommend using both enrichment methods moving forward to maximize the total number of SEPs detected.

Different Methods for SEP Extraction. We compared several distinct methods for isolating SEPs from the lung cancer cell line A549 (i.e., extraction methods) (Figure 3A). We selected another cell line to ensure that our methods translated to the more conventional adherent cells. 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 detected the most SEPs while acid extraction resulted in fewest SEPs detected. Overall, the lysis buffer performs better than water or acid alone, while boiling did not seem to have a strong effect (Figure S5). The number and identity of SEPs detected with or without boiling are similar. 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 3B).

LC−MS/MS Optimization. For SEP discovery, good spectral quality is essential because SEPs are low abundant, with a single peptide detected per SEP in most cases. The confidence of the peptide identification depends on good

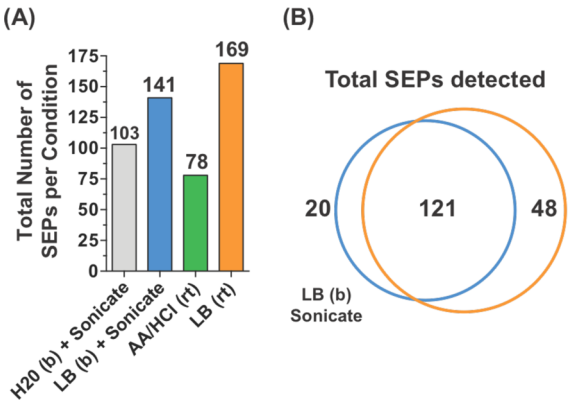


Figure 3. Different extraction methods have a minimal impact on total number of SEPs detected. (A) Total number of SEPs identified from A549 cells using four different SEP extraction methods: boiling (b) in water and sonication; boiling in lysis buffer (LB, 50 mM HCl, 0.1% β-ME, 0.05% Triton X-100) and sonication; acetic acid (AA) and hydrochloric acid (HCl) at room temperature (rt); and lysis buffer at room temperature. (B) Comparison of the extraction methods demonstrated good overlap between the methods with lysis buffer at room temperature capturing the most SEPs.

quality MS/MS spectra—i.e., good sequence coverage and a low background are necessary. Previously, we used an Orbitrap Velos hybrid ion trap mass spectrometer (Thermo Fisher Scientific) with Collision Induced Dissociation (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 MS/MS data, obtained using an Orbitrap, can solve this problem but leads to less sensitivity since more ions are required for detection.

High-energy Collisional Dissociation (HCD) is reported to provide better sequence coverage than CID, provided the HCD energy is adequate for the peptide.[37] Improved sequence coverage can benefit SEP detection by providing more confidence in the SEP peptide detected. We tested whether HCD would improve SEP peptide characterization. For example, MS/MS of a SEP peptide by low-resolution CID and high-resolution HCD on the Fusion Tribrid MS reveals increased sequence coverage using HCD (Figure 4A, B). We found modest improvements in peptide coverage using HCD. For instance, CID identified 11 b-ions and 10 y-ions, while HCD detected 11 b-ions and 12 y-ions. Qualitatively, the HCD spectrum is less noisy, and major peaks in the CID spectra are not assigned (Figure 4A, B). A similar improvement in coverage was observed using HCD with the QE mass spectrometer (Figure S6). These results indicate that HCD provides a slight improvement in sequence coverage of peptides and much lower background, but does not effect the total number of SEPs we detect.

We also optimized the Automatic Gain Control (AGC) and fill time of the Q Exactive to increase coverage in the MS2 spectra. The higher AGC setting and longer max fill times (sensitive) identified 13 b-ions and 17 y-ions, while the default AGC and fill time (standard) settings detected 2 b-ions and 12 y-ions (Figure 4C, D and Figure S7). All data presented herein were collected under the "sensitive" settings to ensure good spectral quality. With the sensitive setting, we observe a marked improvement in the number of detected ions to provide

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

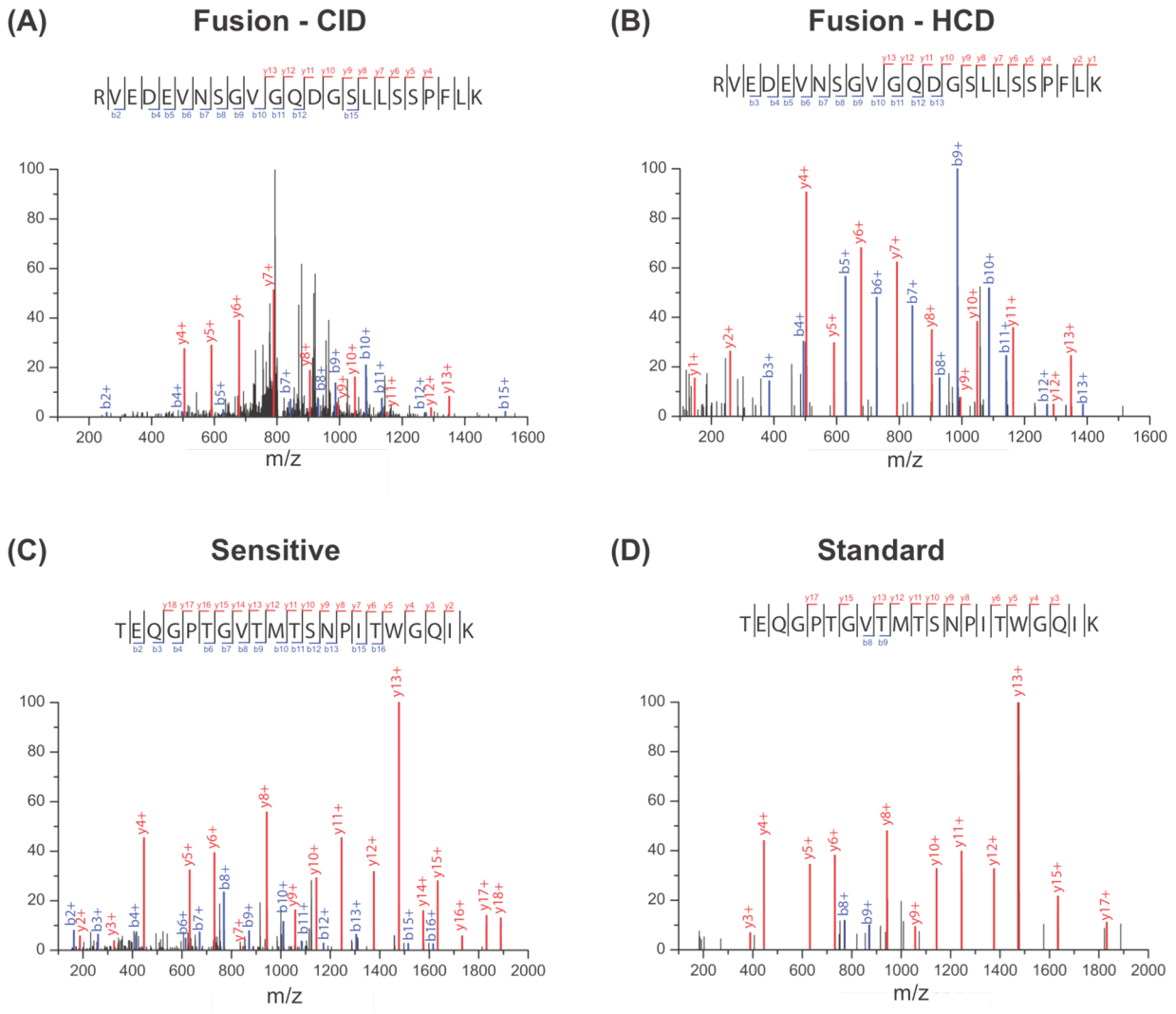


Figure 4. Comparison of MS/MS spectra acquired using different fragmentation methods and automatic gain control. (A) MS/MS spectrum of the same SEP peptide acquired by low resolution CID or (B) high resolution HCD (Fusion Tribrid MS). (C) MS/MS spectrum of the same SEP peptide acquired with sensitive or (D) standard setting (QExactive MS).

significantly better sequence coverage. Therefore, increased fill times and high AGC settings should be used for SEP discovery.

Label-Free SEP Quantitation. We do not obtain many spectral counts for SEPs due to their overall short length, which has prevented us from using spectral counting to quantify SEP levels. Here, we look at using the area under the curve in the MS1 spectra to quantitate SEP levels. We decided to compare SEP levels in control and arsenite-treated HEK293 cells. This system is ideal for these experiments because known increases in heme oxygenase 1 (HO-1) expression can be used as a positive control. Moreover, SCL35A4 mRNA,[38] which includes the SLC35A4 smORF, was reported to be elevated under these conditions, which suggests that arsenite treatment might regulate SEP levels.

Sodium arsenite-treated (10 μM) and untreated HEK293 cells were extracted and analyzed by LC-MS/MS. Heme oxygenase 1 (HO-1) was reported to be up-regulated by arsenite treatment of HEK293 cells in a previous proteomics study,[39] and we validated this change by Western blot showing HO-1 was highly expressed in arsenite-treated samples ($p < 0.01$) (Figure 5A). We looked at HO-1 levels by label-free LC-MS, by quantitating the area under the LC-MS peak for an HO-1 peptide in the MS1 data. We performed label-free quantitative analysis using Skyline software[40,41] that extracts peak area of the detected peptides from MS1 by retention time and accurate mass. Using peak areas allows us to quantitate relative protein or SEP expression level between two conditions. This analysis showed a strong increase in HO-1 peptide levels in the arsenite-treated sample demonstrating that the label-free quantitation is similar to a Western blot (Figure 5A, B).

We measured the levels of three peptides to determine what effect, if any, arsenite has on SEP levels. The peptides included two SEPs, SLC35A4-SEP and SEP257, and cofilin, which was the negative control. As expected, analysis of the area under the curve for a cofilin peptide revealed that cofilin levels were unchanged between the arsenite- and control-treated samples. A similar analysis of SLC35A-SEP and SEP257 demonstrated that these two peptides were unchanged between the control and arsenite-treated samples (Figure 5C, Figure S8). Furthermore, most SEPs have similar ion intensities such that this label-free quantitation method should be general (Figure S9).

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

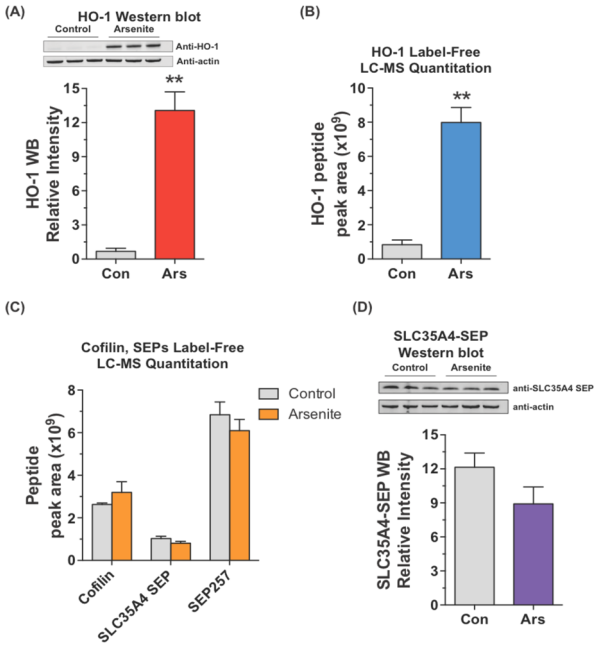


Figure 5. Quantitation of SEPs upon arsenite treatment. (A) HEK293 cells were treated with 10 μM sodium arsenite for 24 h. Western blot analysis revealed increased HO-1 expression upon arsenite treatment (10 μM, 24 h). The intensity of the bands on the Western blot was quantified by LiCor Odyssey CLx and normalized by β-actin. (B) Peak area (MS1) of the HO-1 peptide agrees with Western blot. (C) Peak areas (MS1) of cofilin, SLC35A4-SEP, and SEP257 were unchanged upon arsenite treatment. (D) SLC35A4-SEP levels were also measured by Western blot, which agreed with the proteomics quantitation. The intensity of the bands on the blot was quantified by LiCor Odyssey CLx and normalized by β-actin. (Student's t test, **, $p < 0.01$).

We wanted to confirm that SLC35A4-SEP is unchanged, so we generated an antibody against SLC35A4-SEP, which we used for Western blot analysis. We tested this antibody by overexpressing SLC35A4 and demonstrated that it efficiently detects SLC35A4-SEP (Figure S10). Using this antibody against control and arsenite-treated samples shows that SLC35A4-SEP levels are unchanged (Figure 5D), supporting our quantitative label-free mass spectrometry results. The label-free quantitative method, which measures SEP levels between two different conditions, will be of tremendous use in distinguishing SEPs that are changing under different biological conditions, even though we did not find any changes in this example.

Analysis of Novel Human SEPs. In this study, we detected 37 novel human SEPs (Table S1), which come from smORFs that are not annotated in the RefSeq database. Each of these smORFs represents a novel human gene. The new SEPs are translated from smORFs in the 5′UTR (5 SEPs), 3′UTR (2 SEPs), noncoding RNAs (6 SEPs), and 24 RNAs that were not in the RefSeq database but are present in our RNA-Seq data. Most of the new smORFs (21 in total (55%)) have an AUG start codon, while the remaining 16 SEPs (45%) do not. This observation is in agreement with previous studies,[21,23,24,27] indicating that a significant portion of SEPs can be translated from noncanonical AUG start codon.

A few of the SEPs are unknown isoforms of known proteins. For instance, one of the SEP peptides we detected, GYFDSGDYNMAK, is derived from a 119 amino acid SEP from a nonannotated smORF with an ATG start. When we align this SEP to nonredundant human proteins using pBLAST, it has >85% sequence homology to several α-endosulfine protein isoforms (Figure 6A and Figure S11). Thus, we conclude that

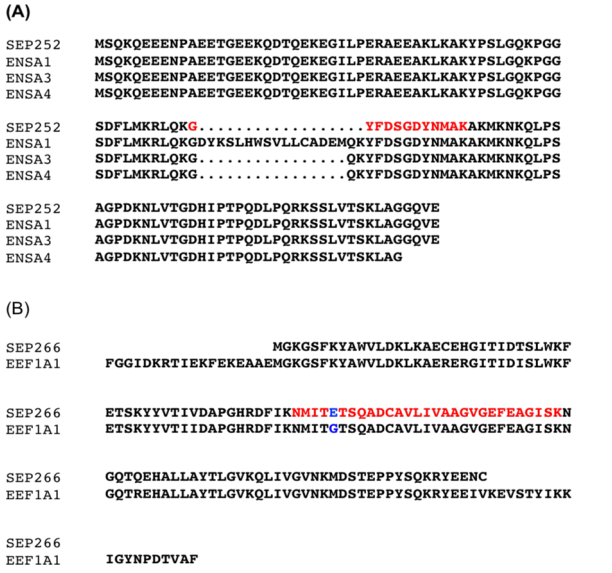


Figure 6. Some novel SEPs are new isoforms of known proteins or fragments of longer proteins. (A) SEP252 is a new isoform of the protein α-endosulfine (ENSA) and this connection was discovered because the SEP peptide (red) is homologous to ENSA but different enough to realize that this peptide is from a nonannotated smORF. Alignment of the entire smORF demonstrates high sequence homology (>80%) to various ENSA isoforms indicating that this SEP is a member of the ENSA family of proteins. (B) SEP266 was identified through a peptide that is homologous (red) to another peptide from Elongation factor 1-α 1 (EEF1A1) but differs by one amino acid (blue) indicating that it belongs to a nonannotated smORF. Alignment of the entire smORF shows high sequence homology (>80%) to part of EEF1A1.

SEP252 is a novel α-endosulfine protein isoform, and we demonstrate how SEP discovery can help find additional, nonannotated, isoforms of known small proteins.

Another group of newly discovered SEPs has sequence homology to known proteins but the SEP and the known protein are different lengths, a part of much longer proteins (Figure 6B and Figure S11). For example, one of the SEPs peptides, NMITETSQADCAVLIVAAGVGEFEAGISK, belongs to a 123 amino acid long SEP with an ATG start. pBLAST of this sequence demonstrated strong sequence homology of this SEP to a 462 amino acid long eukaryotic translation elongation factor 1 from residues 49 to 169. Truncated variants of EEF1A1 have previously been shown to promote[42] or suppress[43] cancer cell growth suggesting that this SEP266 might be an interesting candidate for downstream cell biological studies. The discovery of truncated forms of known proteins, such as EEF1A1, might provide new insight into the biological regulation of these proteins.

CONCLUSIONS

By testing and optimizing several different parameters in the SEP workflow, we have improved the number of SEPs detected, and enhanced the confidence in those assignments. The identification of smORFs and SEPs becomes increasingly important as new biological functions are emerging. For example, new mammalian SEPs that regulate muscle endur-

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

ance[44] and metabolism[10] have recently been discovered. As a potential pool of molecules with roles in fundamental biology, the discovery of smORFs and SEPs is of paramount importance. Here, we highlight the power of proteomics in contributing to this field by defining a new workflow that improves on the enrichment, mass spectrometry, and quantitation of human SEPs.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b00191.

(Figure S1) Acid precipitation and C8 solid phase extraction; (Figure S2) total number of SEPs; (Figure S3) SEP length distribution; (Figure S4) hydropathy score; (Figure S5) pairwise comparison of the extraction methods; (Figure S6) HCD; (Figure S7) MS/MS spectra of three SEP peptides; (Figure S8) peak areas of the detected peptides; (Figure S9) label-free quantitative analysis using Skyline software; (Figure S9) label-free quantitative analysis using Skyline software; (Figure S10) Western blot; (Figure S11) RNA-Seq transcript; (Table S1) full list of 37 non-UNIPROT SEPs (PDF)
(XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: asaghatelian@salk.edu (A.S.).

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

This study was supported by the NIH (R01 GM102491, A.S), the NCI Cancer Center Support Grant P30 (CA014195 MASS core, A.S.), The Leona M. and Harry B. Helmsley Charitable Trust grant (#2012-PG-MED002, A.S.), and Dr. Frederick Paulsen Chair/Ferring Pharmaceuticals (A.S.), and NIH (P41 GM103533 and R01 MH067880, J.K.D., J.R.Y.), and NIH (R01 HG004037, I.J., M.K.) and GENCODE Welcome Trust grant (U41 HG007234, I.J., M.K.).

■ ABBREVIATIONS

SEP	smORF-Encoded Polypeptide
kDa	kilodalton
MWCO	molecular weight cutoff
LC−MS/MS	liquid chromatography−tandem mass spectrometry
MS	mass spectrometry
SPE	solid phase extraction
PAGE	polyacrylamide gel electrophoresis
CDS	coding sequence
UTR	untranslated region
CID	collision induced dissociation
HCD	high-energy collisional dissociation
AGC	automatic gain control
HO-1	heme oxygenase 1

■ REFERENCES

(1) Hashimoto, Y.; Niikura, T.; Tajima, H.; Yasukawa, T.; Sudo, H.; Ito, Y.; Kita, Y.; Kawasumi, M.; Kouyama, K.; Doyu, M.; Sobue, G.; Koide, T.; Tsuji, S.; Lang, J.; Kurokawa, K.; Nishimoto, I. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 6336−6341.
(2) Guo, B.; Zhai, D.; Cabezas, E.; Welsh, K.; Nouraini, S.; Satterthwait, A. C.; Reed, J. C. *Nature* **2003**, *423*, 456−461.
(3) Zhai, D.; Luciano, F.; Zhu, X.; Guo, B.; Satterthwait, A. C.; Reed, J. C. *J. Biol. Chem.* **2005**, *280*, 15815−15824.
(4) Bliss, M. *The Discovery of Insulin*; University of Chicago Press: Chicago, 2013.
(5) Bliss, M.; Purkis, R. *The Discovery of Insulin*; University of Chicago Press: Chicago, 1982.
(6) De Lecea, L.; Kilduff, T.; Peyron, C.; Gao, X.-B.; Foye, P.; Danielson, P.; Fukuhara, C.; Battenberg, E.; Gautvik, V.; Bartlett, F. n. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 322−327.
(7) Sakurai, T.; Amemiya, A.; Ishii, M.; Matsuzaki, I.; Chemelli, R. M.; Tanaka, H.; Williams, S. C.; Richardson, J. A.; Kozlowski, G. P.; Wilson, S. *Cell* **1998**, *92*, 573−585.
(8) Vale, W.; Spiess, J.; Rivier, C.; Rivier, J. *Science* **1981**, *213*, 1394−1397.
(9) Ladoukakis, E.; Pereira, V.; Magny, E. G.; Eyre-Walker, A.; Couso, J. P. *Genome Biol.* **2011**, *12*, R118.
(10) Lee, C.; Zeng, J.; Drew, B. G.; Sallam, T.; Martin-Montalvo, A.; Wan, J.; Kim, S. J.; Mehta, H.; Hevener, A. L.; de Cabo, R.; Cohen, P. *Cell Metab.* **2015**, *21*, 443−454.
(11) Slavoff, S. A.; Heo, J.; Budnik, B. A.; Hanakahi, L. A.; Saghatelian, A. *J. Biol. Chem.* **2014**, *289*, 10950−10957.
(12) Galindo, M. I.; Pueyo, J. I.; Fouix, S.; Bishop, S. A.; Couso, J. P. *PLoS Biol.* **2007**, *5*, e106.
(13) Kondo, T.; Hashimoto, Y.; Kato, K.; Inagaki, S.; Hayashi, S.; Kageyama, Y. *Nat. Cell Biol.* **2007**, *9*, 660−665.
(14) Hemm, M. R.; Paul, B. J.; Miranda-Ríos, J.; Zhang, A.; Soltanzad, N.; Storz, G. *J. Bacteriol.* **2010**, *192*, 46−58.
(15) Hemm, M. R.; Paul, B. J.; Schneider, T. D.; Storz, G.; Rudd, K. E. *Mol. Microbiol.* **2008**, *70*, 1487−1501.
(16) Wadler, C. S.; Vanderpool, C. K. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 20454−20459.
(17) Hanada, K.; Higuchi-Takeuchi, M.; Okamoto, M.; Yoshizumi, T.; Shimizu, M.; Nakaminami, K.; Nishi, R.; Ohashi, C.; Iida, K.; Tanaka, M. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 2395−2400.
(18) Aspden, J. L.; Eyre-Walker, Y. C.; Phillips, R. J.; Amin, U.; Mumtaz, M. A. S.; Brocard, M.; Couso, J.-P. *eLife* **2014**, *3*, e03528.
(19) 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. *PLoS Genet.* **2006**, *2*, e52.
(20) Kastenmayer, J. P.; Ni, L.; Chu, A.; Kitchen, L. E.; Au, W. C.; Yang, H.; Carter, C. D.; Wheeler, D.; Davis, R. W.; Boeke, J. D.; Snyder, M. A.; Basrai, M. A. *Genome Res.* **2006**, *16*, 365−373.
(21) Ma, J.; Ward, C. C.; Jungreis, I.; Slavoff, S. A.; Schwaid, A. G.; Neveu, J.; Budnik, B. A.; Kellis, M.; Saghatelian, A. *J. Proteome. Res.* **2014**, *13*, 1757−1765.
(22) Oyama, M.; Kozuka-Hata, H.; Suzuki, Y.; Semba, K.; Yamamoto, T.; Sugano, S. *Mol. Cell. Proteomics* **2007**, *6*, 1000−1006.
(23) 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. *Nat. Chem. Biol.* **2012**, *9*, 59−64.
(24) Vanderperre, B.; Lucier, J. F.; Bissonnette, C.; Motard, J.; Tremblay, G.; Vanderperre, S.; Wisztorski, M.; Salzet, M.; Boisvert, F. M.; Roucou, X. *PLoS One* **2013**, *8*, e70698.
(25) Vanderperre, B.; Lucier, J. F.; Roucou, X. *Database* **2012**, *2012*, bas025.
(26) Bazzini, A. A.; Johnstone, T. G.; Christiano, R.; Mackowiak, S. D.; Obermayer, B.; Fleming, E. S.; Vejnar, C. E.; Lee, M. T.; Rajewsky, N.; Walther, T. C.; Giraldez, A. J. *EMBO J.* **2014**, *33*, 981−993.
(27) Ingolia, N. T.; Lareau, L. F.; Weissman, J. S. *Cell* **2011**, *147*, 789−802.

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX

(28) Branca, R. M.; Orre, L. M.; Johansson, H. J.; Granholm, V.; Huss, M.; Pérez-Bercoff, Å.; Forshed, J.; Käll, L.; Lehtiö, J. *Nat. Methods* **2013**, *11*, 59−62.
(29) Castellana, N.; Bafna, V. *J. Proteomics* **2010**, *73*, 2124−2135.
(30) Xu, T.; Park, S. K.; Venable, J. D.; Wohlschlegel, J. A.; Diedrich, J. K.; Cociorva, D.; Lu, B.; Liao, L.; Hewel, J.; Han, X.; Wong, C. C.; Fonslow, B.; Delahunty, C.; Gao, Y.; Shah, H.; Yates, J. R., 3rd *J. Proteomics* **2015**, *129*, 16−24.
(31) Cociorva, D.; Tabb, D. L.; Yates, J. R. *Curr. Protoc. Bioinformatics***2007**, *16*.
(32) Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd *J. Proteome Res.* **2002**, *1*, 21−26.
(33) Kozak, M. *Cell* **1986**, *44*, 283−292.
(34) Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd *Anal. Chem.* **2004**, *76*, 4193−4201.
(35) Vale, W.; Vaughan, J.; Jolley, D.; Yamamoto, G.; Bruhn, T.; Seifert, H.; Perrin, M.; Thorner, M.; Rivier, J. *Methods Enzymol.* **1986**, *124*, 389−401.
(36) Vale, W.; Vaughan, J.; Yamamoto, G.; Bruhn, T.; Douglas, C.; Dalton, D.; Rivier, C.; Rivier, J. *Methods Enzymol.* **1983**, *103*, 565−577.
(37) Diedrich, J. K.; Pinto, A. F.; Yates, J. R., 3rd *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1690−1699.
(38) Andreev, D. E.; O'Connor, P. B.; Fahey, C.; Kenny, E. M.; Terenin, I. M.; Dmitriev, S. E.; Cormican, P.; Morris, D. W.; Shatsky, I. N.; Baranov, P. V. *eLife* **2015**, *4*, e03971.
(39) Lau, A. T.; He, Q. Y.; Chiu, J. F. *Biochem. J.* **2004**, *382*, 641−650.
(40) 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. *Bioinformatics* **2010**, *26*, 966−968.
(41) Schilling, B.; Rardin, M. J.; MacLean, B. X.; Zawadzka, A. M.; Frewen, B. E.; Cusack, M. P.; Sorensen, D. J.; Bereman, M. S.; Jing, E.; Wu, C. C.; Verdin, E.; Kahn, C. R.; Maccoss, M. J.; Gibson, B. W. *Mol. Cell. Proteomics* **2012**, *11*, 202−214.
(42) Dahl, L. D.; Corydon, T. J.; Ränkel, L.; Nielsen, K. M.; Füchtbauer, E.-M.; Knudsen, C. R. *Cancer Cell Int.* **2014**, *14*, 17.
(43) Rho, S. B.; Park, Y. G.; Park, K.; Lee, S.-H.; Lee, J.-H. *FEBS Lett.* **2006**, *580*, 4073−4080.
(44) Anderson, D. M.; Anderson, K. M.; Chang, C. L.; Makarewich, C. A.; Nelson, B. R.; McAnally, J. R.; Kasaragod, P.; Shelton, J. M.; Liou, J.; Bassel-Duby, R.; Olson, E. N. *Cell* **2015**, *160*, 595−606.

DOI: 10.1021/acs.analchem.6b00191
Anal. Chem. XXXX, XXX, XXX−XXX