An Optimized Proteomics Approach Reveals Novel Alternative Proteins in Mouse Liver Development

Authors

Ying Yang, Hongwei Wang, Yuanliang Zhang, Lei Chen, Gennong Chen, Zhaoshi Bao, Yang Yang, Zhi Xie, and Qian Zhao

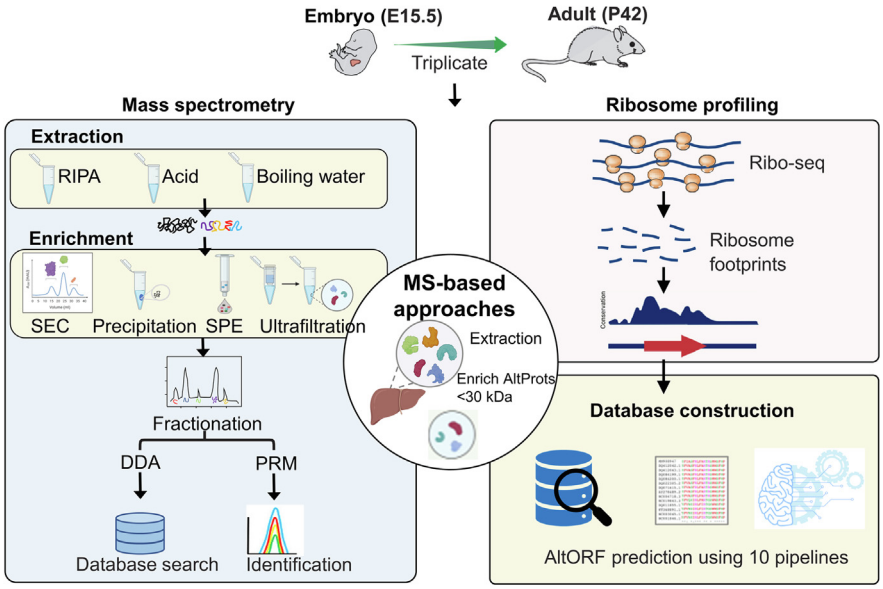
Correspondence

q.zhao@polyu.edu.hk

In Brief

This study proposed a novel and efficient method for the improved discovery and identification of AltProts, integrating RIPA extraction, size-exclusion chromatography (SEC) enrichment, electrostatic repulsion–hydrophilic interaction chromatography (ERLIC) fractionation, MS analysis, and a Ribo-seq–based AltProt database. Importantly, SEC is attractive for simultaneous enrichment and fractionation from complex proteomes. With this strategy, we discovered eighty-nine novel AltProts in embryonic and adult mouse livers, which could play important roles in embryonic development.

Graphical Abstract



Highlights

- Establishment of an efficient method to improve AltProt identification.
- Method enabled simultaneous enrichment and fractionation from complex proteome.
- Eighty-nine novel AltProts were identified to reveal AltORF translation in liver development.
- Establishment of a combined approach of Ribo-seq prediction and targeted MS detection.
- Differential AltProts analysis reveals involvement in development-related biological pathways.

2023, Mol Cell Proteomics *22(1)*, 100480
© 2022 THE AUTHORS. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
https://doi.org/10.1016/j.mcpro.2022.100480

An Optimized Proteomics Approach Reveals Novel Alternative Proteins in Mouse Liver Development

Ying Yang[1,‡], Hongwei Wang[2,‡], Yuanliang Zhang[1], Lei Chen[1], Gennong Chen[2], Zhaoshi Bao[3], Yang Yang[1], Zhi Xie[2], and Qian Zhao[1,*]

Alternative ORFs (AltORFs) are unannotated sequences in genome that encode novel peptides or proteins named alternative proteins (AltProts). Although ribosome profiling and bioinformatics predict a large number of AltProts, mass spectrometry as the only direct way of identification is hampered by the short lengths and relative low abundance of AltProts. There is an urgent need for improvement of mass spectrometry methodologies for AltProt identification. Here, we report an approach based on size-exclusion chromatography for simultaneous enrichment and fractionation of AltProts from complex proteome. This method greatly simplifies the variance of AltProts discovery by enriching small proteins smaller than 40 kDa. In a systematic comparison between 10 methods, the approach we reported enabled the discovery of more AltProts with overall higher intensities, with less cost of time and effort compared to other workflows. We applied this approach to identify 89 novel AltProts from mouse liver, 39 of which were differentially expressed between embryonic and adult mice. During embryonic development, the upregulated AltProts were mainly involved in biological pathways on RNA splicing and processing, whereas the AltProts involved in metabolisms were more active in adult livers. Our study not only provides an effective approach for identifying AltProts but also novel AltProts that are potentially important in developmental biology.

Alternative ORFs (AltORFs) are unannotated coding sequences that are different from any known protein-coding gene documented in database or reference annotation projects (1), also known as nonclassical ORFs, novel ORFs. AltORFs contain any unannotated coding sequence of any reading frame of mRNA or alleged ncRNA (1, 2). The translation products of AltORFs are termed alternative proteins (AltProts), which have no similarity to canonical reference proteins (RefProts) of the same gene. Unlike short proteins/microproteins that are encoded by small ORFs (sORFs) with restrictions of less than 100 amino acids, AltProts do not have an upper limit on length (3). Therefore, AltProts include proteins of less than and greater than 100 amino acids. Recently, AltProts have turned out to play essential roles in a variety of physiological processes or diseases (4–7), such as metabolism (8), cancer immunology (9, 10), transcriptional (11), translational regulation (12), ion signaling (13), and development (12, 14).

However, the discovery of functional AltProts was mostly serendipitous, to date, we still lack a systematic approach to directly identify AltProts from biological specimens in large scale (15). The Ribosome profiling (Ribo-seq) technique sequences ribosome-protected RNA fragments and thus enables the prediction of thousands of AltORFs with bioinformatics pipelines (16, 17). Currently, mass spectrometry (MS) is considered as the only method that allows direct identification of AltProts (18, 19). However, only tens to hundreds of AltProts per sample can be identified by MS (2, 20–22). The big difference in the identification number between the two methods calls for urgent improvement on the MS-based methodologies to detect AltProts. The discovery of AltProts by MS is challenging partly due to their short length and interference from large canonical proteins (19, 23). Another major obstacle is the lack of well-established AltProt databases (9). Using public databases that combine all translational products from various samples, the efficiency of AltProt discovery is far inferior than that of RefProts. Considering the high temporal/spatial specificity of AltProts translation, it is important to use customized database from the same specific samples for mining novel AltProts. Although

From the [1]State Key Laboratory of Chemical Biology and Drug Discovery, Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, SAR, China; [2]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China; [3]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical School, Beijing, China
‡These authors contributed equally to this work.
*For correspondence: Qian Zhao, q.zhao@polyu.edu.hk.

© 2022 THE AUTHORS. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

https://doi.org/10.1016/j.mcpro.2022.100480

several prior works have improved the AltProt sample preparation procedures or database construction, there is still a vast room for improvement (2, 9, 22, 24).

Protein translation plays a crucial role in embryonic development and thus is under precise regulations (12, 25). While a large number of canonical proteins and their mechanisms in developmental biology have been thoroughly investigated, only a few AltProts have been studied (12, 14). Considering AltProts could also play pivotal roles in development, either independently or through the regulation of canonical proteins, the large scale and accurate identification of AltProts is crucial for our understanding of the mechanisms in embryonic development.

We herein report an optimized approach integrating MS and Ribo-seq techniques to identify AltProts with improved depth and efficiency. With the optimized approach, we were able to discover and quantify stage-dependent AltProts from embryonic and adult livers that were enriched in specific biological pathways. Our study not only provided an approach but also novel AltProts as new players in liver development.

EXPERIMENTAL PROCEDURES

Chemicals and Reagents

Acetonitrile, methanol, formic acid (FA), trichloroacetic acid, water (HPLC grade), 16% Tricine gel, and Tricine SDS running buffer were from Thermo Fisher Scientific. Acetic acid (AA), ethanol, and chloroform were from DUKSAN. Lysyl endopeptidase (Lys-C, mass spectrometry grade) and trypsin (sequencing grade) were purchased from Promega. Ammonium formate, ammonium bicarbonate, DL-DTT, and iodoacetamide were from Sigma-Aldrich, and all other reagents were from Sigma-Aldrich.

Animals and Tissue Collection

To compare AltProt enrichment and fractionation methods from liver total lysates, C57BL/6 mice weighing between 18 and 22 g were purchased from Centralized Animal Facilities, The Hong Kong Polytechnic University, Hong Kong. Adult mice were anaesthetized and then perfused with isotonic saline containing protease inhibitors (0.120 mM EDTA, 0.2 mM PMSF, and Roche Complete Protease Inhibitor tablets, pH 7.4) before decapitation. Livers were quickly dissected and immediately snap-frozen in liquid nitrogen. All animal experiments were approved by the Hong Kong Polytechnic University Animal Subjects Ethics Subcommittee (Approval No: 20-21/275-ABCT-R-STUDENT) and were performed in accordance with the Institutional Guidelines and Animal Ordinance of the Department of Health.

For discovery of AltProts in embryonic liver development, livers were harvested separately from embryonic (E15.5) and adult (P42) C57BL/6 mice and immediately snap-frozen in liquid nitrogen. Mice were purchased from the Guangdong Medical Experimental Animal Center (Guangdong, China; License No: SCXK (YUE) 2018 0002). All experimental procedures were approved by the Animal Ethics Committee of the Zhongshan Ophthalmic Center, Sun Yat-sen University (Guangzhou, China; License No: SYXK (YUE) 2018 0189) and in accordance with the institutional animal welfare guidelines and Animal Protection Law of China.

Protein Extraction and AltProt Enrichment

Mouse liver tissues were obtained from The Hong Kong Polytechnic University. Three different AltProt extraction methods were compared: (1) RIPA lysis buffer (50 mM Tris-HCl, 150 mM sodium chloride, 2 mM EDTA, 1% NP40, 1% sodium deoxycholate), (2) acid lysis buffer (50 mM hydrochloric acid (HCl), 0.1% β-mercaptoethanol; 0.05% Triton X-100) (22), and (3) boiling water (22). Then, the extracts were centrifuged at 16,000 g for 20 min at 4 °C to remove residual debris.

We tested 10 enrichment methods in triplicates from four categories, (1) precipitation, (2) size selection, (3) solid phase extraction (SPE) enrichment method, (4) hexagonal mesoporous silica materials, using equal amounts of lysates. For the precipitation, these methods used were based upon previously described protocols. AA 0.25% or AA 25 % precipitation: AA (0.25%, v/v) (22) or (25%, v/v) (2) was added to the supernatant followed by centrifugation at 16, 000g for 20 min at 4 °C. For the trichloroacetic acid (TCA) precipitation, 20% TCA was added to the samples as 1:1 (v/v), followed by the addition of chloroform ($CHCl_3$) 1:1 (v/v). Samples were centrifuged at 1500 g for 10 min at 4 °C and transfer the supernatant to a new tube. The lower samples were then washed with 100 μl of Milli-Q water and 100 μl of methanol, followed by vortex and centrifugation at 1500 g at 4 °C for 10 min. Subsequently, both supernatants were combined (26). For acetonitrile (ACN) precipitation, 3.2 volumes of ACN plus 0.1% TFA was added to the sample (27). For methyl tert-butyl ether (MTBE)-based sequential precipitation, single-phase buffer MTBE/methanol/water (5:3:1, v/v) and two-phase buffer MTBE/methanol/water (5:1:1, v/v) were applied for sequential precipitation and delipidation as described previously (28). For the method of size selection category, the first one is the 30-kDa-molecular weight cut-off ultrafiltration (30-kDa-MWCO), the lysate was loaded into a 30-kDa-MWCO (Millipore), and the flow through was collected (22). Another method is size-exclusion chromatography (SEC) enrichment, to isolate proteins <30 kDa from larger proteins in liver lysates, a GE AKTA Explorer FPLC System (GE Healthcare) was combined with a Sephadex 75 Increase 5/150 Gl column (GE Healthcare) for enrichment and fractionation of small proteins. Low molecular weight standards (GE Healthcare) were used for mass calibration. Each SEC separation run was performed at a flow rate of 0.2 ml/min at a wavelength of 254 nm for 15 min. Only fractions between 8 min and 15 min of retention time, which corresponded to proteins of molecular weight <30 kDa and had a total volume of 1.6 ml, were collected into a low protein binding tube (Eppendorf); for SEC enrichment purpose, these fractions were combined into one tube and lyophilized before use. For the method of SPE category, the liver lysates were enriched using C8 SPE cartridges (Agilent Technologies) or hydrophilic-lipophilic-balanced SPE (HLB SPE, Waters) cartridges. The first method is C8 SPE-based enrichment (22), cartridges were activated with one column volume (CV) of methanol and then equilibrated with two CVs of triethylammonium formate (TEAF) buffer (pH 3.0) before the lysate was applied. The cartridges were then washed with two CVs of TEAF buffer (pH 3.0) and the enriched proteins were eluted with ACN:TEAF buffer (3:1, pH 3.0). The other method is HLB SPE-based enrichment, cartridges were activated with methanol and then equilibrated with water before the lysate was applied. The cartridges were then washed with water and eluted with 60% ACN. Lastly, hexagonal mesoporous silica materials MCM-41 were mixed with lysates and small proteins were extracted as described by Du *et al* (29). Detailed protocol on enrichment method is available in the supplemental materials.

Protein Sample Cleanup with SP3 Method

For each 20 μg of sample, Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophobic and Sera-Mag SpeedBeads Carboxyl

Magnetic Beads, hydrophilic (GE Healthcare) were gently combined in a ratio of 1:1 (v/v) and used as described by Hughes *et al* (30). Samples were reduced and alkylated using DTT and iodoacetamide, respectively. Next, the bead slurries were transferred to the samples. Then, absolute ethanol was added to a final concentration of 50% (v/v) to induce protein binding. Beads were resuspended in 50 mM ammonium bicarbonate supplemented with Lys-C enzymes at an enzyme to protein ratio of 1:100 (w/w). After 4 h incubation, trypsin was added at an enzyme to protein ratio of 1:20 (w/w), as 1:25 was recommended by Hughes *et al*. for complete digestion, and the sample was incubated at 37 °C for 12 h. Peptide concentration was determined using the Pierce Quantitative Fluorometric Peptide assay (Thermo Fisher Scientific). From each sample, peptides were labeled with TMT-6plex (includes the following channels: 126, 127N, 127C, 128N, 128C, 129N, Thermo Fisher Scientific) according to the manufacturer's instructions.

SDS-PAGE Gel Analysis of Enriched AltProts Samples

After enrichment, protein content was quantified by the Bradford assay and the same amount (25 μg) of protein was loaded on each lane of the gel. Samples were analyzed using 16% tricine-SDS-PAGE and separated at a constant 60 V until they completely entered the separating gel from the stacking gel. Then, a constant 110 V was maintained until the tracking dye reached the bottom of the gel. Finally, the gel was stained with Coomassie brilliant blue R-250 (Bio-Rad).

Comparison of Fractionation Methods after SEC Enrichment

SEC Enrichment into Four Fractions—Mouse liver samples were loaded on the SEC column and then final four fractions of the low molecular weight range were collected and finally they were injected separately into MS for detection.

High-pH Reversed-Phase Fractionation—After SEC enrichment, the obtained proteins were digested and then peptides were fractionated using a Waters Acquity UPLC Peptide BEH C18 column (2.1 × 100 mm, 1.7 μM, Waters) on an Agilent 1290 Infinity LC system (Agilent Technologies) operating at 50 μl/min. Buffer A consisted of 10 mM ammonium formate and buffer B consisted of 10 mM ammonium formate and 90% ACN, both buffers were adjusted to pH 9 with ammonium hydroxide as described previously (31). Fractions were collected every 1 min from 6 min to 100 min retention time (96 fractions, finally concatenated into eight fractions). Peptides were separated by a linear gradient as follows: 0 to 10 min, 1% B; 10 to 38 min, 1 to 8% B; 38 to 75 min, 8 to 62% B; 75 to 85 min, 62 to 95% B; 85 to 100 min, 95% B. The final eight fractions were concentrated and analyzed by LC-MS/MS.

ERLIC Fractionation—After SEC enrichment, the obtained proteins were digested and then peptides were fractionated using an Agilent 1290 Infinity LC system equipped with a PolyWAX ERLIC column (200 × 2.1 mm, 5 μM, 300 Å, PolyLC) as described previously (32). Buffer A consisted of 90% acetonitrile and 0.1% AA and buffer B consisted of 30% acetonitrile, 0.1% FA. From 6 min to 100 min retention time, fractions were collected every 1 min (96 fractions, finally concatenated into eight fractions). Peptides were separated by a stepwise gradient as follows: 0 to 10 min, 0% B; 10 to 22 min, 0 to 8% B; 22 to 38 min, 8 to 45% B; 38 to 50 min, 45 to 80% B; 50 to 68 min, 80 to 98% B; 68 to 100 min, 98% B. The final eight fractions were concentrated and analyzed by LC-MS/MS.

LC-MS/MS Analysis

For data-dependent acquisition, all mass spectrometry data were collected on an Orbitrap Exploris 480 mass spectrometry equipped with the FAIMS interface and coupled with an Ultimate 3000 RSLC nano system (Thermo Fisher Scientific). The digested samples were redissolved in 0.1% FA and separated on a self-packed capillary column packed with Reprosil-Pur C18 1.9 μM particles (Dr Maisch GmbH). Mobile phase A (0.1% FA) and mobile phase B (80% ACN and 0.1% FA) were used to separate peptides with the following gradients: 2 min, 8 - 10% B; 2 to 120 min, 10 - 35% B, 120 to 140 min, 35 to 90% B; 140 to 150 min, 90%B in bottom-up proteomics, at a constant flow rate of 300 nl/min. The full scan spectra were measured with a resolution of 120,000 within 50 ms maximum injection time, followed by MS2 scans with a resolution of 30,000 within 55 ms maximum injection time. The isolation window of the MS2 scan was set to 1.6 *m/z*, and only ions with 2 to 6 charges were triggered for the MS2 event. The normalized collision energy was set as 32. The dynamic exclusion time was set as 45 s. Compensation voltages were set at -45 V and -65 V to remove singly charged ions.

Construction of AltProts Database

This study used the Ribo-seq dataset we reported previously (12). Briefly, preprocessing of Ribo-seq raw data included adaptor removal using Cutadapt (33) (v 2.4, with parameters "–minimum-length 6 –discard-untrimmed –match-read-wildcards –max-n = 0.5"), low-quality trimming using Sickle (34) (v 1.33, with parameters "se -x -t sanger"). rRNA and tRNA contaminants were removed by aligning trimmed reads to mouse tRNA and rRNA sequences (5S, 5.8S, 18S, and 28S) using Bowtie 2 (35) (v1.0.1, with command "-q -L 20 –phred33 –end-to-end"). All remaining reads were mapped to the mouse reference genome GRCm 38 with a GTF annotation file (GENCODE vM25) using STAR (v 2.7.2 a) (36), and further unique mapped reads were extracted. Ten pipelines, RiboTISH (v 0.2.1) (37), ORFquant (v 0.99.0) (38), ORFRATER (39), RiboCode (v 1.2.11) (40), riboHMM (41), Ribotricer (v 1.3.1) (42), RiboWave (v 1.0) (43), RP-BP (v 2.0.0) (44), RibORF (v 1.0) (45), and PRICE (v 1.0.3 b) (46), were used to perform ORF and AltORF detection with the longest strategy under the default threshold setting (supplemental Table S1). The final set of actively translated ORFs with all near-cognate start codons (AUG, TUG, CUG, and GUG) followed by an in-frame stop codon in annotated transcripts was stringently filtered based on the requirement of a minimum length of 18 nucleotides and the expression of the ORF-containing gene at an above-background level, as described in a previous report (12). Those ORFs that pass above filtering criteria were classified into several categories based on their relative location with nearest annotated coding sequence (CDS), as described previously (1). In the classification result, ORFs were defined as annotated proteins. Upstream ORFs (uORFs) and downstream ORFs were defined as AltORFs originating from the 5′ untranslated regions (UTRs) and 3′UTRs of annotated protein-coding genes, respectively; long noncoding RNA ORFs (lncRNA-ORFs) were defined as AltORFs originating from transcripts currently annotated as lncRNAs; upstream overlapping ORFs (uoORFs), downstream overlapping ORFs, and internal out-of-frame ORFs were defined as AltORFs located upstream, downstream, and intermediate of CDS and out-frame overlapping with annotated CDSs, respectively. Finally, nucleic acid sequences of all actively translated AltORFs were converted into amino acid sequences in the FASTA format for the construction of protein databases.

Identification of Canonical Proteins and AltProts

The LC-MS/MS raw data were analyzed with MSFragger (version 3.3). The common parameters were set as below: precursor mass tolerance: 10 ppm, fragment mass tolerance: 0.02 Da; trypsin as enzyme; two missed cleavages; oxidation (methionine), acetyl (protein N-term), and TMT-6plex (N terminus) as variable modifications; carbamidomethylation (cysteine) and TMT-6plex (lysine) as fixed modification; the validation was performed using PeptideProphet; the FDR was set as 1%. Two different protein databases were used in this study: (1) Mouse OpenProt and sORF database were used for comparison of

enrichment methods. Mouse OpenProt protein database was derived from OpenProt (https://openprot.org, version number 1.6, 01 September 2020) (47) and contains 563,275 entries consisting of RefProts, novel isoforms, and AltProts predicted from both Ensembl and RefSeq. There were 503,679 entries in the *Mus musculus* AltProt protein database from sORF.org (http://www.sorfs.org, downloaded on 01 June 2021) (48); (2) in-house mouse AltProt database had 146,461 entries, which were used for AltProt discovery in TMT-labeled embryonic and adult livers. Identification of AltProts was always based on a peptide specific to the AltProt sequence and not common with the RefProts. The results from the custom database search were further filtered against the reference mouse proteins database (RefProt, containing Ensembl, NCBI RefSeq, and UniProtKB) using a stringent string-searching-based mapping algorithm to ensure that we did not report any known protein degradation, mutants, or isoforms.

We performed Gene Ontology (GO) analysis mainly based on annotated AltORFs, which are in the same genes that encode the related uORFs, downstream ORFs, and uoORFs, as well as lncRNA-ORFs that were encoded by the retained introns of protein-coding genes with known functions. GO analysis was performed with R package clusterProfiler (v4.0.5).

Validation of Novel AltProts with Parallel Reaction Monitoring

For parallel reaction monitoring (PRM), the samples were separated on the same LC-MS system by a 150 min gradient. Full scan spectra were measured with a resolution of 120,000 within a 50 ms maximum injection time, followed by targeted peptide MS2 scans with a resolution of 30,000 within a 60 ms maximum injection time under the 1.2 *m/z* isolation window. The normalized collision energy was set as 30. PRM data (tier 3 level) were processed with Skyline (version 21.1) software as described previously (49). Predicted retention time and MS/MS spectra were calculated based on two deep learning tools, DeepRT (50) and pDeep2 (51), respectively.

Identification of More AltProts Using the PRM Method

Twenty-seven AltProts were selected from the Ribo-seq-based AltProt database for targeted PRM analysis (tier 3 level) to identify additional AltProts. Briefly, a fragmentation inclusion list of theoretically predicted tryptic peptides in the selected AltProt was generated to identify more novel AltProts using high-resolution data-dependent scanning. A total of 51 unique peptide targets (corresponding to 27 AltProts) were selected in the inclusion list based on the following stringent screening criteria, including peptides uncommon to RefProts, sequence length greater than 7 amino acids, and the absence of methionine oxidation.

Experimental Design and Statistical Rationale

To test the performance of different AltProt enrichment methods, we performed triplicates for each enrichment method using adult C57BL/6 mice liver samples. To investigate AltProt expression during liver development, livers of embryonic (E15.5) and adult (P42) C57BL/6 mice in triplicates were used. Data were analyzed by a two-tailed unpaired Student's *t* test (unless otherwise indicated), and $p < 0.05$ was selected as the statistical limit of significance. We selected * and ** for $p < 0.05$ and $p < 0.01$, respectively. Unless otherwise stated, all the data in the graphs were expressed as arithmetic mean ± the SD from at least three repeated experiments.

RESULTS

Optimization of the Workflow for Microprotein Discovery

Considering the distinct lengths and properties of canonical RefProts and AltProts (52), the identification of AltProts with classical proteomic methods is analytically challenging. Therefore, we sought to improve the proteomics workflow at multiples steps, including protein extraction, AltProts enrichment, and peptides fractionation by comparing various conditions (Fig. 1). First, three widely employed protein extraction methods, RIPA lysis buffer, acidic lysis buffer, and boiling water, were tested for extracting AltProts from mouse liver homogenates. Significant protein loss was observed with acid lysis buffer and boiling water although they have been reported for extraction of small proteins by preferentially causing aggregation of high molecular weight proteins (22, 23). In contrast, RIPA lysis buffer offered much higher efficiency for total protein extraction and therefore was adopted in all following experiments (supplemental Fig. S1*A*).

SEC is the Most Efficient Method to Enrich AltProts

Next, we tested 10 methods from four categories to find the most efficient method for enriching AltProts from total proteins. In the first category “precipitation”, organic solvent or acids precipitated high molecular weight proteins and subsequently enriched AltProts. In the second category “size selection”, ultrafiltration tubes and SEC enabled separation of proteins by size. In the third category “solid phase separation”, the nonpolar reversed-phase sorbent trapped large hydrophobic proteins, while small and polar proteins were eluted and enriched. The fourth category was hexagonal mesoporous silica materials MCM-41, which enabled selectively enriched peptides and small protein through size selectivity and adsorptive mechanism. The efficiency of methods was compared side by side based on gel images and or MS analysis of the enriched proteins. Based on the tricine gel and glycine gel image, most methods were able to remove proteins larger than 40 kDa efficiently. However, the proteins enriched with these methods display vastly different profiles (Figs. 2*A* and S1, *B–F*). TCA precipitation, AA precipitation, C8 SPE, HLB SPE, 30-kDa-MWCO, and SEC resulted in strong protein bands and therefore were chosen for the following comparison with MS.

With equal protein amounts, the highest identification number was achieved by using SEC enrichment, with an average of 51 AltProts identified, which was more than twice that of the other methods (Fig. 2*B*). Meanwhile, although the intensity of RefProts was similar across all tested methods, the intensity of AltProts after SEC enrichment was five folds higher than that of other methods. SEC greatly reduced the difference between RefProts and AltProts in terms of MS intensity, which demonstrated its effectiveness in concentrating AltProts out of total lysates (Fig. 2*C*).

Characteristics of AltProts Enriched with Various Methods

Given the complementary nature of these enrichment methods, there were only a few AltProts commonly identified by using different categories of methods (Figs. 3*A* and S2). Although individual method did not yield a high number of AltProts, the methods collectively contributed more varieties

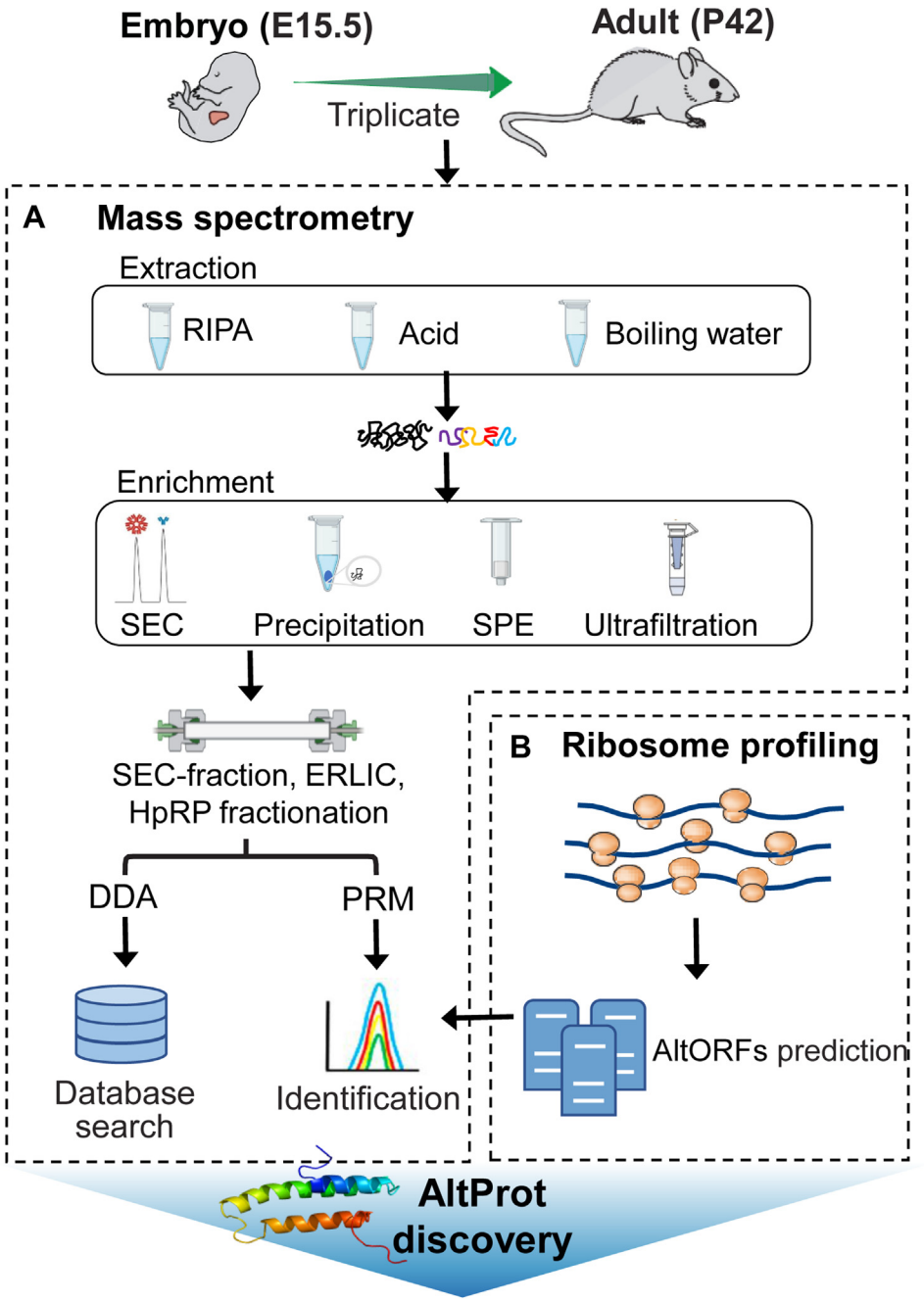


FIG. 1. **Schematic illustration of the workflow for MS-based discovery of AltORFs-encoded AltProts from mouse liver tissues.** *A*, designed workflow including extraction, enrichment, and fractionation methods for the discovery of AltProts. *B*, construction of AltProt database. The Ribo-seq data was screened by 10 different bioinformatics pipelines to find all possible translational AltORFs, which were then translated into potential translational products AltProts. AltORFs, alternative ORFs; AltProts, alternative proteins; DDA, data-dependent acquisition; MS, mass spectrometry; PRM, parallel reaction monitoring; SEC, size-exclusion chromatography; SPE, solid-phase extraction.

of AltProts. In our study, we found that No-enrich and SEC method were actually complementary in identifying different categories of AltProts. The reproducibility was higher within the same category than between categories. For example, over 60% of AltProts identified with TCA precipitation were reproducibly identified with AA precipitation. Among all the methods, SEC was found to be the most comprehensive. For AltProts that were identified by multiple enrichment methods, SEC resulted in the highest intensities (highlighted in red in Fig. 3*A*). Next, we analyzed the hydrophobicity and isoelectric point (pI) to investigate whether AltProt identification was associated with their biophysical properties (Fig. 3, *B* and *C*).

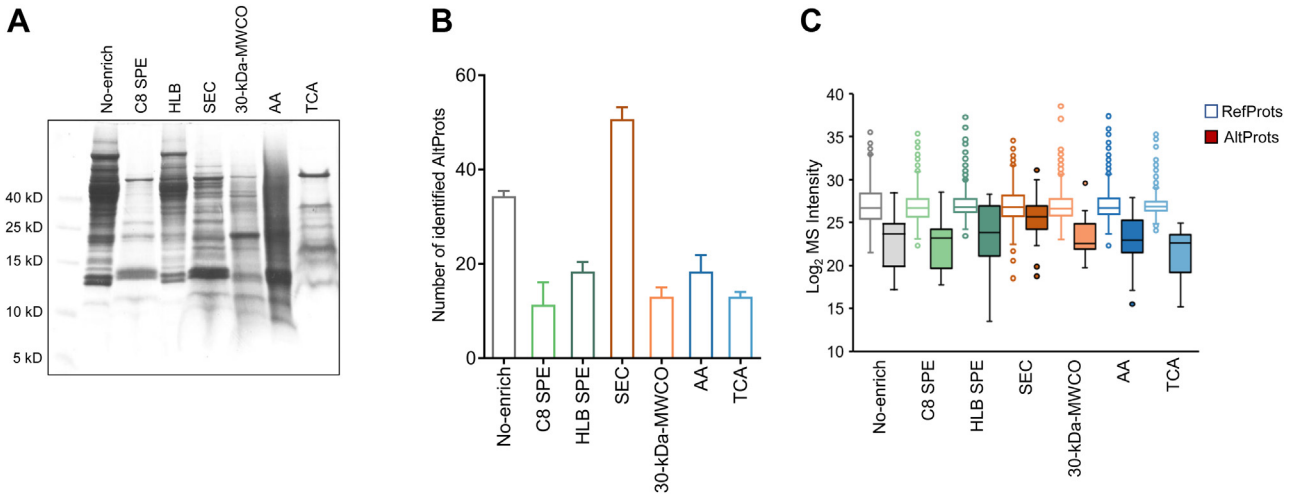


FIG. 2. **Comparison of different enrichment methods for AltProts in mouse livers.** *A*, liver lysates were lysed in RIPA lysis buffer followed by enrichment including C8 SPE, HLB SPE, SEC, 30-kDa-MWCO, AA precipitation, or TCA precipitation. The results from these enrichments were analyzed by SDS-PAGE (Coomassie stain). *B*, average number of detected AltProts using different enrichment methods. *C*, MS intensity of identified AltProts and RefProts in each enrichment method. 30-kDa-MWCO, 30-kDa-molecular weight cut-off ultrafiltration; AltProts, alternative proteins; AA, acetic acid; C8 SPE, C8 solid-phase extraction; HLB SPE, hydrophilic-lipophilic-balanced solid-phase extraction; MS, mass spectrometry; RefProts, reference proteins; SEC, size-exclusion chromatography; TCA, trichloroacetic acid.

As expected, the acid precipitation methods enriched more hydrophilic AltProts with lower GRAVY scores (Fig. 3*B*). TCA precipitation and AA precipitation preferentially enriched more AltProts with a high pI than other methods (Fig. 3*C*). Such differential biophysical properties partially explained the observation that a complementary pool of AltProts was enriched with different methods. SEC-based method enriched AltProts with evenly distributed hydrophobicity and pI and therefore was the most efficient method.

Fractionation Improves AltProt Discovery

Peptide fractionation using electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) and high pH reverse phase (HpRP) has been reported to improve the discovery of AltProts in prior studies (32, 53). SEC, which was found to be the most efficient and unbiased method for the enrichment of AltProts in our study, could also serve for protein fractionation to obtain four fractions of different molecular weight ranges. Therefore, we evaluated four fractionation methods for improving the depth of AltProt discovery, including SEC enrichment without fractionation (SEC), SEC enrichment into 4 fractions (SEC-fraction), SEC enrichment followed by ERLIC fractionation (SEC-ERLIC), and SEC enrichment followed by HpRP fractionation (SEC-HpRP) (Fig. 4*A*). SEC-fraction and ERLIC fractionation increased the number of AltProts by 1.4 to 1.6 folds. SEC-ERLIC led to the highest number of AltProts, while SEC-fraction was the most time- and effort-effective, as it could enrich and fractionate AltProts simultaneously within 15 min (Fig. 4*A*). The intensities of AltProts even showed slight increase after SEC-fraction and SEC-ERLIC (Fig. 4*B*).

Optimized Workflow Enables Discovery of AltProts in Embryonic Liver Development

Next, we applied the optimized workflow in combination with TMT-based quantification to investigate AltProts expression during liver development (Fig. 5*A*). Total protein lysates were extracted from the livers of embryonic (E15.5) or adult (P42) C57BL/6 mice in triplicates followed by SEC-ERLIC and TMT-based quantification. As we previously studied the protein translation landscape of mouse livers during development by using Ribo-seq, we were able to construct a liver-specific protein database based on Ribo-seq results and search the MS data against it. Although our customized database was much smaller than public databases (47, 48), we were able to detect 5146 RefProts and 89 AltProts reproducibly from embryonic and adult mouse livers (supplemental Table S2). Representative mass spectra of AltProt peptides were listed in supplemental Figs. S3 and S4. Despite the fact that MS and Ribo-seq were two completely different techniques, the measured fold change between embryonic and adult livers showed a positive correlation with R equals to 0.71 (Fig. 5*B*), indicating that both techniques can precisely capture the overall changes of proteome during development. A large majority of the AltProts identified were encoded by lncRNA-ORFs (74%) and uoORFs (22%) and some were from uORF, downstream ORF, and internal out-of-frame ORFs (Fig. 5*C*). The identified AltProts showed similar hydrophobicity with RefProts (supplemental Fig. S5). Furthermore, 39 AltProts were found to be differentially expressed (Fig. 5*D*). GO analysis of AltORFs showed that AltProts upregulated in embryonic livers were involved in RNA splicing and processing, whereas AltProts upregulated in

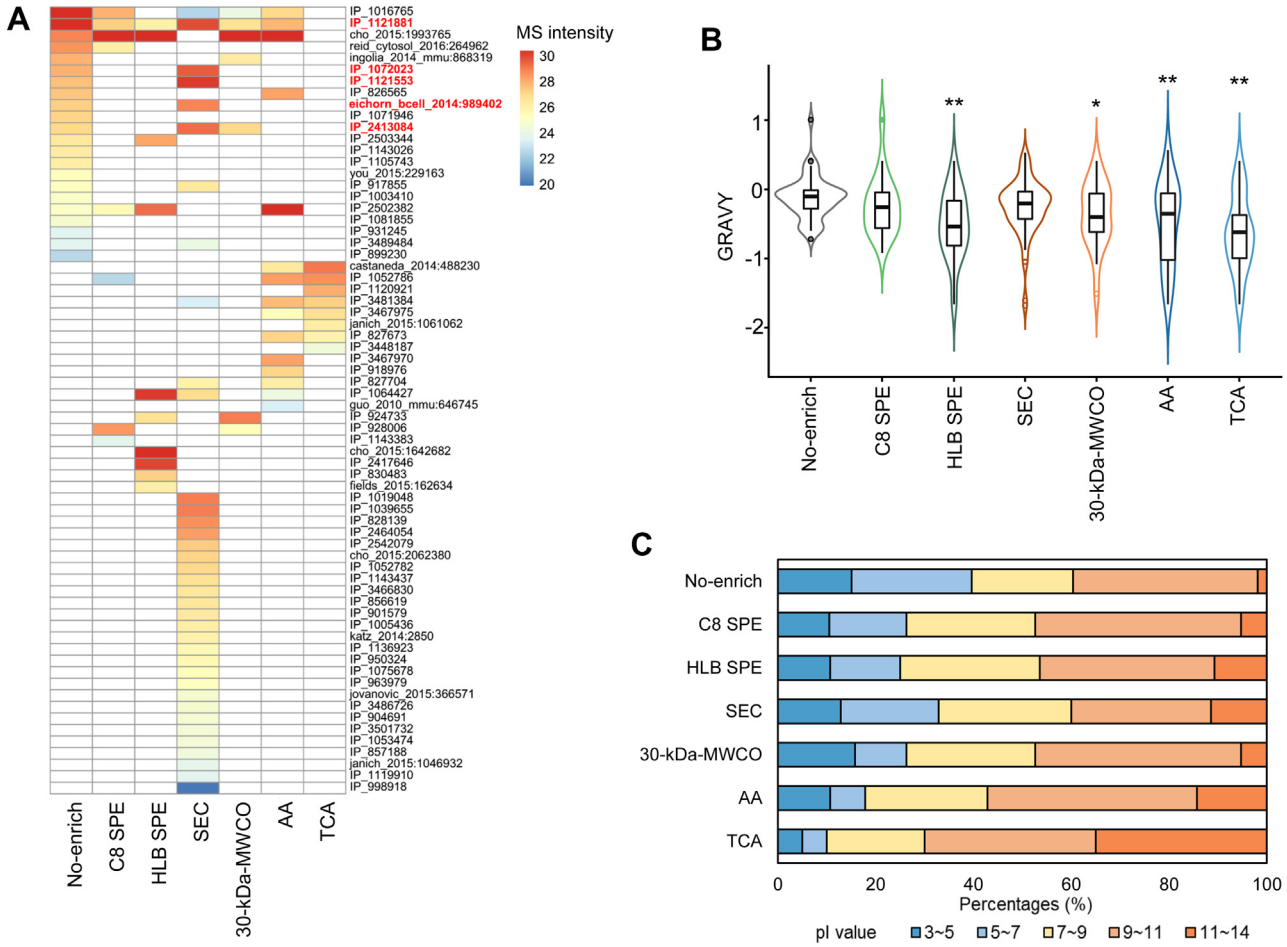


FIG. 3. **Performance of different approaches for enriching AltProts.** Distribution of MS intensity (*A*), hydrophobicity (*B*), and isoelectric point (*C*) of identified AltProts from different enrichment methods. *$p < 0.05$ *versus* No-enrich; **$p < 0.01$ *versus* No-enrich. 30-kDa-MWCO, 30-kDa-molecular weight cut-off ultrafiltration; AltProts, alternative proteins; AA, acetic acid; C8 SPE, C8 solid-phase extraction; GRAVY, grand average of hydropathicity index; HLB SPE, hydrophilic-lipophilic-balanced solid-phase extraction; pI, isoelectric point; MS, mass spectrometryy; SEC, size-exclusion chromatograph; TCA, trichloroacetic acid.

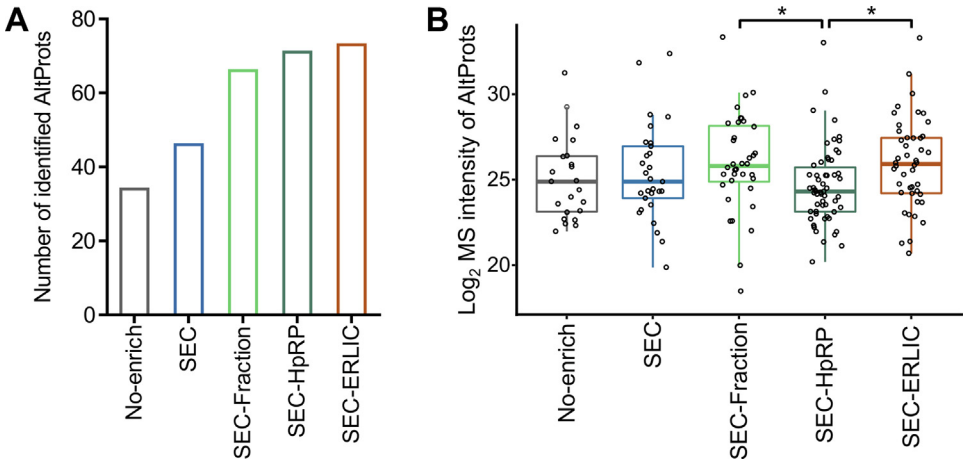


FIG. 4. **The effect of fractionation methods on AltProt discovery.** The number of identified AltProts (*A*) and MS intensity of AltProts (*B*) before and after fractionation. * indicates $p < 0.05$ for comparison. AltProts, alternative Proteins; MS, mass spectrometry; SEC, size-exclusion chromatography; SEC-fraction, SEC enrichment into 4 fractions; SEC-ERLIC, SEC enrichment followed by ERLIC fractionation; SEC-HpRP, SEC enrichment followed by HpRP fractionation.

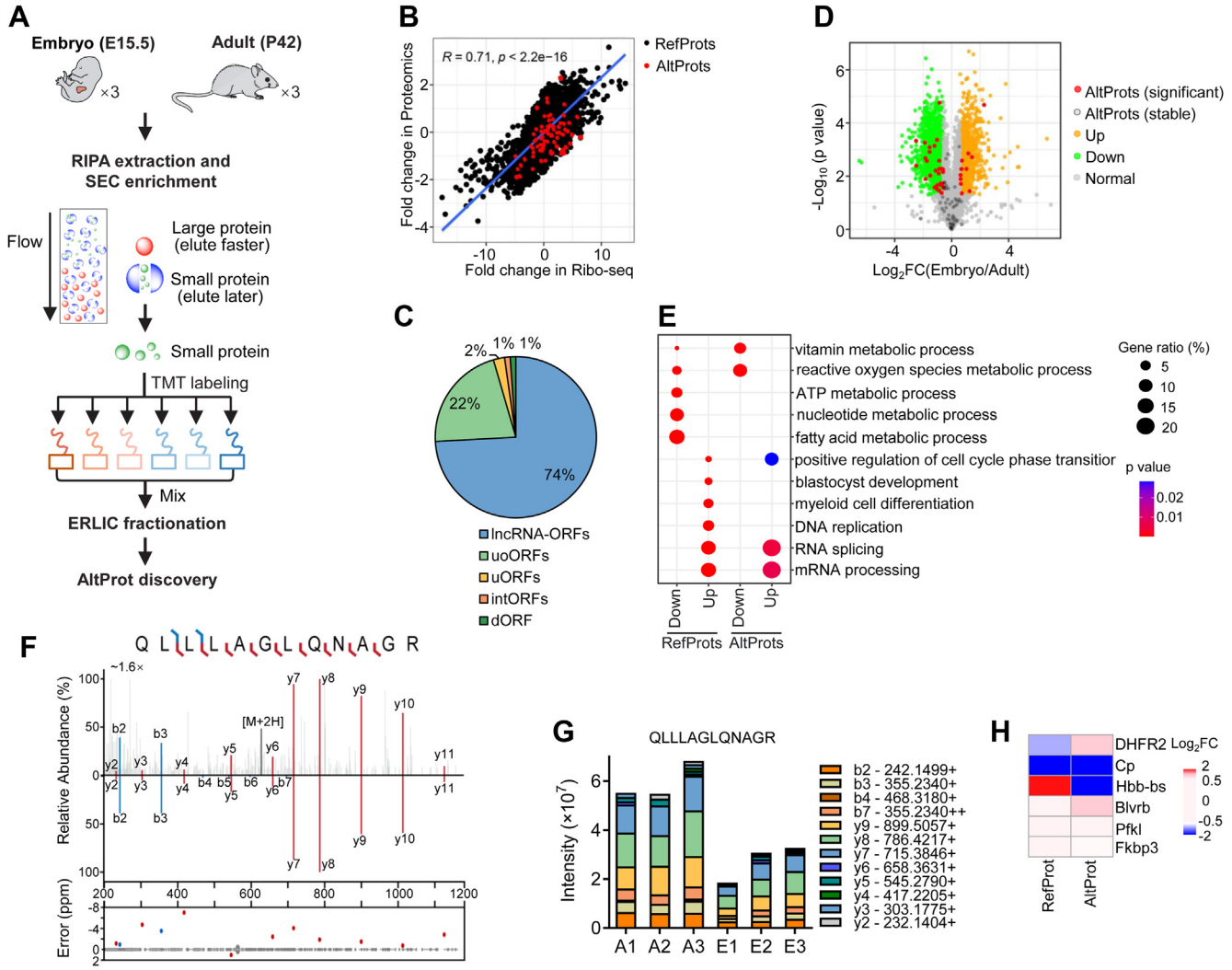


FIG. 5. **Discovery of AltProts in embryonic and adult livers.** *A*, SEC-ERLIC workflow for the discovery of AltProts in adult and embryonic livers using a TMT-based MS approach. *B*, the correlation of protein expression differences from embryonic and adult mice detected using two techniques Ribo-seq and MS-based proteomic approaches. *C*, RNA type distribution of identified AltProts. *D*, volcano plot of identified RefProts and AltProts in adult and embryonic livers. *Orange* and *green* dots represent the upregulated and downregulated RefProts, respectively. *Red* and *dark gray* dots represent the significant changed AltProts and stable AltProts, respectively. (*p* values <0.05; |fold change (FC)| > 1.5). *E*, GO analysis of the significantly changed RefProts and AltProts. *F*, an example of the experimental spectrum and the predicted spectrum of noncanonical peptide QLLLAGLQNAGR. *G*, the corresponding peak areas of the representative noncanonical peptide QLLLAGLQNAGR in embryonic and adult livers using PRM method. *H*, heatmap of MS intensity of pairs of AltProts and their primary RefProts from the same gene. AltProts, alternative protein; Cp, ceruloplasmin; DHFR, dihydrofolate reductases; ERLIC, electrostatic repulsion-hydrophilic interaction chromatography; Hbb-bs, beta-globin; lncRNA, long noncoding RNA; MS, mass spectrometry; RefProts, reference proteins; Ribo-seq, ribosome profiling; SEC, size-exclusion chromatography.

adult livers were enriched in metabolic pathways (Fig. 5*E*). The biological pathways were consistent with that of RefProts, suggesting the functional importance of AltProts in liver development. We further employed an alternative MS strategy, PRM, to validate the identification and quantification of novel AltProts (supplemental Figs. S6 and S7). For example, the MS2 spectrum of the noncanonical peptide QLLLAGLQNAGR highly agreed with its predicted spectrum (Fig. 5*F*). The amount of this peptide was significantly downregulated in three embryonic livers compared to adult livers (Fig. 5*G*). In the end, we sought to understand the relationship between AltProts and RefProts. We specifically searched for actively translated AltORFs within the 5′- and 3′-UTRs of canonical ORFs. With stringent criteria, six pairs of AltProts and their primary RefProts from the same gene were detected by MS in the same experiment (Fig. 5*H* and supplemental Table S3). Among them, dihydrofolate reductase, ceruloplasmin, and beta-globin (Hbb-bs) and their corresponding AltProts were

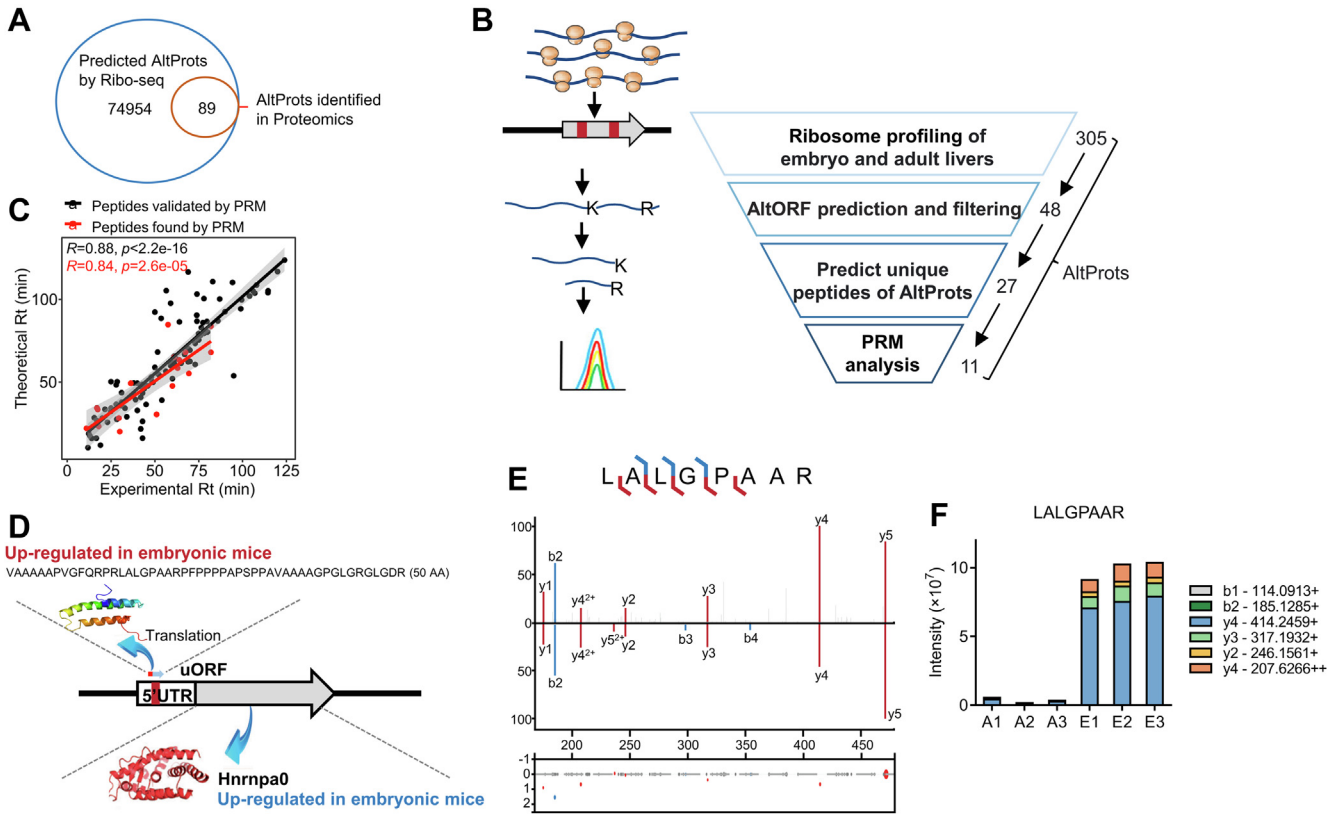


FIG. 6. **Discovery of additional AltProts using PRM method.** *A*, Venn diagram of AltProts predicted by Ribo-seq and AltProts detected in targeted MS-based proteomics. *B*, flow chart of the AltProt discovery using PRM method. *C*, the correlation between the experimental retention time and the theoretical retention time of identified AltProts. *Red* dots represent the peptides found by targeted PRM, *black* dots represent the peptides found by SEC-ERLIC workflow and validated by PRM method. *D*, a typical schematic diagram of an AltProt expressed on the uORF of Hnrnpa0 gene encoding HnRNPA0. *E*, an example of the experimental spectrum and the predicted spectrum of noncanonical peptide LALGPAAR. *F*, the peak areas of the representative noncanonical peptide LALGPAAR using PRM method. AltProts, alternative proteins; ERLIC, electrostatic repulsion-hydrophilic interaction chromatography; HnRNPA0, heterogeneous nuclear ribonucleoprotein A0; MS, mass spectrometry; PRM, parallel reaction monitoring; Rt, retention time; SEC, size-exclusion chromatography; uORFs, upstream ORFs.

significantly changed between embryonic and adult mice, indicating a potential *cis* gene regulatory effect between AltORFs and the corresponding primary ORFs.

Integrating Ribo-seq and PRM to Discover Additional AltProts

It is noteworthy that the number of AltORFs being translated predicted with Ribo-seq was dramatically higher than that detected by MS (Fig. 6*A*). Therefore, we tested an alternative approach by integrating Ribo-seq with targeted MS method to discover additional AltProts that were undetectable using conventional shotgun proteomics (Fig. 6*B*). To provide a precise list of AltORFs, we used 10 different bioinformatics pipelines to predict possible translational AltORFs and kept only those that were reproducibly reported with at least two pipelines. The full-length sequences of AltProts were subsequently generated by using 3-frame translation. Out of the 27 selected AltProts with unique peptides, 11 were detectable with PRM (supplemental Table S4). The retention time showed a high correlation between theoretical and experimental values (R = 0.84-0.88), indicating a high confidence in AltProt identification (Fig. 6*C*). Even though the identification rate was 40% with this approach, it could serve as a supplement to traditional shotgun proteomics and possibly allow detection of AltProts with low abundance. For example, peptide LALGPAAR was from a novel AltProt with 50 amino acids. This AltProt was encoded by the 5′-UTR sequence of Hnrnpa0 gene encoding Heterogeneous nuclear ribonucleoprotein A0 (HnRNPA0) (Fig. 6*D*). Both this AltProt and RefProt HnRNPA0 were significantly upregulated in embryonic livers (Fig. 6, *E* and *F*). Hnrnpa0 plays an important role in myeloid cell differentiation (54) as well as neurodevelopment (55). The identification of its upstream AltORF could lead to novel regulatory mechanisms of this important protein.

DISCUSSION

In this study, we tested various methods and found “RIPA extraction/SEC enrichment/ERLIC fractionation” was the most

efficient strategy for identifying AltProts with MS. With this strategy, we investigated novel AltProts in embryonic and adult mouse livers.

Although a few elegant works using MS for AltProt detection have been reported in recent years, but the number of AltProt identified to our knowledge still varies widely, from tens (2, 20, 28) to hundreds (21, 22, 24). This is probably explained by the different enrichment and analysis methods used and the sample variation. In our study, 89 novel AltProts were identified and compared between embryonic and adult mice, although not the highest, it is based on only one sample type. Our study is so far the most comprehensive one to optimize multiple steps and various combinations for AltProt identification and we found that different workflows favor different types of AltProts. According to our results, the SEC-based enrichment outperformed other methods in terms of identification number, specificity, and reproducibility of low-abundant AltProts. In contrast, sample loss and batch-to-batch variability were observed in the 30-kDa-MWCO method, probably due to nonspecific protein binding to the filter membrane (23). SPE cartridges extracted a limited number of peptides (56) probably due to the undesired retention of relatively large and hydrophobic proteins by nonpolar materials.

An additional advantage of the SEC approach is its capability to separate AltProts by size (31, 57), enabling analysis depth comparable to HpRP or ERLIC fractionation without extra cost of time and effort. Besides, SEC does not require specific buffer conditions and therefore is usually compatible with downstream experiments like top-down MS and functional characterization. SEC-based approach has great potential in future AltProts studies.

We also compared two types of SEC columns for AltProt enrichment, considering that flow rate, particle pore size, sample volume, and CV could all influence the separation efficiency. Conventional SEC requires relatively large amounts of proteins and more time due to the large CV (31, 58). Scaling-up the volumes would also dilute the proteins of interest, which impeded the detection sensitivity. We found that SEC column with smaller CV (3 ml) outperformed the one with larger CV (24 ml). Smaller column is also more efficient to complete the enrichment and fractionation simultaneously within 15 min.

We acknowledged that the identification number and confidence of AltProts are highly dependent on the size and quality of database and therefore decided to use only a non-inflated, customized database. In this study, 89 AltProts were identified from embryonic and adult mouse livers. Our results showed that many AltProts that were upregulated in embryonic livers were involved in RNA splicing, RNA processing, and regulation of cell cycle transition (Fig. 5*E*). RNA splicing is a crucial process for changing mature mRNA into functional protein, a process that is required during mammalian embryogenesis to generate a viable organism from a single cell (59). RNA processing maintains protein synthesis during early developmental stages (60). Cell cycle transition determines cell-fate transition and embryonic development (61). All of these biological pathways are important in the embryonic development.

One of the important roles that AltORFs play is to regulate the translation of downstream canonical ORFs (12, 62). The translation of uORFs of GCN4 promoted the release of ribosomes from the same transcript, preventing ribosomes from reaching start codon and subsequent inhibiting translation of the GCN4 gene (63). Some other uORFs positively regulated the translation of the downstream canonical ORFs (64). In our study, two uORFs and corresponding canonical ORFs of hnRNPA0 and hnRNPA2/B1 showed significant activation in embryonic livers. The observation was highly consistent in both MS and Ribo-seq results. hnRNPA0 was reported to affect myeloid cell differentiation and neurodevelopment (54, 55). hnRNPA2/B1 regulated mammalian embryonic development (65). We speculate that AltProts encoded by uORF could promote the expression of downstream CDS, thereby regulating liver development. The detailed relationship in functions and mechanisms will be studied in due course.

Although we have discovered interesting AltProts involved in embryonic development with an optimized approach, the total identification number of AltProts was not comparable to that of RefProts. One possible reason is that we used a small, specific database and stringent cut-offs to filter the findings. However, the intrinsic short length and likely low abundance of AltProts are more important factors. Therefore, improvement in MS instrumentation with high sensitivity is needed in the future studies of AltProts.

DATA AVAILABILITY

The data that support the findings of this study have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) *via* the PRIDE partner repository with the dataset identifier PXD033940.

Supplemental data—This article contains supplemental data (2, 22, 26-29).

Acknowledgments—We acknowledge the funding support from Research Grants Council-GRF 15305821, CRF Equipment C5033-19E, and RGC-RIF R5050-18, support from Laboratory for Synthetic Chemistry and Chemical Biology Limited (LSCCB) and Centre for Eye and Vision Research (CEVR) under the Health@InnoHK Programme launched by ITC, HKSAR. We thank the Prof. Mankin Wong, PolyU Research Facilities ULS and UCEA, and research institute/center RiFood and RCMI for technical support.

Authors contributions—Yi. Y., H. W., L. C., Z. X., and Q. Z. methodology; Yi. Y., Y. Z., L. C., and Ya. Y. investigation; Yi. Y., Y. Z., and Q. Z. formal analysis; Yi. Y. and Q. Z. writing–original draft.

Conflict of Interest—The authors declare that there are no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: 30-kDa-MWCO, 30-kDa-molecular weight cut-off ultrafiltration; AA, acetic acid; ACN, acetonitrile; AltORF, alternative ORF; AltProt, alternative protein; CV, column volume; ERLIC, electrostatic repulsion-hydrophilic interaction chromatography; GO, gene ontology; HnRNPA0, Heterogeneous nuclear ribonucleoprotein A0; HpRP, high pH reverse phase; lncRNA, long noncoding RNA; MS, mass spectrometry; MTBE, methyl tert-butyl ether; PRM, parallel reaction monitoring; RefProt, reference protein; SEC, size-exclusion chromatography; sORF, small ORF; SPE, solid phase extraction; TCA, trichloroacetic acid; TEAF, triethylammonium formate; uoORF, upstream overlapping ORF; uORFs, upstream ORFs.

Received June 21, 2022, and in revised form, November 15, 2022
Published, MCPRO Papers in Press, December 7, 2022, https://doi.org/10.1016/j.mcpro.2022.100480

REFERENCES

1. Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., *et al*. (2022) Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999
2. Cardon, T., Hervé, F., Delcourt, V., Roucou, X., Salzet, M., Franck, J., *et al*. (2020) Optimized sample preparation workflow for improved identification of ghost proteins. *Anal. Chem.* **92**, 1122–1129
3. Cardon, T., Fournier, I., and Salzet, M. (2021) Shedding light on the ghost proteome. *Trends Biochem. Sci.* **46**, 239–250
4. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.-P., *et al*. (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516
5. Laumont, C. M., Daouda, T., Laverdure, J. P., Bonneil, É., Caron-Lizotte, O., Hardy, M. P., *et al*. (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **7**, 10238
6. Ruiz Cuevas, M. V., Hardy, M. P., Hollý, J., Bonneil, É., Durette, C., Courcelles, M., *et al*. (2021) Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815
7. Yin, X., Hu, J., and Xu, H. (2018) Distribution of micropeptide-coding sORFs in transcripts. *Chin. Chem. Lett.* **29**, 1029–1032
8. Zhang, S., Reljić, B., Liang, C., Kerouanton, B., Francisco, J. C., Peh, J. H., *et al*. (2020) Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat. Commun.* **11**, 1312
9. Cleyle, J., Hardy, M.-P., Minati, R., Courcelles, M., Durette, C., Lanoix, J., *et al*. (2022) Immunopeptidomic analyses of colorectal cancers with and without microsatellite instability. *Mol. Cell Proteomics* **21**, 100228
10. Apavaloaei, A., Hesnard, L., Hardy, M.-P., Benabdallah, B., Ehx, G., Thériault, C., *et al*. (2022) Induced pluripotent stem cells display a distinct set of MHC I-associated peptides shared by human cancers. *Cell Rep.* **40**, 111241
11. Koh, M., Ahmad, I., Ko, Y., Zhang, Y., Martinez, T. F., Diedrich, J. K., *et al*. (2021) A short ORF-encoded transcriptional regulator. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2021943118
12. Wang, H., Wang, Y., Yang, J., Zhao, Q., Tang, N., Chen, C., *et al*. (2021) Tissue- and stage-specific landscape of the mouse translatome. *Nucl. Acids Res.* **49**, 6165–6180
13. Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., *et al*. (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271–275
14. Pauli, A., Norris, M. L., Valen, E., Chew, G. L., Gagnon, J. A., Zimmerman, S., *et al*. (2014) Toddler: an embryonic signal that promotes cell movement *via* apelin receptors. *Science* **343**, 1248636
15. Kustatscher, G., Collins, T., Gingras, A.-C., Guo, T., Hermjakob, H., Ideker, T., *et al*. (2022) Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Met.* https://doi.org/10.1038/s41592-41022-01454-x
16. Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020) Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468
17. Franceschetti, T., Zhao, Q., Vincent, K., Perreault, C., Milosevic, S., and Sommermeyer, D. (2021) Targetable immunogenic tumor specific antigens can be identified in non-coding regions of the genome. *Cancer Res.* **81**, 1520
18. Hu, D. C., Liu, Q., Xu, H. B., Cui, H. R., Yu, S. W., Yang, X. D., *et al*. (2005) A novel protein found in selenium-rich silkworm pupas. *Chin. Chem. Lett.* **16**, 1347–1350
19. Thibault, P., and Perreault, C. (2022) Immunopeptidomics: reading the immune signal that defines self from nonself. *Mol. Cell Proteomics* **21**, 100234
20. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., *et al*. (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64
21. Fabre, B., Choteau, S. A., Duboé, C., Pichereaux, C., Montigny, A., Korona, D., *et al*. (2022) Depth exploration of the alternative proteome of Drosophila melanogaster. *Front. Cell Dev. Biol.* **10**, 901351
22. Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., *et al*. (2016) Improved identification and analysis of small open reading frame encoded polypeptides. *Anal. Chem.* **88**, 3967–3975
23. Cassidy, L., Kaulich, P. T., Maaß, S., Bartel, J., Becher, D., and Tholey, A. (2021) Bottom-up and top-down proteomic approaches for the identification, characterization, and quantification of the low molecular weight proteome with focus on short open reading frame-encoded peptides. *Proteomics* **21**, e2100008
24. Zhang, Q., Wu, E., Tang, Y., Cai, T., Zhang, L., Wang, J., *et al*. (2021) Deeply mining a universe of peptides encoded by long noncoding RNAs. *Mol. Cell Proteomics* **20**, 100109
25. Buszczak, M., Signer, R. A. J., and Morrison, S. J. (2014) Cellular differences in protein synthesis regulate tissue homeostasis. *Cell* **159**, 242–251
26. Dingess, K. A., van den Toorn, H. W. P., Mank, M., Stahl, B., and Heck, A. J. R. (2019) Toward an efficient workflow for the analysis of the human milk peptidome. *Anal. Bioanal. Chem.* **411**, 1351–1363
27. Cassidy, L., Kaulich, P. T., and Tholey, A. (2019) Depletion of high-molecular-mass proteins for the identification of small proteins and short open reading frame encoded peptides in cellular proteomes. *J. Proteome Res.* **18**, 1725–1734
28. Li, N., Zhou, Y., Wang, J., Niu, L., Zhang, Q., Sun, L., *et al*. (2020) Sequential precipitation and delipidation enables efficient enrichment of low-molecular weight proteins and peptides from human plasma. *J. Proteome Res.* **19**, 3340–3351
29. Du, Y., Wu, D., and Guan, Y. (2016) Further investigation of a peptide extraction method with mesoporous silica using high-performance liquid chromatography coupled with tandem mass spectrometry. *J. Sep. Sci.* **39**, 2156–2163
30. Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., and Krijgsveld, J. (2019) Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85
31. Harney, D. J., Hutchison, A. T., Su, Z., Hatchwell, L., Heilbronn, L. K., Hocking, S., *et al*. (2019) Small-protein enrichment assay enables the rapid, unbiased analysis of over 100 low abundance factors from human plasma. *Mol. Cell Proteomics* **18**, 1899–1915
32. Ma, J., Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., *et al*. (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.* **13**, 1757–1765
33. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet. J.* **17**, 10–12
34. Joshi, N., and Fass, J. (2011) Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for Fastq Files (version 1.33) [software].

35. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25
36. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., *et al*. (2013) Star: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
37. Zhang, P. A.-O., He, D., Xu, Y., Hou, J., Pan, B. F., Wang, Y. A.-O., *et al*. (2017) Genome-wide identification and differential analysis of translational initiation. *Nat. Commun.* **8**, 1749
38. Calviello, L., Hirsekorn, A., and Ohler, U. (2020) Quantification of translation uncovers the functions of the alternative transcriptome. *Nat. Struct. Mol. Biol.* **27**, 717–725
39. Fields, A. P., Rodriguez, E. H., Jovanovic, M., Stern-Ginossar, N., Haas, B. J., Mertins, P., *et al*. (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell* **60**, 816–827
40. Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., and Yang, X. (2018) De novo annotation and characterization of the translatome with ribosome profiling data. *Nucl. Acids Res.* **46**, e61
41. Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., *et al*. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5**, e13328
42. Choudhary, S., Li, W., and Smith, A. D. (2020) Accurate detection of short and long active ORFs using Ribo-seq data. *Bioinformatics* **36**, 2053–2059
43. Xu, Z., Hu, L., Shi, B., Geng, S., Xu, L., Wang, D., *et al*. (2018) Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. *Nucl. Acids Res.* **46**, e109
44. Malone, B., Atanassov, I., Aeschimann, F., Li, X., Großhans, H., and Dieterich, C. (2017) Bayesian prediction of RNA translation from ribosome profiling. *Nucl. Acids Res.* **45**, 2960–2972
45. Ji, Z., Song, R., Regev, A., and Struhl, K. (2015) Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890
46. Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D. J., Weekes, M. P., *et al*. (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat. Met.* **15**, 363–366
47. Brunet, M. A., Brunelle, M., Lucier, J. F., Delcourt, V., Levesque, M., Grenier, F., *et al*. (2019) OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucl. Acids Res.* **47**, D403–D410
48. Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucl. Acids Res.* **46**, D497–D502
49. Guo, H., Yang, Y., Zhang, Q., Deng, J.-R., Yang, Y., Li, S., *et al*. (2022) Integrated mass spectrometry reveals celastrol as a novel catechol-O-methyltransferase inhibitor. *ACS Chem. Biol.* https://doi.org/10.1021/acschembio.1022c00011
50. Ma, C., Ren, Y., Yang, J., Ren, Z., Yang, H., and Liu, S. (2018) Improved peptide retention time prediction in liquid chromatography through deep learning. *Anal. Chem.* **90**, 10881–10888
51. Zeng, W.-F., Zhou, X.-X., Zhou, W.-J., Chi, H., Zhan, J., and He, S.-M. (2019) MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. *Anal. Chem.* **91**, 9724–9731
52. Samandi, S., Roy, A. V., Delcourt, V., Lucier, J.-F., Gagnon, J., Beaudoin, M. C., *et al*. (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* **6**, e27860
53. Cassidy, L., Prasse, D., Linke, D., Schmitz, R. A., and Tholey, A. (2016) Combination of bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS enhances coverage of proteome and low molecular weight short open reading frame encoded peptides of the archaeon methanosarcina mazei. *J. Proteome Res.* **15**, 3773–3783
54. Young, D. J., Stoddart, A., Nakitandwe, J., Chen, S.-C., Qian, Z., Downing, J. R., *et al*. (2014) Knockdown of Hnrnpa0, a del (5q) gene, alters myeloid cell fate in murine cells through regulation of AU-rich transcripts. *Haematologica* **99**, 1032–1040
55. Gillentine, M. A., Wang, T., Hoekzema, K., Rosenfeld, J., Liu, P., Guo, H., *et al*. (2021) Rare deleterious mutations of HNRNP genes result in shared neurodevelopmental disorders. *Genome Med.* **13**, 63
56. Kononikhin, A. S., Starodubtseva, N. L., Bugrova, A. E., Shirokova, V. A., Chagovets, V. V., Indeykina, M. I., *et al*. (2016) An untargeted approach for the analysis of the urine peptidome of women with preeclampsia. *J. Proteomics* **149**, 38–43
57. Liu, Y., Xun, X.-H., Yi, J.-M., Xiang, Y., and Hua, J. (2017) Discovery of lung squamous carcinoma biomarkers by profiling the plasma peptide with LC/MS/MS. *Chin. Chem. Lett.* **28**, 1093–1098
58. Müller, S. A., Findeiß, S., Pernitzsch, S. R., Wissenbach, D. K., Stadler, P. F., Hofacker, I. L., *et al*. (2013) Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics. *J. Proteomics* **86**, 27–42
59. Revil, T., Gaffney, D., Dias, C., Majewski, J., and Jerome-Majewska, L. A. (2010) Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* **11**, 399
60. Macaulay, A., Scantland, S., and Robert, C. (2011) *RNA Processing during Early Embryogenesis: Managing Storage, Utilisation and Destruction*. IntechOpen, London
61. Zhao, J., Lu, P., Wan, C., Huang, Y., Cui, M., Yang, X., *et al*. (2021) Cell-fate transition and determination analysis of mouse male germ cells throughout development. *Nat. Commun.* **12**, 6839
62. Zhang, H., Wang, Y., and Lu, J. (2019) Function and evolution of upstream ORFs in eukaryotes. *Trends Biochem. Sci.* **44**, 782–794
63. Gunišová, S., Beznosková, P., Mohammad, M. P., Vlčková, V., and Valášek, L. S. (2016) In-depth analysis of cis-determinants that either promote or inhibit reinitiation on GCN4 mRNA after translation of its four short uORFs. *RNA* **22**, 542–558
64. Starck, S. R., Tsai, J. C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., *et al*. (2016) Translation from the 5' untranslated region shapes the integrated stress response. *Science* **351**, aad3867
65. Kwon, J., Jo, Y.-J., Namgoong, S., and Kim, N.-H. (2019) Functional roles of hnRNPA2/B1 regulated by METTL3 in mammalian embryonic development. *Sci. Rep.* **9**, 8640