Towards a universal sample preparation method for denaturing top-down proteomics of complex proteomes

Zhichang Yang[1], Xiaojing Shen[1], Daoyang Chen[1], Liangliang Sun[1,*]

[1] Department of Chemistry, Michigan State University, 578 S Shaw Ln, East Lansing, MI 48824 USA

* Corresponding author. E-mail: lsun@chemistry.msu.edu

Abstract

A universal and standardized sample preparation method becomes vital for denaturing top-down proteomics (dTDP) to advance the scale and accuracy of proteoform delineation in complex biological systems. It needs to have high protein recovery, minimum bias, good reproducibility, and compatibility with downstream mass spectrometry (MS) analysis. Here we employed a lysis buffer containing sodium dodecyl sulfate (SDS) for extracting proteoforms from cells, and for the first time, compared membrane ultrafiltration (MU), chloroform-methanol precipitation (CMP), and single-spot solid-phase sample preparation using magnetic beads (SP3) for proteoform cleanup for dTDP. The MU method outperformed CMP and SP3 methods, resulting in high and reproducible protein recovery from both *E. coli* cell (59±3%) and human HepG2 cell (86±5%) samples without a significant bias. Single-shot capillary zone electrophoresis (CZE)-MS/MS analyses of the prepared *E. coli* and HepG2 cell samples using the MU method identified 821 and 516 proteoforms, respectively. Nearly 30% and 50% of the identified *E. coli* and HepG2 proteins are membrane proteins. CZE-MS/MS identified 94 histone proteoforms from the HepG2 sample with various post-translational modifications, including acetylation, methylation, and phosphorylation. Our results suggest that combining the SDS-based protein extraction and the MU-based protein cleanup could be a universal sample preparation method for dTDP. The MS raw data have been deposited to the ProteomeXchange Consortium with the data set identifier PXD018248.

Keywords: sample preparation, denaturing top-down proteomics, SDS, membrane ultrafiltration, chloroform-methanol precipitation, SP3, CZE-MS/MS, membrane proteins, histone, PTMs

Introduction

Denaturing top-down proteomics (dTDP) aims to delineate proteoforms in cells with high throughput.[1-3] It is becoming an important tool for gaining a better understanding of protein function in disease and development.[3,4] For mass spectrometry (MS)-based dTDP, tremendous efforts have been made in boosting proteoform liquid-phase separation,[5-15] improving MS instrumentation,[8,16-18] and developing new bioinformatics tools for proteoform identifications (IDs) through database search,[19-21] leading to thousands of proteoform IDs from a complex proteome. The Kelleher group integrated three dimensional (3D) liquid-phase separations (isoelectric focusing (IEF), gel-eluted liquid fraction entrapment electrophoresis (GELFrEE), and reversed-phase liquid chromatography (RPLC)) and a 12T FT-ICR mass spectrometer for large-scale dTDP of human cells, enabling over 3 000 proteoform IDs.[5] Anderson *et al.* showed that coupling 2D GELFrEE-RPLC separation to a 21T FT-ICR mass spectrometer identified over 3 000 proteoforms from human cancer cells.[8] The Ge group combined 2D size exclusion chromatography (SEC)-RPLC separation and a Q-TOF mass spectrometer for dTDP, detecting 5 000 different proteoforms from heart tissues.[9] Our group coupled a 3D SEC-RPLC-capillary zone electrophoresis (CZE) separation to an Orbitrap mass spectrometer for dTDP and identified nearly 6 000 proteoforms from *E. coli* cells.[11] The Wu group developed a 2D-RPLC system for high-capacity proteoform separation, and identified 2778 proteoforms from HeLa cell lysates.[12] The Paša-Tolić group developed a high-capacity RPLC system for proteoform separation via using an 80-cm long RPLC column, enabling 1665 proteoform IDs from bacteria with an Orbitrap mass spectrometer.[13] Recently, our group employed a 1.5-meters long capillary for CZE separation of proteoforms and coupling the CZE separation to an Orbitrap mass spectrometer enabled the identification and quantification of thousands of proteoforms from zebrafish brain samples using hundreds of nanograms of protein materials.[14]

The development of large-scale dTDP underlines the importance of a standardized and universal sample preparation method to achieve comprehensive extraction of proteins from biological samples with high recovery, good reproducibility, minimum bias and absence of MS incompatible salts, chaotropes and detergents. Protein extraction using a cell lysis buffer containing chaotropic agents or detergents, and protein sample cleanup

before MS with ultrafiltration or precipitation have been suggested as efficient approaches for preparation of protein samples for MS.[22] Sodium dodecyl sulfate (SDS) is an extremely efficient detergent for solubilizing and denaturing proteins, making it widely used in proteomics studies for protein extraction.[23] However, higher than 0.01% (w/v) SDS can be detrimental to chromatography separation and suppress the ESI.[24] Highly efficient depletion of SDS before MS analysis is critical. Multiple methods have been evaluated for SDS removal for bottom-up proteomics and/or dTDP, including membrane ultrafiltration,[25] chloroform-methanol precipitation (CMP),[26] and single-spot solid-phase sample preparation using magnetic beads (SP3).[27,28]

Membrane ultrafiltration (MU) has been widely used by the bottom-up proteomics community for the filter-aided sample preparation (FASP) method to remove SDS before enzymatic digestion of proteins.[25] Basically, a protein sample in 1-5% (w/v) SDS solution is loaded onto a commercialized membrane filter unit with a 10-30-kDa molecular weight cut off (MWCO), followed by washing with a 8 M urea solution to remove SDS, which is based on the fact that 8 M urea can destroy the hydrophobic interaction between SDS and proteins. The MU has also been routinely deployed for buffer exchange for TDP sample preparation.[22] CMP is a well-recognized method for removing SDS from proteins in the dTDP workflow, and the Kelleher group has utilized the CMP for cleaning the protein samples after GELFrEE fractionation in their large-scale dTDP works.[5,6,8] Briefly, a protein sample dissolved in a SDS solution is mixed with methanol, chloroform, and water. After centrifugation, three phases form and the proteins precipitate at the interphase. After removing the upper phase, more methanol is added and the purified protein pellet is obtained after centrifugation. SP3 has been suggested as an efficient sample preparation method for bottom-up proteomics and various detergents can be removed from proteins using the SP3 method.[27,28] Recently, the Webb group evaluated the SP3 method for preparing intact protein samples for dTDP, demonstrating the great potential of the SP3 method as a universal sample preparation method for both bottom-up proteomics and dTDP.[29] For SP3, a protein sample in a SDS buffer was mixed with magnetic beads and acetonitrile (ACN). Under a high concentration of ACN, proteins are adsorbed on the beads. Then the beads are washed with organic solvents (i.e., ethanol and ACN) to clean

up the proteins, followed by on-bead digestion for bottom-up proteomics[27,28] or recovering proteins from beads with cold 80% (v/v) formic acid for dTDP.[29]

In this work, for the first time, we compared the MU with a 30-kDa MWCO membrane, CMP, and SP3 methods for cleaning up proteins extracted from *E. coli* cells using 1% (w/v) SDS for dTDP. The MU method showed the best results regarding the protein recovery and compatibility with the follow-up MS analysis. We further tested the MU method for human cells (HepG2). We analyzed the prepared *E. coli* and HepG2 samples using our CZE-MS/MS system. Our data demonstrated that coupling the SDS-based protein extraction with the MU-based sample cleanup could be a universal sample preparation method for dTDP with high protein recovery, no significant protein bias, good reproducibility, and great compatibility with follow-up MS analysis.

Experimental section

Details of materials and reagents are listed in **Supporting Information I**.

Protein Extraction from *Escherichia coli* and HepG2 cells

Escherichia coli (*E. coli*, strain K-12 substrain MG1655) was cultured in the LB (Luria-Bertani) medium at 37 °C until OD600 reached 0.7. The *E. coli* cells were harvested by centrifugation at 4 000 rpm for 10 min. The cell pellet was washed with PBS (phosphate buffered saline) buffer for three times to remove the leftover culture medium. After that, 400 µL of a lysis buffer containing 1% (w/v) SDS, 100 mM NH_4HCO_3, protease inhibitors, and phosphatase inhibitors (pH 8.0) was added into the Eppendorf tube containing the *E. coli* cells. The cells were pipetted up and down a couple of times and lysed by ultrasonication (Branson Sonifier 250, VWR Scientific, Batavia, IL) on ice for 10 min. After cell lysis, the cell lysates were then centrifuged at 14 000 g for 5 min. After that, the protein concentration of the supernatant was measured with the BCA (Bicinchoninic acid) assay. The supernatant was then aliquoted into 100 µg/tube (4 mg/mL protein concentration) and stored at -80 °C before use. The cultured HepG2 cells were kindly provided by Prof. David Lubman at the Department of Surgery Research of University of Michigan. After cell culture, the HepG2 cells were harvested through centrifugation at 100 *g* for 5 min and were washed with the PBS buffer for three times. The cell lysis protocol was the same as the *E. coli* cells described above. After the BCA assay for protein concentration

measurement, the extracted proteins were aliquoted into 100 µg/tube (4 mg/mL protein concentration) and stored at -80 °C before use.

Protein sample cleanup with various methods before MS analysis

SP3 method

The SP3 procedure was performed according to the literature with some modifications.[27,28] 10 µg, 100 µg and 500 µg of the two types of Carboxylate-modified paramagnetic beads were added into 100-µg *E. coli* protein extraction followed by addition of acetonitrile (ACN) ensuring ACN concentration higher than 70% (v/v). *E. coli* protein extraction was incubated in presence of magnetic beads and ACN for 18 min at room temperature and then was placed on a magnet for 2 min. The supernatant was taken out, dried down and the protein concentration was measured through the BCA assay. 200 µL of ethanol was used to rinse the beads twice and 200 µL ACN was used to rinse the beads once. 60 µL of 100 mM NH_4HCO_3 (pH 8) was then added into the beads and sonicated for 10 min. The solution was then placed onto a hotplate at 95 °C for 15 min. The supernatant containing proteins was taken out and the protein concentration was measured with the BCA assay. The SP3 method was also applied on the HepG2 cell lysate with the same procedure.

CMP method

The CMP procedure was processed based on the literature.[26] Briefly, 400 µL methanol, 100 µL chloroform and 300 µL water were added into 100-µg *E. coli* cell lysate (1 µg/µL, 1% (w/v) SDS) successively. Every addition of reagent was followed by a thorough vortex. The mixture was then centrifuged at 14 000 g for 1 min. Solution separated into three layers after centrifugation. The top aqueous layer was carefully removed without disturbing the protein flake. 400 µL of methanol was then added into the solution followed by a thorough vortex. The mixture was then centrifuged at 20 000 g for 5 min. Supernatant was removed. The protein pellet was suspended in a 50-µL buffer containing 100 mM NH_4HCO_3 (pH 8) with or without 1% (w/v) SDS with gentle pipetting. We also vortexed and sonicated the sample solution gently for a short period of time to improve the protein recovery. After centrifugation, the protein solution was analyzed by the BCA assay to determine the protein concentration.

MU method

100 µL of an 8 M urea solution in 100 mM NH_4HCO_3 (pH 8) was first added into 100-µg of *E. coli* cell lysate, producing a protein solution with about 0.80 mg/mL protein concentration. The mixture was then loaded onto a membrane filtration unit (30 kDa MWCO membrane). The filtration unit was centrifuged at 14 000 g to make sure that all the solution went through the membrane. The membrane was then washed with 100 µL of 8 M urea in 100 mM NH_4HCO_3 twice followed by membrane washing with 100-µL 100 mM NH_4HCO_3 for three times. After the washing, 50 µL of 100 mM NH_4HCO_3 was loaded onto the membrane, followed by pipetting up and down a few times. The filtration unit was then vortexed for 5 min and flipped over followed by a quick spin-down to recover the proteins from the membrane. The protein concentration in the collected solution was measured through the BCA assay. The same procedure was utilized for the HepG2 cell lysate.

SDS-PAGE and CZE-MS/MS analysis

The *E. coli* and HepG2 cell lysates before and after cleanup with the three methods were analyzed by SDS-PAGE according to the procedure in the literature.[30]

For CZE-MS/MS, a linear polyacrylamide (LPA)-coated capillary (50/360 µm i.d./o.d.) with one end etched by hydrofluoric acid was used for CZE separation.[31-33] The commercialized electrokinetically pumped sheath flow CE-MS interface (EMASS II, CMP scientific, Brooklyn, NY) was used to couple CZE to MS.[34,35] The automated CZE operations were implemented with an ECE-001 autosampler (CMP scientific). A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for all CZE-MS/MS analyses. A data-dependent acquisition (DDA) method was employed. The details of SDS-PAGE and CZE-MS/MS analysis are described in **Supporting Information I.**

Data analysis

The TopPIC (Top-down mass spectrometry based proteoform identification and characterization) software was applied for proteoform IDs via database search for all *E. coli* and HepG2 data.[19] Briefly, the RAW files were converted into mzML files using the msconvert tool.[36] The mzML files was then processed by the TopFD (Top-down mass spectrometry feature detection) tool for spectral deconvolution. The resulted msalign files was then processed by TopPIC (v1.3.1) for database searching. UniProt databases of *E.*

coli (UP000000625) and Human (UP000005640) were used for search. For the database search, the maximum number of mass shift was 1. All other parameters were kept as default. The target-decoy approach was employed to evaluate the false discovery rate (FDR) of proteoform spectrum match (PrSM) and proteoform IDs.[37,38] The database search results were filtered with a 1% PrSM-level FDR and a 5% proteoform-level FDR. The proteoforms identified from *E. coli* and HepG2 cells are listed in the **Supporting Information II**. The MS raw data have been deposited to the ProteomeXchange Consortium via the PRIDE[39] partner repository with the data set identifier PXD018248.

The Retrieve/ID mapping tool from the UniProt was used for Gene Ontology (GO) analysis. Grand average of hydropathy (GRAVY) values of proteoforms were calculated through a GRAVY Calculator (http://www.gravy-calculator.de/). Positive GRAVY values suggest hydrophobic and negative values indicate hydrophilic. The transmembrane domains (TMDs) of identified membrane proteins were predicted using the TMHMM software (http://www.cbs.dtu.dk/services/TMHMM/).

Results and discussion

Comparison of MU, CMP and SP3 methods for cleanup of cell lysates containing SDS before MS

SDS has been widely used in proteomic studies to facilitate protein extraction from cells and protein solubilization. However, trace amount of SDS could be detrimental to downstream processes such as enzymatic digestion in bottom up proteomics, chromatographic separation, and MS detection.[24,40] It is vital to remove SDS from cell lysates before top-down MS analysis. MU, CMP, and SP3 methods have been used in dTDP for removing detergents (*e.g.,* SDS) from proteins.[5,6,8,22,29] Here, for the first time, we compared the MU, CMP, and SP3 methods for preparation of *E. coli* and human (HepG2) cell lysates containing 1% (w/v) SDS for dTDP regarding protein recovery and protein bias. For each method, 100 µg of proteins dissolved in 1% (w/v) SDS were used as the starting material. The BCA assay and SDS-PAGE were used to evaluate the performance of the three methods. To make the sample preparation method compatible with follow-up dynamic pH junction-based CZE-MS/MS analysis,[41] 100 mM NH_4HCO_3 (pH 8.0) was used to redissolve the proteins after removing SDS with the three methods.

For the SP3 method, we first tested the loading capacity of magnetic beads by incubating 100 µg of *E. coli* proteins with three different amounts of magnetic beads, 10 µg, 100 µg and 500 µg. The protein recovery based on the BCA assay was about 60% and had no obvious difference among the three different bead amounts, **Figure 1A**. We also measured the amount of proteins that were not bound to the magnetic beads at the first step with the BCA assay, **Figure 1B**. The unbound protein amount was about 5 µg, indicating that the magnetic beads captured proteins with high efficiency. Considering the recovered proteins (~60 µg) and unbound proteins (~ 5 µg), we noted that about 35% of the loaded proteins were lost somewhere during the SP3 process. We speculated that those proteins were still adsorbed on the magnetic beads and were not eluted by the 100 mM NH_4HCO_3 (pH 8.0) buffer. We further analyzed the proteins prepared by the SP3 method with the three different bead amounts using SDS-PAGE, **Figure S1** in **Supporting Information I**. The three *E. coli* protein samples after the SP3 cleanup show no significant difference regarding the molecular weight (MW) distributions. The results indicate that 10 µg of magnetic beads are good enough to prepare 100-µg proteins from a complex proteome, which agrees well with the data in the literature.[27,28] We utilized 10-µg beads for all the following SP3 experiments. We also noted that the SP3 method-based sample cleanup introduced an obvious bias against large proteins (higher than 50 kDa) compared to the sample before cleanup, **Figure S1**. The bias was also observed in the HepG2 human cell lysate processed by the SP3 method, **Figure S2**. We used 100 mM NH_4HCO_3 (pH 8.0) to extract the proteins from the beads in order to make the method compatible with follow-up CZE-MS/MS analysis, which might lead to relatively low efficiency of redissolving large proteins, because it has been suggested that a buffer containing detergents is essential for completely extracting proteins bound to beads in SP3.[27-29]

We then employed the MU, CMP, and SP3 methods for preparing aliquots of the *E. coli* cell lysate dissolved in 1% (w/v) SDS. Each aliquot contained 100-µg proteins, and four aliquots were prepared by each method. The MU and SP3 methods generated much higher protein recovery than the CMP method (~60% *vs.* 5%) with good reproducibility (RSD<12%) when a solution containing 100 mM NH_4HCO_3 (pH 8.0) was used to redissolve the protein pellet from CMP, **Figure 1C**. We noted that the protein pellet from

CMP was hard to be dissolved in the NH_4HCO_3 buffer, which resulted in a low protein recovery. We further tried to use a buffer containing 1% (w/v) SDS and 100 mM NH_4HCO_3 (pH 8.0) to redissolve the protein pellet and obtained a 50% protein recovery with high precision (RSD, 4%). We then analyzed the *E. coli* cell lysates before and after cleanup using the three methods by SDS-PAGE, **Figure 1D**. For the CMP method, we used the protein sample redissolved in the 1% (w/v) SDS solution for SDS-PAGE. Two batches of prepared samples with the three methods were analyzed. The MU and CMP method show comparable protein MW distributions, which are similar to the original *E. coli* sample without cleanup. As we discussed before, the SP3 method had trouble recovering large proteins with the 100 mM NH_4HCO_3 (pH 8.0) buffer. All the three methods show good reproducibility regarding the SDS-PAGE data. Based on the discussed protein recovery, protein bias, and compatibility with the CZE-MS/MS analysis of the three sample cleanup methods, the MU method outperformed the CMP and SP3 methods. We further employed the MU method for preparation of the HepG2 cell lysate in 1% (w/v) SDS. The SDS-PAGE and BCA assay data clearly show that the MU method can achieve reproducible preparation of the human cell lysate with high protein recovery and precision (86±5%), **Figure 1E** and **1F**. All the results demonstrate that the MU method could be a universal method for sample preparation in dTDP of complex proteomes. We obtained a higher protein recovery for the human cell lysate than the *E. coli* cell lysate (86% *vs.* 60%) using the MU method, presumably due to the fact that *E. coli* proteins tend to be smaller than human proteins in the length range of 1-250 amino acids based on the data in Swiss-Prot database, **Figure S3**, resulting in a higher chance for protein flow-through the membrane (30-kDa MWCO) for the *E. coli* sample.

We also noted that for the MU method, when the centrifugal force is too high (*i.e.,* 16 800 *g*), the protein recovery can be reduced drastically compared to the typical centrifugal force (14 000 *g*) used in the procedure (33% *vs.* 86%), possibly due to membrane clogging by proteins or impurities in the extraction solution. We suggest a pre-centrifugation operation for protein samples to remove any precipitate before the MU procedure, which will ensure the straightforward MU operations and good protein recovery.

Coupling SDS-based protein extraction and MU-based sample cleanup to CZE-MS/MS for dTDP

We further coupled the SDS-based protein extraction and the MU-based protein sample cleanup to our dynamic pH junction-based CZE-MS/MS for dTDP of *E. coli* cells. About 500 nL of the *E. coli* protein solution in 100 mM NH_4HCO_3 (pH 8.0) after cleanup was injected into the CZE capillary for analysis. The injected protein amount was roughly 400 ng. The BGE of CZE was 20% (v/v) acetic acid. We performed CZE-MS/MS analysis of two batches of the *E. coli* sample prepared by the MU method. **Figure 2A** shows the base peak electropherograms of the two *E. coli* samples and **Figure 2B** shows the numbers of identified proteins, proteoforms, and PrSMs. The whole workflow shows good reproducibility regarding the CZE separation profile, base peak intensity (**Table S1**), and identifications. Single-shot CZE-MS/MS identified 832±65 proteoforms (n=2) with a 5% proteoform-level FDR. When we used a 1% proteoform-level FDR, 821± 67 (n = 2) proteoforms corresponding to 219± 21 proteins were identified in a single CZE-MS/MS run, **Figure 2B**. On average, about 20 fragment ions were matched to each identified proteoform, **Figure 2C**, suggesting the high confidence of the proteoform identifications. We noted that mass of identified proteoforms ranged from 1 kDa to 25 kDa and over 70% of the identified proteoforms had mass smaller than 6 kDa. We also analyzed the GO information of the identified proteins, **Figure 2D**, and about 30% of the proteins were membrane proteins. We finally analyzed the hydrophobicity of the identified proteoforms and compared it with our previous work, in which 8M urea was used for protein extraction from *E. coli* cells.[41] As shown in **Figure 2E**, the *E. coli* proteoforms identified in this work show higher hydrophobicity than the ones identified in our previous work, most likely due to the fact that SDS has stronger solubility for hydrophobic proteins than 8M urea. We also noted that compared to the protein samples extracted with 8M urea,[41] the samples from the 1% (w/v) SDS extraction required a higher acetic acid concentration in the BGE of CZE (20% *vs.* 5% (v/v) acetic acid) to achieve reproducible CZE separations, which might be due to the higher hydrophobicity of proteoforms from the 1% (w/v) SDS extraction. We need to point out that when high concentration of acetic acid (*i.e.,* 20%) is used as the BGE for CZE separation, the sample dissolved in the NH_4HCO_3 buffer in a sample vial could be acidified by the BGE during the sample injection process, which will

influence the dynamic pH junction sample stacking obviously. When the sample volume is small (*i.e.,* <5 µL), the issue becomes severe. Immersing the sample injection end of the capillary in a 100 mM NH_4HCO_3 buffer for seconds before moving it into the sample vial for sample injection can eliminate the issue based on our experience.

We also analyzed the HepG2 cell proteins prepared by the MU method using our dynamic pH junction-based CZE-MS/MS. The same CZE and MS conditions as the *E. coli* samples were used here except that we employed 40% (v/v) acetic acid as the BGE of CZE due to much higher complexity of the human cell line sample compared to the *E. coli* sample. The CZE-MS/MS identified 534 proteoforms and 248 proteins in a single run with a 5% proteoform-level FDR. When a 1% proteoform-level FDR was used, 516 proteoforms corresponding to 241 proteins were identified. **Figure 3A** shows the base peak electropherogram of the CZE-MS/MS run. The mass of identified proteoforms ranged from about 1 kDa to roughly 24 kDa, **Figure 3B**. Over 200 proteoforms had mass higher than 10 kDa. Out of the 248 identified proteins, 125 proteins are membrane proteins, 112 proteins are located in nucleus, and 22 proteins belong to chromatin according to the information from the UniProt Knowledgebase (https://www.uniprot.org/). Sequences and fragmentation patterns of two transmembrane proteins (6.8 kDa mitochondrial proteolipid and Cytochrome c oxidase subunit 6A1, mitochondrial) are shown in **Figures 3C** and **3D**. The two membrane proteins were identified with high confidence and the TMDs were cleaved reasonably well in gas phase by HCD. **Figures 3E** and **3F** show the mass spectrum and fragmentation pattern of one proteoform of C1QBP (Complement component 1 Q subcomponent-binding protein, mitochondrial) having a mass of 23767.7 Da. The proteoform had clear signal in the mass spectrum and was identified by MS/MS through the database search with 18 matched fragment ions and a 2.53E-11 E-value. An N-terminal truncation was determined for the proteoform.

The CZE-MS/MS data further indicate that the sample preparation procedure (SDS-based protein extraction and MU-based sample cleanup) is efficient for extraction and preparation of proteins including membrane proteins from bacterial and human cells. The sample preparation procedure should be also compatible with widely used RPLC-MS/MS, although we only used CZE-MS/MS in this work.

Proteoforms with post-translational modifications (PTMs)

We also performed another CZE-MS/MS run of the prepared *E. coli* sample from the MU method under very clean CZE and MS conditions to pursue a higher number of proteoform identifications, leading to an identification of 1,336 proteoforms corresponding to 301 proteins with a 1% proteoform-level FDR. Various protein modifications were detected, including but not limited to N-terminal methionine removal, N-terminal truncation, N-terminal acetylation, and disulfide bond, **Figure 4A**. Two truncated proteoforms of 50S ribosomal protein L7/L12 at the N-terminus with or without lysine methylation are shown in **Figures 4B** and **4C**. The fragmentation patterns show extensive backbone cleavages of the two proteoforms. We also observed that the abundance of the methylated proteoform was about 50% of the non-methylated proteoform according to the mass spectrum in **Figure 4D**. The methylation at Lys-82 detected in our work agrees well with the data in the literature.[42] We identified 15 proteoforms with one or two disulfide bonds and those proteoforms are listed in **Supporting Information II**. Sequences and fragmentation patterns of two proteoforms with one and two disulfide bonds are shown in **Figures 4E** and **4F**. Interestingly, for **Figure 4E**, the location of the disulfide bond was previously reported as zinc ion binding position.[43] For the 50S ribosomal protein L31, the literature data suggested that the C16 was responsible for zinc ion binding, but our data show that the C16, C18, C37 and C40 form two disulfide bonds, **Figure 4F**. The disulfide bonds might form endogenously or develop after cell lysis due to the loss of zinc ions during sample preparation.

We identified proteoforms with various PTMs in the HepG2 data, including but not limited to N-terminal acetylation (205), phosphorylation (11), and disulfide bonds (8), **Figure 5A**. The proteoforms with phosphorylation and disulfide bond are listed in **Supporting Information II**. We identified one proteoform of programmed cell death protein 5 with N-terminal acetylation and one serine phosphorylation (**Figure 5B**), one proteoform of 60S acidic ribosomal protein P2 with two serine phosphorylations (**Figure 5C**), and one proteoform of small ubiquitin-related modifier 1 with both acetylation and phosphorylation at the N-terminal serine residue (**Figure 5D**). The PTM information of these three proteoforms match well with the UniProt Knowledgebase (https://www.uniprot.org/). We noted that the three serine residues marked in red in the underlined region in **Figure 5C** could be phosphorylated according to the UniProt Knowledgebase, and our data show

that only two of them are actually phosphorylated in the proteoform. We also identified one proteoform of 60S ribosomal protein L32 with one disulfide bond, **Figure S4**, which is not reported in the literature according to the UniProt Knowledgebase. Prothymosin alpha (PTMA) is a histone binding protein and it can regulate gene transcription.[44] Prothymosin alpha has eight phosphorylation sites according to the UniProt Knowledgebase. Our data revealed one phosphorylation site (mass shift 79.97 Da) in the underlined region (S85 or T87) in **Figure S5A**, which is not reported previously. We also compared the relative abundance of the identified phosphorylated proteoform of PTMA and the corresponding unphosphorylated proteoform based on the extracted base peak electropherogram, **Figure S5B**. The unphosphorylated proteoform had about 5-times higher abundance than the phosphorylated one. Additionally, CZE separated the phosphorylated and unphosphorylated proteoforms very well with an 8-min difference in migration time and the phosphorylated one migrated obviously slower than the unphosphorylated one in CZE due to the charge reduction from the phosphorylation, which agrees well with the previous reports.[45-48] The migration time shift between unphosphorylated and phosphorylated proteoforms provides additional evidence for the phosphorylation PTM. **Figures S5C** and **S5D** show mass spectra of the unphosphorylated and phosphorylated proteoforms, indicating a difference between them regarding charge distribution. We speculate that the phosphorylation could influence the ESI of prothymosin alpha to some extent.

Histone PTMs are extremely important for regulating gene expression and dTDP is an invaluable approach for delineating the histone code in a proteoform specific manner.[49-52] In this work, we identified 94 histone proteoforms from the HepG2 sample in a single CZE-MS/MS run without any histone purification. The histone proteoforms are listed in **Supporting Information II**. The 94 histone proteoforms covered the five major histone variants, H1 (11), H2A (39), H2B (36), H3 (1) and H4 (7), **Figure 6A**. We observed various PTMs on the histone proteoforms, including acetylation, methylation, and phosphorylation, **Figures 6B-F**. Sequences and fragmentation patterns of two histone H4 proteoforms are shown in **Figure 6B** and **C**. We observed both N-terminal acetylation and a 28-Da mass shift most likely corresponding to two methylations within the underlined region in the two proteoforms. Due to the limited backbone cleavage

coverages for the two proteoforms, it is difficult to localize the methylation PTM. Interestingly, there are no literature reports about methylation or di-methylation PTM in the two regions of histone H4 underlined in **Figures 6B** and **C** according to the UniProt Knowledgebase. We also identified one histone H4 proteoform with a 337-Da mass shift, **Figure 6D**. The mass shift corresponds to a region with four lysine residues (K6, K9, K13 and K17). According to the UniProt Knowledgebase, these four lysine residues could have acetylation (+42 Da), propionylation (+56 Da), crotonylation (+68 Da), butyrylation (+70 Da), succinylation (+100 Da), and glutarylation (+114 Da). We speculate that the 337-Da mass shift is most likely produced by a combination of these various PTMs. The data further suggest the importance of improving the backbone cleavage coverage for comprehensive characterization of proteoforms.

We identified one proteoform of Histone H2A type 1-J with a 122-Da mass shift in the underlined region, **Figure 6E**. We speculate that the mass shift corresponds to an acetylation (+42 Da) and a phosphorylation (+80 Da). It has been reported that the K6 and K10 residues could be acetylated.[53] However, no literature information about the phosphorylation at T17, S19 or S20 in the mass shift corresponding region according to the UniProt Knowledgebase. We also identified one Histone H2A type 1 proteoform with an 83-Da mass shift in the underlined region, **Figure 6F**. The K96 and K100 in the mass shift corresponding region could be acetylated based on the previous reports [53,54] and the information from PhosphoSitePlus® v6.5.8 (https://www.phosphosite.org/). Two lysine acetylation modifications produce an 84-Da mass shift, which is 1-Da heavier than the observed mass shift. The 1-Da difference could be due to a misassignment of the monoisotopic peak of the protein, which resulted in a 1-Da error of the proteoform's monoisotopic mass. Therefore, the observed 83-Da mass shift is most likely due to the acetylation at both K96 and K100.

Conclusions

We performed comprehensive comparisons of the MU, CMP, and SP3 methods for cleanup of proteome samples in a lysis buffer containing SDS regarding protein recovery, protein bias, and compatibility with follow-up MS analysis. Our data indicate that the SDS-based protein extraction and the MU-based protein cleanup could be a universal sample preparation procedure for dTDP of complex proteome samples. The procedure produced

reproducible sample preparation with high protein recovery for both *E. coli* and human cell line samples. Single-shot CZE-MS/MS analysis of the prepared *E. coli* and HepG2 cell proteome samples (400-ng proteins consumed) identified up to 1 336 proteoforms (301 proteins) and 516 proteoforms (241 proteins) with a 1% proteome-level FDR, respectively. Single-shot CZE-MS/MS analysis of the HepG2 cell sample identified 125 membrane proteins and 94 histone proteoforms. The sample preparation procedure including the SDS-based protein extraction and the MU-based protein cleanup should be also compatible with the widely used RPLC-MS/MS approach, although we only used CZE-MS/MS in this work.

We need to point out that when the sample complexity and protein hydrophobicity increase, the BGE composition of CZE needs to be adjusted to ensure good solubility of proteins during CZE separation. We are working on optimizations of CZE-MS conditions for characterization of proteome samples with high hydrophobicity.

Acknowledgements

We thank Prof. Heedeok Hong's group at the Department of Chemistry of Michigan State University and Prof. David M. Lubman's group at the University of Michigan for kindly providing the *E. coli* cells and HepG2 human cells for this project. We thank the support from the National Science Foundation (CAREER Award, Grant DBI1846913) and the National Institutes of Health (Grant R01GM125991).

Supporting Information

The following supporting information is available free of charge at ACS website

Supporting Information I.pdf:

Suppl Experimental Section

Table S1. Intensity and migration time information of 8 high intense peaks from the CZE-MS/MS analyses of two batches of *E. coli* samples prepared with the MU method.

Figure S1. SDS-PAGE data of *E. coli* proteins processed by the SP3 method with three different amounts of beads.

Figure S2. SDS-PAGE data of HepG2 proteins processed by the SP3 method.

Figure S3. Cumulative distribution of the length of *E. coli* proteins and human proteins.

Figure S4. Sequence and fragmentation pattern of one proteoform of 60S ribosomal protein L32 with one disulfide bond.

Figure S5. Protein phosphorylation data of prothymosin alpha.

Suppl References

Supporting Information II.xlsx: Information of the identified proteoforms from the *E. coli* and HepG2 samples prepared with the MU method.

References

1. Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P., Proteoform: a single term describing protein complexity. *Nature methods* **2013,** *10* (3), 186-7.
2. Toby, T. K.; Fornelli, L.; Kelleher, N. L., Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annual review of analytical chemistry* **2016,** *9* (1), 499-519.
3. Smith, L. M.; Kelleher, N. L., Proteoforms as the next proteomics currency. *Science* **2018,** *359* (6380), 1106-1107.
4. Ntai, I.; Fornelli, L.; DeHart, C. J.; Hutton, J. E.; Doubleday, P. F.; LeDuc, R. D.; van Nispen, A. J.; Fellers, R. T.; Whiteley, G.; Boja, E. S.; Rodriguez, H.; Kelleher, N. L., Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk. *Proceedings of the National Academy of Sciences of the United States of America* **2018,** *115* (16), 4140-4145.
5. Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** *480* (7376), 254-8.
6. Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L., Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Molecular & cellular proteomics : MCP* **2013,** *12* (12), 3465-73
7. Ljiljana, Paša-Tolić.; Pamela, K. Jensen.; Gordon, A. Anderson.; Mary, S. Lipton.; Kim, K. Peden.; Suzana, Martinović.; Nikola, Tolić.; James, E. Bruce.; Richard, D. Smith., High Throughput Proteome-Wide Precision Measurements of Protein Expression Using Mass Spectrometry. *J. Am. Chem. Soc.* **1999**, *121*(34), 7949-7950

8. Anderson, L. C.; DeHart, C. J.; Kaiser, N. K.; Fellers, R. T.; Smith, D. F.; Greer, J. B.; LeDuc, R. D.; Blakney, G. T.; Thomas, P. M.; Kelleher, N. L.; Hendrickson, C. L., Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. *Journal of proteome research* **2017,** *16* (2), 1087-1096.

9. Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y., Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. *Analytical chemistry* **2017,** *89* (10), 5467-5475.

10. Liang, Y.; Jin, Y.; Wu, Z.; Tucholski, T.; Brown, K. A.; Zhang, L.; Zhang, Y.; Ge, Y., Bridged Hybrid Monolithic Column Coupled to High-Resolution Mass Spectrometry for Top-Down Proteomics. *Analytical chemistry* **2019,** *91* (3), 1743-1747.

11. McCool, E. N.; Lubeckyj, R. A.; Shen, X.; Chen, D.; Kou, Q.; Liu, X.; Sun, L., Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia coli Proteome. *Analytical chemistry* **2018,** *90* (9), 5529-5533.

12. Yu, D.; Wang, Z.; Cupp-Sutton, K. A.; Liu, X.; Wu, S., Deep Intact Proteoform Characterization in Human Cell Lysate Using High-pH and Low-pH Reversed-Phase Liquid Chromatography. *Journal of the American Society for Mass Spectrometry* **2019,** *30* (12), 2502-2513.

13. Ansong, C.; Wu, S.; Meng, D.; Liu, X.; Brewer, H. M.; Deatherage Kaiser, B. L.; Nakayasu, E. S.; Cort, J. R.; Pevzner, P.; Smith, R. D.; Heffron, F.; Adkins, J. N.; Pasa-Tolic, L., Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. *Proceedings of the National Academy of Sciences of the United States of America* **2013,** *110* (25), 10153-8.

14. Lubeckyj, R. A.; Basharat, A. R.; Shen, X.; Liu, X.; Sun, L., Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples. *Journal of the American Society for Mass Spectrometry* **2019,** *30* (8), 1435-1445.

15. Liu, Z.; Wang, R.; Liu, J.; Sun, R.; Wang, F., Global Quantification of Intact Proteins via Chemical Isotope Labeling and Mass Spectrometry. *Journal of proteome research* **2019,** *18* (5), 2185-2194.

16. Riley, N. M.; Westphall, M. S.; Coon, J. J., Activated Ion-Electron Transfer Dissociation Enables Comprehensive Top-Down Protein Fragmentation. *Journal of proteome research* **2017,** *16* (7), 2653-2659.

17. Shaw, J. B.; Li, W.; Holden, D. D.; Zhang, Y.; Griep-Raming, J.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S., Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. *Journal of the American Chemical Society* **2013,** *135* (34), 12646-51.

18. Shaw, J. B.; Malhan, N.; Vasil'ev, Y. V.; Lopez, N. I.; Makarov, A.; Beckman, J. S.; Voinov, V. G., Sequencing Grade Tandem Mass Spectrometry for Top-Down Proteomics Using Hybrid Electron Capture Dissociation Methods in a Benchtop Orbitrap Mass Spectrometer. *Analytical chemistry* **2018,** *90* (18), 10819-10827.

19. Kou, Q.; Xun, L.; Liu, X., TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* **2016,** *32* (22), 3495-3497.

20. Cai, W.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X.; Ge, Y., MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. *Molecular & cellular proteomics : MCP* **2016,** *15* (2), 703-14.

21. Sun, R. X.; Luo, L.; Wu, L.; Wang, R. M.; Zeng, W. F.; Chi, H.; Liu, C.; He, S. M., pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Analytical chemistry* **2016,** *88* (6), 3082-90.

22. Donnelly, D. P.; Rawlins, C. M.; DeHart, C. J.; Fornelli, L.; Schachner, L. F.; Lin, Z.; Lippens, J. L.; Aluri, K. C.; Sarin, R.; Chen, B.; Lantz, C.; Jung, W.; Johnson, K. R.; Koller, A.; Wolff, J. J.; Campuzano, I. D. G.; Auclair, J. R.; Ivanov, A. R.; Whitelegge, J. P.; Pasa-Tolic, L.; Chamot-Rooke, J.; Danis, P. O.; Smith, L. M.; Tsybin, Y. O.; Loo, J. A.; Ge, Y.; Kelleher, N. L.; Agar, J. N., Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nature methods* **2019,** *16* (7), 587-594.

23. Speers, A. E.; Wu, C. C., Proteomics of integral membrane proteins--theory and application. *Chemical reviews* **2007,** *107* (8), 3687-714.

24. Botelho, D.; Wall, M. J.; Vieira, D. B.; Fitzsimmons, S.; Liu, F.; Doucette, A., Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation. *Journal of proteome research* **2010,** *9* (6), 2863-70.

25. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nature methods* **2009,** *6* (5), 359-62.

26. Wessel, D.; Flugge, U. I., A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Analytical biochemistry* **1984,** *138* (1), 141-3.

27. Hughes, C. S.; Foehr, S.; Garfield, D. A.; Furlong, E. E.; Steinmetz, L. M.; Krijgsveld, J., Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular systems biology* **2014,** *10*, 757.

28. Hughes, C. S.; Moggridge, S.; Muller, T.; Sorensen, P. H.; Morin, G. B.; Krijgsveld, J., Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature protocols* **2019,** *14* (1), 68-85.

29. Dagley, L. F.; Infusini, G.; Larsen, R. H.; Sandow, J. J.; Webb, A. I., Universal Solid-Phase Protein Preparation (USP(3)) for Bottom-up and Top-down Proteomics. *Journal of proteome research* **2019,** *18* (7), 2915-2924.

30. Blancher, C.; Jones, A., SDS -PAGE and Western Blotting Techniques. *Methods in molecular medicine* **2001,** *57*, 145-62.

31. McCool, E. N.; Lubeckyj, R.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., Large-scale Top-down Proteomics Using Capillary Zone Electrophoresis Tandem Mass Spectrometry. *Journal of visualized experiments: JoVE* **2018,** *140*, e58644.

32. Zhu, G.; Sun, L.; Dovichi, N. J., Thermally-initiated free radical polymerization for reproducible production of stable linear polyacrylamide coated capillaries, and their application to proteomic analysis using capillary zone electrophoresis-mass spectrometry. *Talanta* **2016,** *146*, 839-43.

33. Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J., Ultrasensitive and fast bottom-up analysis of femtogram amounts of complex proteome digests. *Angewandte Chemie* **2013,** *52* (51), 13661-4.

34. Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J., Third-generation electrokinetically pumped sheath-flow nanospray interface with improved stability and sensitivity for automated capillary zone electrophoresis-mass spectrometry analysis of complex proteome digests. *Journal of proteome research* **2015,** *14* (5), 2312-21.

35. Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J., Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid communications in mass spectrometry : RCM* **2010,** *24* (17), 2554-60.

36. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008,** *24* (21), 2534-6.

37. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* **2002,** *74* (20), 5383-92.

38. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **2007,** *4* (3), 207-14.

39. Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A., The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research* **2019,** *47* (D1), D442-D450.

40. Arribas, J.; Castano, J. G., Kinetic studies of the differential effect of detergents on the peptidase activities of the multicatalytic proteinase from rat liver. *The Journal of biological chemistry* **1990,** *265* (23), 13969-73.

41. Lubeckyj, R. A.; McCool, E. N.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 Escherichia coli Proteoforms. *Analytical chemistry* **2017,** *89* (22), 12059-12067.

42. Chang, C. N.; Chang, N., Methylation of the ribosomal proteins in Escherichia coli. Nature and stoichiometry of the methylated amino acids in 50S ribosomal proteins. *Biochemistry* **1975,** *14* (3), 468-77.

43. Monaco, H. L.; Crawford, J. L.; Lipscomb, W. N., Three-dimensional structures of aspartate carbamoyltransferase from Escherichia coli and of its complex with cytidine

triphosphate. *Proceedings of the National Academy of Sciences of the United States of America* **1978,** *75* (11), 5276-80.

44. Karetsou, Z.; Kretsovali, A.; Murphy, C.; Tsolas, O.; Papamarcaki, T., Prothymosin alpha interacts with the CREB-binding protein and potentiates transcription. *EMBO reports* **2002,** *3* (4), 361-6.

45. Chen, D.; Lubeckyj, R. A.; Yang, Z.; McCool, E. N.; Shen, X.; Wang, Q.; Xu, T.; Sun, L., Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics. *Analytical chemistry* **2020,** *92* (5), 3503-3507.

46. Wojcik, R.; Vannatta, M.; Dovichi, N. J., Automated enzyme-based diagonal capillary electrophoresis: application to phosphopeptide characterization. *Analytical chemistry* **2010,** *82* (4), 1564-7.

47. Kim, J.; Zand, R.; Lubman, D. M., Electrophoretic mobility for peptides with post-translational modifications in capillary electrophoresis. *Electrophoresis* **2003,** *24* (5), 782-93.

48. Faserl, K.; Sarg, B.; Gruber, P.; Lindner, H. H., Investigating capillary electrophoresis-mass spectrometry for the analysis of common post-translational modifications. *Electrophoresis* **2018,** *39* (9-10), 1208-1215.

49. Zheng, Y.; Huang, X.; Kelleher, N. L., Epiproteomics: quantitative analysis of histone marks and codes by mass spectrometry. *Current opinion in chemical biology* **2016,** *33*, 142-50.

50. Janssen, K. A.; Sidoli, S.; Garcia, B. A., Recent Achievements in Characterizing the Histone Code and Approaches to Integrating Epigenomics and Systems Biology. *Methods in enzymology* **2017,** *586*, 359-378.

51. Wang, T.; Holt, M. V.; Young, N. L., Early butyrate induced acetylation of histone H4 is proteoform specific and linked to methylation state. *Epigenetics* **2018,** *13* (5), 519-535.

52. Gargano, A. F. G.; Shaw, J. B.; Zhou, M.; Wilkins, C. S.; Fillmore, T. L.; Moore, R. J.; Somsen, G. W.; Pasa-Tolic, L., Increasing the Separation Capacity of Intact Histone Proteoforms Chromatography Coupling Online Weak Cation Exchange-HILIC to Reversed Phase LC UVPD-HRMS. *Journal of proteome research* **2018,** *17* (11), 3791-3800.

53. Wu, Q.; Cheng, Z.; Zhu, J.; Xu, W.; Peng, X.; Chen, C.; Li, W.; Wang, F.; Cao, L.; Yi, X.; Wu, Z.; Li, J.; Fan, P., Suberoylanilide hydroxamic acid treatment reveals crosstalks among proteome, ubiquitylome and acetylome in non-small cell lung cancer A549 cell line. *Scientific reports* **2015,** *5*, 9520.

54. Zhao, S.; Xu, W.; Jiang, W.; Yu, W.; Lin, Y.; Zhang, T.; Yao, J.; Zhou, L.; Zeng, Y.; Li, H.; Li, Y.; Shi, J.; An, W.; Hancock, S. M.; He, F.; Qin, L.; Chin, J.; Yang, P.; Chen, X.; Lei, Q.; Xiong, Y.; Guan, K. L., Regulation of cellular metabolism by protein lysine acetylation. *Science* **2010,** *327* (5968), 1000-4.

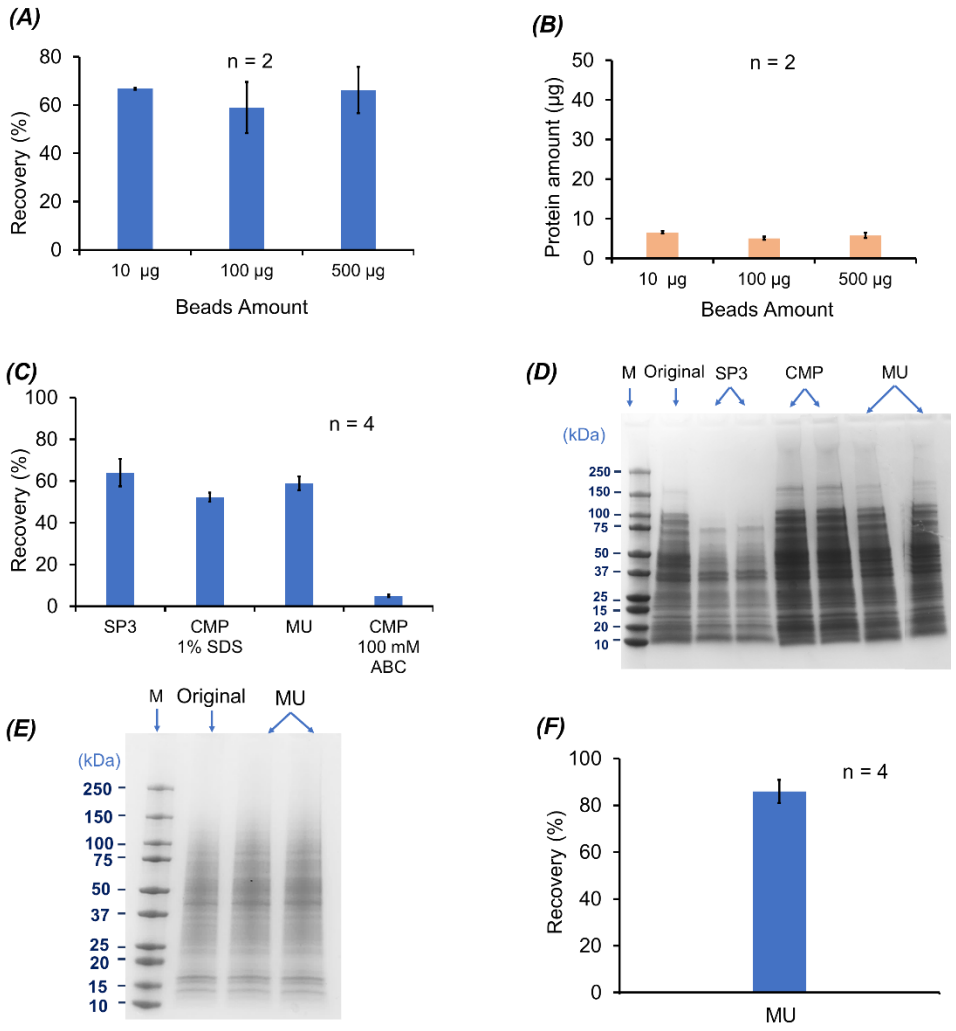


Figure 1. BCA and SDS-PAGE results on the *E. coli* cell proteins (A-D) and HepG2 cell proteins (E and F) when different SDS removal methods were applied. (A) Protein recovery (%) of the SP3 method for 100-µg *E. coli* proteins when different amounts of magnetic beads were used (n=2). (B) Amounts of unbound proteins to magnetic beads as a function of the magnetic bead amount (n=2). (C) Protein recovery (%) of the SP3, CMP and MU methods. The protein pellets from the CMP method were dissolved in 100 mM NH_4HCO_3 (ABC is short for ammonium bicarbonate) (pH 8) with or without 1% (w/v) SDS (n=4). (D) SDS-PAGE data of the recovered *E. coli* proteins using the SP3, CMP and MU methods (n=2) as well as the *E. coli* cell lysate in 1% (w/v) SDS before sample cleanup (Original). For the CMP method, the protein pellet dissolved in 100 mM NH_4HCO_3 (pH 8) with 1% (w/v) SDS was used for the analysis. For each sample, an aliquot of 10-µg proteins was loaded for SDS-PAGE. (E) SDS-PAGE data of the HepG2 cell protein

samples before (Original) and after sample cleanup with the MU method (n=2). For each sample, an aliquot of 6-μg proteins was loaded for SDS-PAGE. (F) Protein recovery data of the HepG2 cell samples after the MU method-based sample cleanup (n=4). The error bars in the figures represent the standard deviations of protein recovery or protein amount.

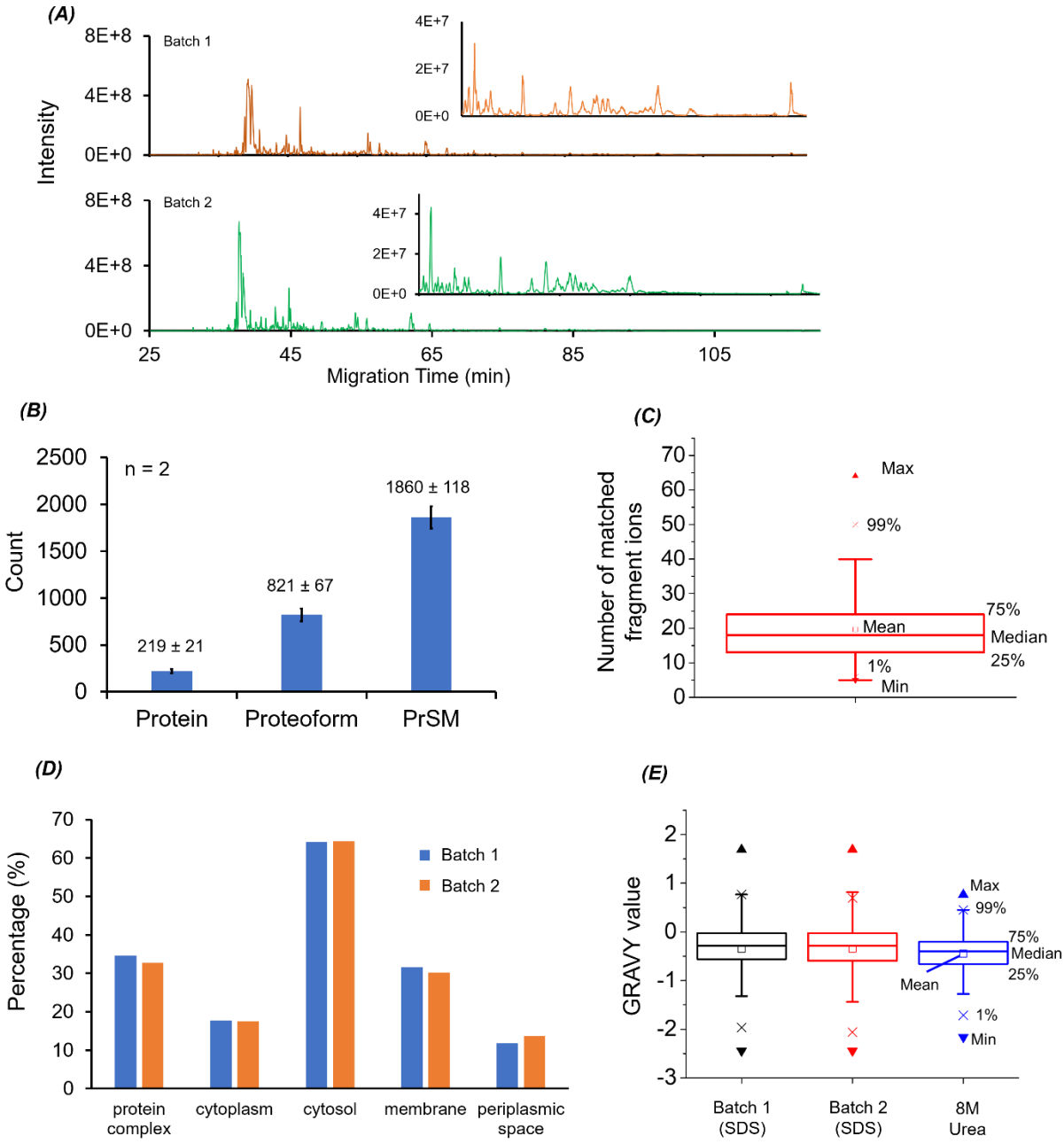


Figure 2. CZE-MS/MS data of *E. coli* samples prepared with the MU method. (A) Base peak electropherograms of two batches of prepared *E. coli* protein samples after CZE-MS/MS analysis. (B) Numbers of protein, proteoform, and PrSM identifications from the two CZE-MS/MS runs. The error bars represent the standard deviations of the number of identifications. (C) Box chart of the number of matched fragment ions of identified *E. coli* proteoforms. (D) Gene Ontology cellular component analysis of identified *E. coli* proteins from the two CZE-MS/MS analyses. (E) Box charts of GRAVY values of the identified

proteoforms from the two CZE-MS/MS analyses in this work (SDS-batch 1 and SDS-batch 2) and from our previous work in reference 41 (8M urea).

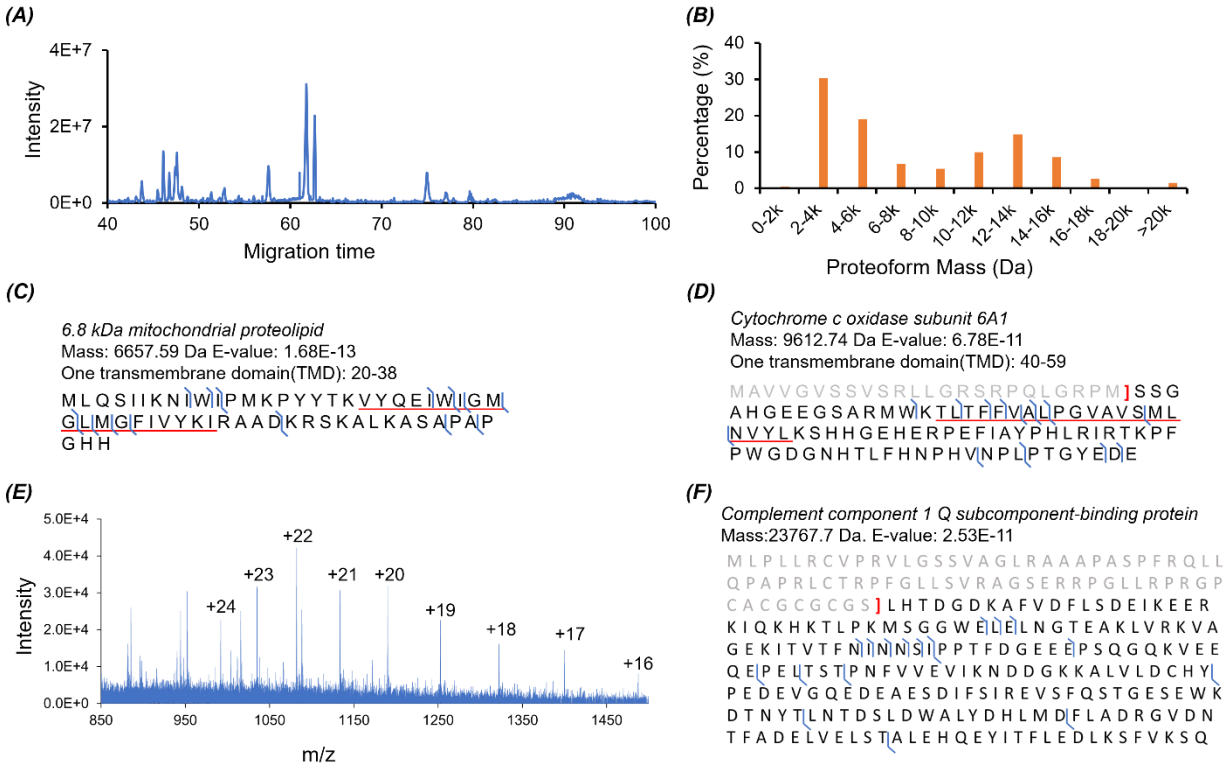


Figure 3. CZE-MS/MS data of the HepG2 cell protein sample prepared with the MU method. (A) Base peak electropherogram of the protein sample after CZE-MS/MS analysis. (B) Mass distribution of the identified proteoforms from the HepG2 protein sample. (C) and (D): Sequences and fragmentation patterns of two transmembrane proteins with one TMD. The regions corresponding to TMDs are underlined. (E) Mass spectrum of the identified proteoform of C1QBP (Complement component 1 Q subcomponent-binding protein, mitochondrial) with a mass of 23767.7Da. (F) Sequence and fragmentation pattern of the C1QBP proteoform in (E).

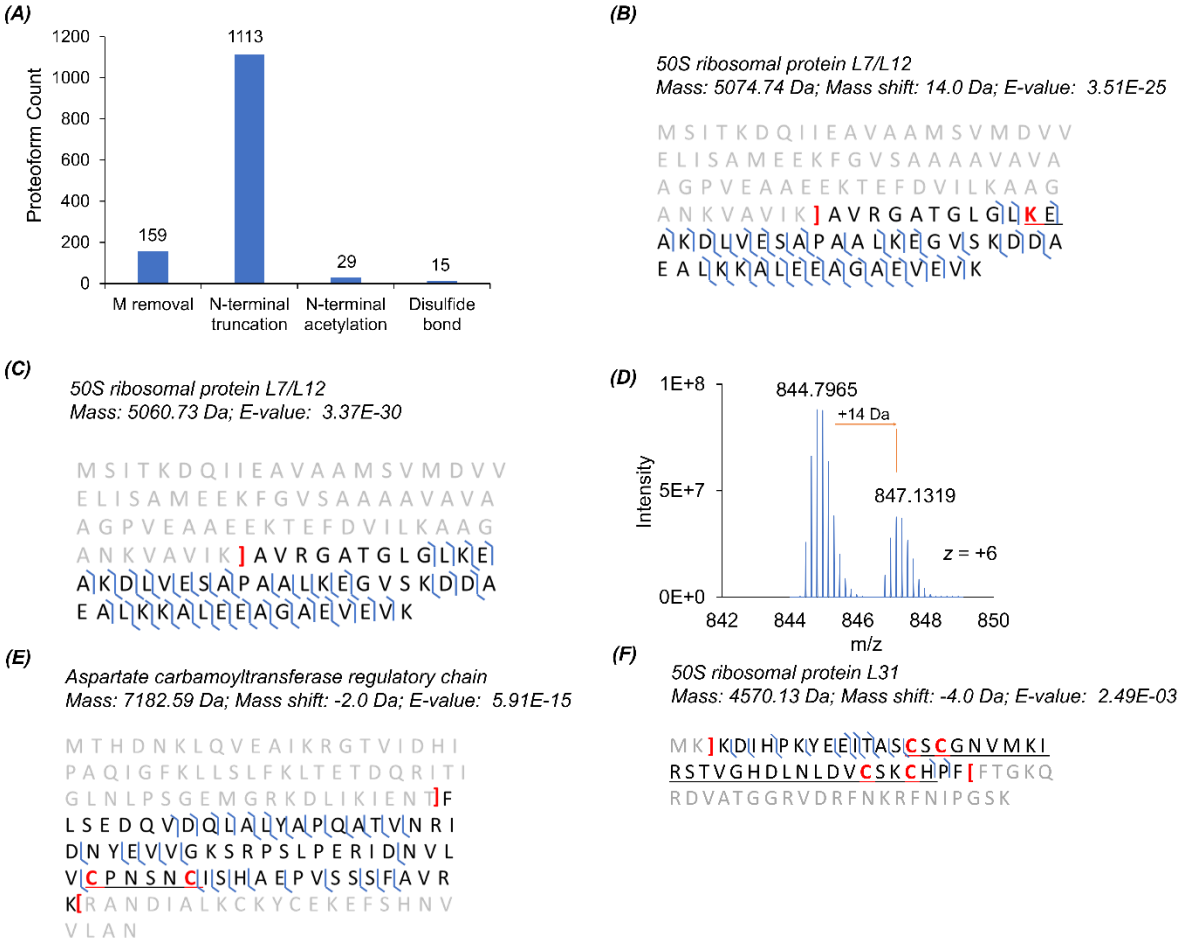


Figure 4. CZE-MS/MS data of the *E. coli* sample regarding PTMs. (A) Distribution of some modifications on the identified proteoforms. (B) Sequence and fragmentation pattern of the 50S ribosomal protein L7/L12 proteoform with one methylation at the marked lysine residue and N-terminal truncation. (C) Sequence and fragmentation pattern of the 50S ribosomal protein L7/L12 proteoform with only N-terminal truncation. (D) Mass spectrum of one charge state (+6) of the 50S ribosomal protein L7/L12 proteoforms in (B) and (C). (E) Sequence and fragmentation pattern of one proteoform of aspartate carbamoyltransferase regulatory chain with one disulfide bond between the two marked cystine residues and truncations at the termini. (F) Sequence and fragmentation pattern of one proteoform of 50S ribosomal protein L31 with two disulfide bonds among the four marked cystine residues, removal of two amino acid residues at the N-terminus, and truncation at the C-terminus.

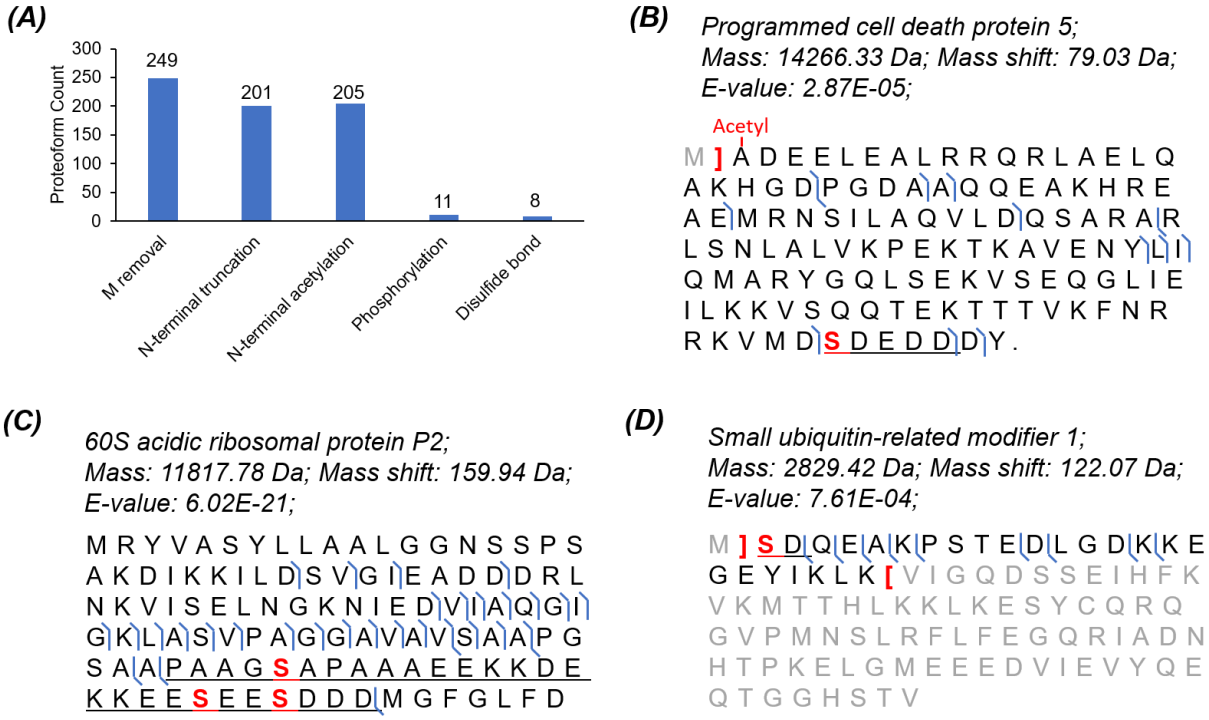


Figure 5. CZE-MS/MS data of the HepG2 sample regarding PTMs. (A) Distribution of some modifications on the identified proteoforms. Sequences and fragmentation patterns of some proteoforms with one phosphorylation site and N-terminal acetylation (B), with two phosphorylation sites (C), and with phosphorylation and acetylation on the N-terminal serine residue (D).

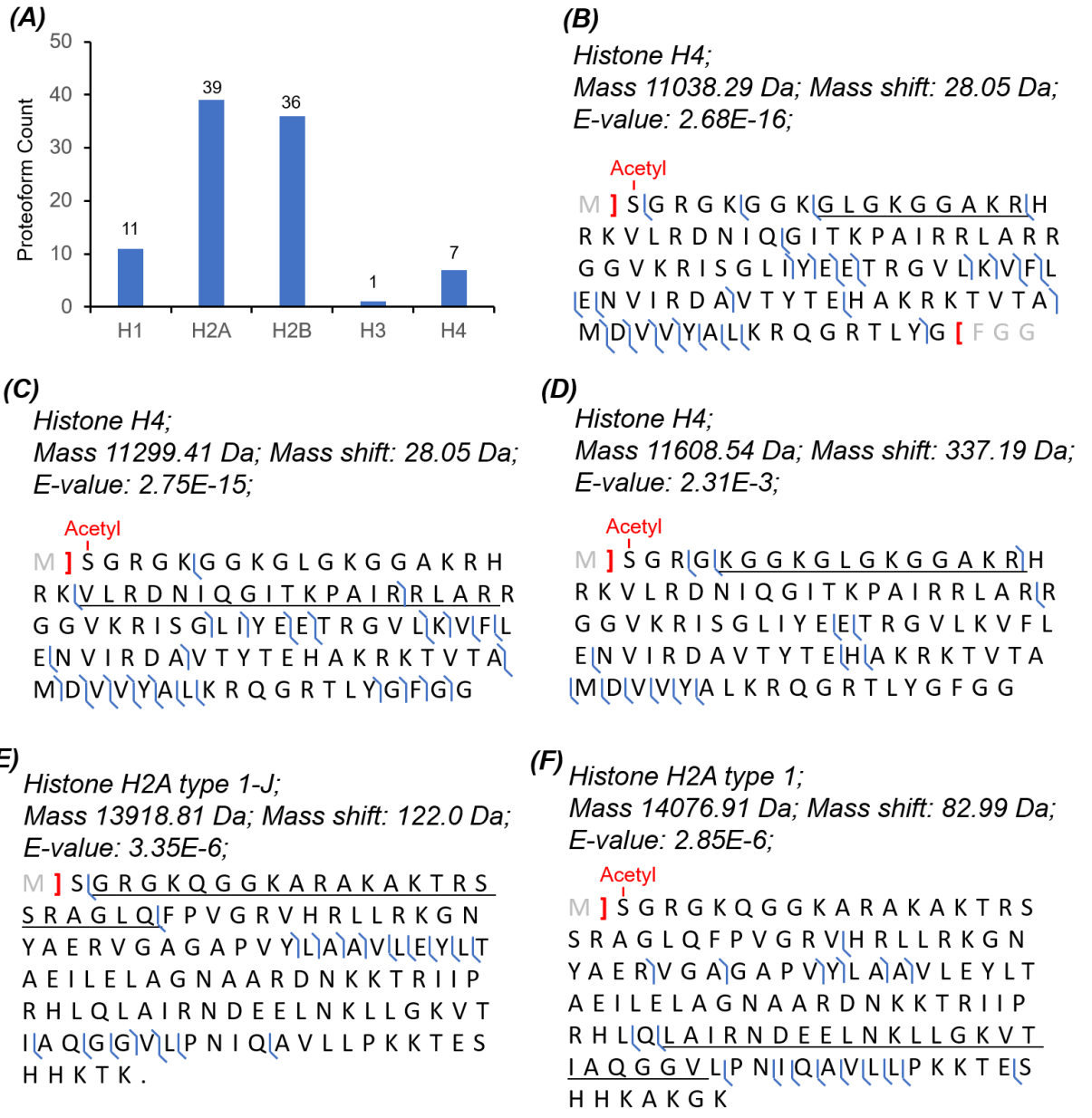
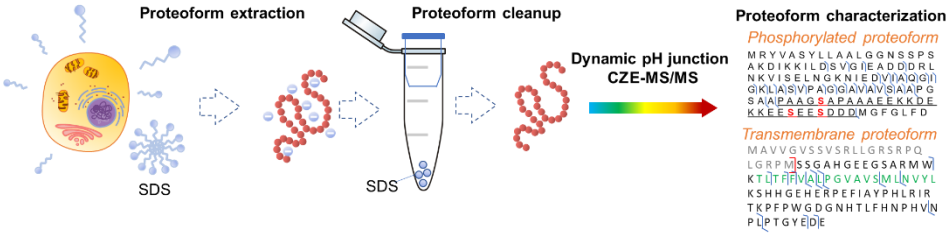


Figure 6. CZE-MS/MS data of the HepG2 sample regarding histone proteoforms. (A) Distribution of the identified histone proteoforms as a function of major histone variants. Sequences and fragmentation patterns of three H4 proteoforms with a 28-Da mass shift (B), a 28-Da mass shift (C), and a 337-Da mass shift (D). Sequences and fragmentation patterns of histone H2A type 1-J proteoform with a 122-Da mass shift (E) and histone H2A type 1 proteoform with an 83-Da mass shift (F).

For TOC:


Proteoform extraction
SDS
Proteoform cleanup
SDS
Dynamic pH junction
CZE-MS/MS
Proteoform characterization
Phosphorylated proteoform
Transmembrane proteoform