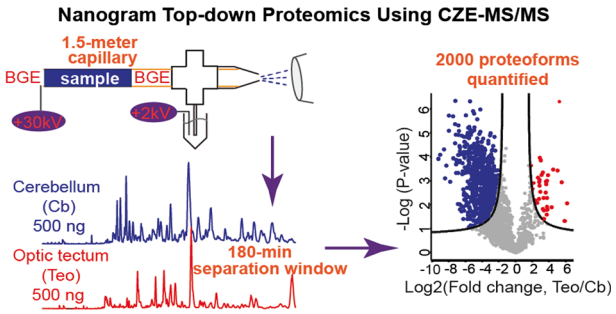
J. Am. Soc. Mass Spectrom. (2019) 30:1435–1445
DOI: 10.1007/s13361-019-02167-w

FOCUS: EMERGING INVESTIGATORS: RESEARCH ARTICLE

Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples

Rachele A. Lubeckyj,[1] Abdul Rehman Basharat,[2] Xiaojing Shen,[1] Xiaowen Liu,[2,3] Liangliang Sun[1]

[1]Department of Chemistry, Michigan State University, 578 S Shaw Ln, East Lansing, MI 48824, USA
[2]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA
[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA



Abstract. Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry (CZE-ESI-MS/MS) has attracted attention recently for top-down proteomics because it can achieve highly efficient separation and very sensitive detection of proteins. However, separation window and sample loading volume of CZE need to be boosted for a better proteome coverage using CZE-MS/MS. Here, we present an improved CZE-MS/MS system that achieved a 180-min separation window and a 2-μL sample loading volume for top-down characterization of protein mixtures. The system obtained highly efficient separation of proteins with nearly one million theoretical plates for myoglobin and enabled baseline separation of three different proteoforms of myoglobin. The CZE-MS/MS system identified 797 ± 21 proteoforms and 258 ± 7 proteins (n = 2) from an *Escherichia coli* (*E. coli*) proteome sample in a single run with only 250 ng of proteins injected. The system still identified 449 ± 40 proteoforms and 173 ± 6 proteins (n = 2) from the *E. coli* sample when only 25 ng of proteins were injected per run. Single-shot CZE-MS/MS analyses of zebrafish brain cerebellum (Cb) and optic tectum (Teo) regions identified 1730 ± 196 proteoforms (n = 3) and 2024 ± 255 proteoforms (n = 3), respectively, with only 500-ng proteins loaded per run. Label-free quantitative top-down proteomics of zebrafish brain Cb and Teo regions revealed significant differences between Cb and Teo regarding the proteoform abundance. Over 700 proteoforms from 131 proteins had significantly higher abundance in Cb compared to Teo, and these proteins were highly enriched in several biological processes, including muscle contraction, glycolytic process, and mesenchyme migration.

Keywords: Capillary zone electrophoresis-tandem mass spectrometry, Top-down proteomics, Mass-limited samples, Proteoform quantification, *Escherichia coli*, Zebrafish cerebellum, Zebrafish optic tectum

Received: 14 January 2019/Revised: 18 February 2019/Accepted: 18 February 2019/Published Online: 9 April 2019

Introduction

Reversed-phase liquid chromatography-electrospray ionization-tandem mass spectrometry (RPLC-ESI-MS/MS) is routinely employed for top-down proteomics. Drastic progress has been achieved for top-down characterization of complex proteomes using RPLC-MS/MS with the production of thousands of proteoform identification and relative

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13361-019-02167-w) contains supplementary material, which is available to authorized users.

Correspondence to: Liangliang Sun; *e-mail:* lsun@chemistry.msu.edu

quantification [1–9]. However, several challenges remain for the RPLC-MS/MS approach. One is the high-capacity separations of complex proteoform mixtures. Another one is the large-scale characterization of proteoforms in mass-limited proteome samples. The state-of-the-art RPLC-MS/MS approach typically requires low micrograms to hundreds of micrograms of protein materials for identification of thousands of proteoforms from complex proteome samples [1–9]. Alternative top-down proteomic platforms that enable high-capacity proteoform separation and highly sensitive proteoform detection are vital.

Capillary zone electrophoresis (CZE) is a simple and highly efficient separation method that separates analytes based on their size-to-charge ratios [10]. CZE-MS has been well recognized as a valuable platform for characterization of intact proteins because it can achieve highly efficient separation and sensitive detection of proteins [11–20]. Haselberg et al. reported high-resolution separation and highly sensitive detection of intact pharmaceutical proteins using CZE-MS with detection of 250 different proteoforms from recombinant human erythropoietin [17]. Han et al. reported a separation and detection of nine subunits of the Dam1 complex using CZE-MS with only 2.5-ng proteins loaded for analysis, and they also demonstrated that CZE-MS achieved similar signal-to-noise ratios (S/N) of Dam1 subunits to RPLC-MS with 100-fold less sample consumption (2.5 ng vs. 250 ng) [13]. In addition, detection of low-amole amounts of intact proteins using CZE-MS was reported in 1996 [11].

Recently, CZE-MS/MS has been applied for top-down proteomics of complex proteome samples [12, 14, 15, 20]. Zhao et al. coupled CZE to MS with an electro-kinetically pumped sheath flow interface [21, 22] for top-down proteomics of a *Mycobacterium marinum* secretome and a yeast cell lysate [14, 20]. They identified 58 proteoforms from the *Mycobacterium marinum* secretome in a single CZE-MS/MS run [20] and identified 600 proteoforms from the yeast cell lysate by coupling RPLC fractionation to CZE-MS/MS [14]. Han et al. coupled CZE to MS with a sheathless interface [23] for top-down proteomics of a *Pyrococcus furiosus* lysate, resulting in the identification of 291 proteoforms with RPLC-CZE-MS/MS [12]. Li et al. employed CZE-MS/MS for top-down characterization of 30–80 kDa proteins in a *P. aeruginosa PA01* whole cell lysate via the electrokinetically pumped sheath flow CE-MS interface and identified 30 proteins in the mass range of 30–80 kDa [15].

The low sample loading volume and narrow separation window of CZE had impeded CZE-MS/MS for large-scale top-down proteomics of complex proteome samples. More recently, our group improved the sample loading volume and separation window of CZE drastically via employing a 1-m-long separation capillary with high-quality neutral coating on its inner wall and one highly efficient protein stacking method based on a dynamic pH junction principle [24–26]. Our CZE-MS/MS system achieved a 90-min separation window and a 500-nL sample loading volume for top-down proteomics of *an Escherichia coli* (*E. coli*) proteome, resulting in an identification of 600 proteoforms in a single run. [25] We identified nearly 6000 proteoforms from the *E. coli* proteome with size exclusion chromatography (SEC)-RPLC fractionation and the CZE-MS/MS [27].

Building upon our previous work, here, we tried to boost the separation window and sample loading volume of our CZE-MS/MS system further via employing a much longer separation capillary compared to our previous work (1.5 m vs. 1 m). A much longer separation capillary leads to obviously lower electric field across the capillary and produces significantly lower electrophoretic velocity of proteoforms, resulting in a much wider separation window for acquisition of more MS/MS spectra of proteoforms for identifications. Besides a much wider separation window, a longer separation capillary also allows a larger sample loading volume without loss of separation efficiency theoretically. First, we compared the CZE-MS/MS systems with a 1.5-m-long capillary and a 1-m-long capillary regarding the separation performance of a standard protein mixture. We also evaluated the reproducibility and sample loading volume of the CZE-MS/MS system with a 1.5-m capillary for separation of the standard protein mixture. Second, we tested the CZE-MS/MS system for large-scale top-down proteomics of the *E. coli* proteome using 1 μg and 100 ng of *E. coli* proteins as the starting materials. Third, we applied the CZE-MS/MS system for quantitative top-down proteomics of zebrafish brain cerebellum (Cb) and optic tectum (Teo) regions.

Experimental Section

Materials and Reagents

Acrylamide was purchased from Acros Organics (NJ, USA). Standard proteins, ammonium bicarbonate (NH_4HCO_3), urea, dithiothreitol (DTT), iodoacetamide (IAA), and 3-(Trimethoxysilyl)propyl methacrylate were purchased from Sigma-Aldrich (St. Louis, MO). LC/MS grade water, acetonitrile (ACN), methanol, formic acid, and HPLC-grade acetic acid were purchased from Fisher Scientific (Pittsburgh, PA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were obtained from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail (provided in EASYpacks) was bought from Roche (Indianapolis, IN). Mammalian cell-PE LB™ buffer containing NP-40 detergent was purchased from G-Biosciences (St. Louis, MO) for protein extraction from zebrafish brain samples.

Sample Preparation

A mixture of standard proteins consisting of myoglobin (myo, 16.9 kDa, pI 7.0, 0.1 mg/mL, equine), carbonic anhydrase (CA, 29 kDa, pI 5.1, 0.5 mg/mL, bovine), and bovine serum albumin (BSA, 66.5 kDa, pI 5.0, 1.0 mg/mL) was prepared in LC-MS grade water and used as a stock solution. The stock solution was diluted by a factor of 100 with 50 mM NH_4HCO_3 (pH 8.0) for the CZE-MS experiment.

Escherichia coli (*E. coli*, strain K-12 substrain MG1655) was cultured in LB medium at 37 °C with 225 rpm shaking until OD600 reached 0.7. *E. coli* cells were harvested by centrifugation at 4000 rpm for 10 min. Then, the *E. coli* cells were washed with PBS three times, followed by cell lysis in a lysis buffer containing 8 M urea, 100 mM Tris-HCl (pH 8.0), and protease inhibitors. Sonication with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) was performed on ice for 10 min to reach complete cell lysis. The supernatant containing the extracted proteins was collected after centrifugation at 18000*g* for 10 min. A small aliquot of the extracted proteins was used for bicinchoninic acid (BCA) assay to determine the protein concentration. The leftover proteins were stored at − 80° °C before use.

One milligram of *E. coli* proteins in 8 M urea and 100 mM Tris-HCl (pH 8.0) was denatured at 37 °C, reduced with DTT by adding 1.7 μL of 1 M DTT solution, and alkylated with IAA by adding 4.0 μL of 1 M IAA solution. Then, the proteins were desalted with a C4-trap column (Bio-C4, 3 μm, 300 Å, 4.0 mm i.d., 10 mm long) from Sepax Technologies, Inc. (Newark, DE). An HPLC system (Agilent Technologies, 1260 Infinity II) was used. The HPLC eluate containing the *E. coli* proteins and 80% (*v*/*v*) ACN from the trap column were collected and lyophilized with a vacuum concentrator (Thermo Fisher Scientific). The dried protein sample was redissolved in 50 mM NH_4HCO_3 (pH 8.0) to reach a 2 mg/mL protein concentration, as determined by the BCA assay, for CZE-MS/MS analyses.

Zebrafish brain cerebellum (Cb) and optic tectum (Teo) regions were collected from three mature female zebrafish (AB/Tuebingen line). The zebrafish brain samples were kindly provided by Professor Jose Cibelli's group at the Department of Animal Science of Michigan State University. The whole protocol related to the zebrafish was performed following guidelines defined by the Institutional Animal Care and Use Committee of Michigan State University. Zebrafish brains were frozen in liquid nitrogen immediately after the sample collection and then transferred to a − 80° °C freezer for storage. After washing with PBS for a couple of times to remove the blood, the three Cb and three Teo samples from three fishes were pooled to get one Cb sample and one Teo sample, followed by protein extraction with the mammalian cell-PE LB™ buffer plus complete protease inhibitors. Homogenization with a Homogenizer 150 (Fisher Scientific, Pittsburgh, PA) on ice and sonication with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 10 min were performed to assist the protein extraction. After centrifugation at 18000*g* for 10 min, the supernatant containing the extracted proteins was collected, and a small aliquot of the proteins was used for BCA assay to determine the protein concentration. The leftover proteins were used for the experiments.

Approximately 1 mg of zebrafish proteins in the lysis buffer was denatured at 37 °C, reduced with DTT by adding 1.5 μL of 1 M DTT solution, and alkylated with IAA by adding 3.8 μL of 1 M IAA solution. Next, the proteins were transferred to Microcon-30 kDa centrifugal filter units for cleanup. The proteins on the membrane were washed with 8 M urea for three times to remove the NP-40 detergent and then washed with 50 mM NH_4HCO_3 (pH 8.0) three times to remove urea. Finally, the proteins from the Cb and Teo regions were redissolved in 50 mM NH_4HCO_3 buffer on the membrane via gently shaking for 30 min at room temperature. The Cb and Teo samples with a 1-mg/mL protein concentration were analyzed by CZE-MS/MS in triplicate.

CZE-ESI-MS/MS Analysis

An ECE-001 CE autosampler from the CMP Scientific (Brooklyn, NY) was used for automated CE operation. The CE system was coupled to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) through a commercialized electrokinetically pumped sheath flow CE-MS interface (CMP Scientific, Brooklyn, NY) [21, 22]. A fused silica capillary (50 μm i.d., 360 μm o.d., 1 m or 1.5 m in length) was used for CZE separation. The inner wall of the capillary was coated with linear polyacrylamide (LPA) based on references [26, 28]. One end of the capillary was etched with hydrofluoric acid based on reference [29] to reduce the outer diameter of the capillary. *(Caution: use appropriate safety procedures while handling hydrofluoric acid solutions.)* The background electrolyte (BGE) used for CZE was 10% (*v*/*v*) acetic acid (pH ~ 2.2). The sheath buffer was 0.2% (*v*/*v*) formic acid containing 10% (*v*/*v*) methanol. Sample injection was carried out by applying pressure (5–10 psi) at the sample injection end, and the injection periods were calculated based on the Poiseuille's law for different sample loading volumes. High voltage (30 kV) was applied at the injection end of the separation capillary for separation, and 2–2.2 kV was applied in the sheath buffer vial for ESI. In the end of each CZE-MS run, we flushed the capillary with BGE by applying 20-psi pressure for 10 min. The ESI emitters were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d., and 10 cm length) with a Sutter P-1000 flaming/brown micropipet puller. The opening size of the ESI emitters was 30–40 μm.

The Q-Exactive HF mass spectrometer was used for all the experiments. For the standard protein mixture, the MS parameters were as follows: The mass resolution was 120,000 (at *m*/*z* 200), the number of microscans was one, the AGC target value was 1E6, the maximum injection time was 50 ms, and the scan range was 600–2000 *m*/*z*. For the *E.coli* and fish brain samples, top 8 data-dependent acquisition (DDA) methods were used. For MS, we used a 240,000 mass resolution (at *m*/*z* 200), three microscans, 1E6 AGC target value, 50 ms maximum injection time, and 600–2000 *m*/*z* scan range. For MS/MS, the mass resolution was 120,000 (at *m*/*z* 200), the number of microscans was 3, the AGC value was 1E5, the maximum injection time was 200 ms, the isolation window was 4 *m*/*z*, and the normalized collision energy (NCE) was 20%. The top 8 most intense ions in one MS spectrum were sequentially isolated in the quadrupole, followed by higher energy collision dissociation (HCD). Only ions in each MS spectrum with intensities higher than 1E5 and charge states higher than 2 (for zebrafish brain samples) or higher than 5 (for the *E. coli* samples) were selected

for HCD fragmentation. The dynamic exclusion was enabled and was set to 30 s. The "exclude isotopes" function was turned on.

Data Analysis

The standard protein, *E. coli*, and zebrafish brain data were analyzed using Xcalibur software (Thermo Fisher Scientific) to get intensity and migration time of proteins. The electropherograms were exported from Xcalibur and were further formatted using Adobe Illustrator to make the final figures.

All the *E. coli* and zebrafish RAW files were analyzed by the TopPIC (TOP-down mass spectrometry based proteoform identification and characterization) pipeline for proteoform identification and quantification [30]. The RAW files were first converted into mzML files with msconvert tool [31]. Then, a TopFD (TOP-down mass spectrometry feature detection) tool was used to perform spectral deconvolution and generate msalign files. Finally, TopPIC (version 1.2.2) was used for database searching with msalign files as input. UniProt databases of *E. coli* (UP000000625) and zebrafish (AUP000000437) were used for the database search. Cysteine carbamidomethylation was set as a fixed modification, and the maximum number of unexpected modifications was 1 for the zebrafish data or 2 for *E. coli* data. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da. The target-decoy approach was used to estimate the false discovery rates (FDRs) of proteoform identifications [32, 33]. A 5% proteoform-level FDR was used to filter the proteoform identifications. To reduce the redundancy of proteoform identifications, if the proteoforms were identified by multiple spectra that corresponded to the same proteoform feature reported by TopFD or these proteoforms were from the same protein and had smaller than 1.2-Da precursor mass differences, we considered these proteoforms as one proteoform identification.

For label-free quantification of zebrafish brain Cb and Teo regions, the TopFD tool grouped top-down spectral peaks into isotopomer envelopes and combined isotopomer envelopes from the same proteoform with different migration times and charge states. These combined envelopes were then reported as CZE-MS features. The peak intensity of a feature was calculated as the sum of the intensities of its corresponding peaks and was used for proteoform quantification to compare the proteoform abundance between the Cb and Teo. Migration time alignment was employed to correct migration time shifts and find matched features between CZE-MS/MS runs. All proteoform identifications from the six CZE-MS/MS runs (three runs for the CB and three runs for the Teo) were combined for proteoform quantifications. Two proteoforms identified from two runs were considered as the same identification if their CZE-MS features were matched. For each identified proteoform, we found the feature of the proteoform and searched matched features in the other five CZE-MS/MS runs. There were three cases. (a) If a matched feature was found and there were identified MS/MS spectra for the precursor, the feature intensity and the scan number of an identified MS/MS spectrum were reported for the proteoform. (b) If a matched feature was found and there were no identified MS/MS spectra for the precursor, the feature intensity was reported for the proteoform, and the scan number was reported as blank. (c) If a matched feature was not found, the feature intensity and scan number were reported as blank for the proteoform. Only proteoforms having feature intensities in all the six CZE-MS/MS runs were considered as quantified proteoforms for comparison in this work. The output of the proteoform quantification data was further analyzed by the Perseus software to perform basic processing, *t* test analyses, and generate the volcano plot [34].

Results and Discussion

Evaluation of the CZE-MS System with a 1.5-M-Long Separation Capillary Using a Standard Protein Mixture

We first compared a 1.5-m capillary and a 1-m capillary regarding the separation window and protein intensity using a standard protein mixture containing myoglobin (myo, 16.9 kDa), carbonic anhydrase (CA, 29 kDa), and bovine serum albumin (BSA, 66.5 kDa), Figure 1 a. The sample injection volume per CZE-MS run was 500 nL. The three proteins were baseline separated by CZE using both the 1.5-m and 1-m capillaries. The 1.5-m capillary produced much wider separation window than the 1-m capillary for the three proteins (11 min vs. 5 min) because the proteins migrated slower in the 1.5-m capillary, evidenced by their much longer migration time. The base peak intensities of the three proteins from the 1.5-m capillary were roughly twofold lower than that from the 1-m capillary because the proteins diffused for a much longer time in the 1.5-m capillary. There was another strong peak after the CA peak. That peak represented one impurity in the sample, and the impurity was identified as superoxide dismutase (SD) based on our MS/MS data, Figure S1 in supporting information I. Interestingly, we observed that the intensity of SD was lower than that of CA and myo in the 1-m-capillary data, which was different from the 1.5-m-capillary data. Based on our replicated runs, relative intensities of SD and other proteins from the 1-m capillary experiment (data not shown) were not as consistent as that from the 1.5-m capillary experiment (Figure 1b). The phenomenon may be due to the much narrower protein peaks in the 1-m capillary data, leading to much smaller numbers of data points across the peaks.

We further evaluated the reproducibility of our CZE-MS system with the 1.5-m capillary for separation of the standard protein mixture, Figure 1 b. The system showed reasonably good reproducibility regarding the separation profiles, base peak intensities of proteins with relative standard deviations (RSDs) less than 25%, and migration time of proteins with RSDs less than 2%.

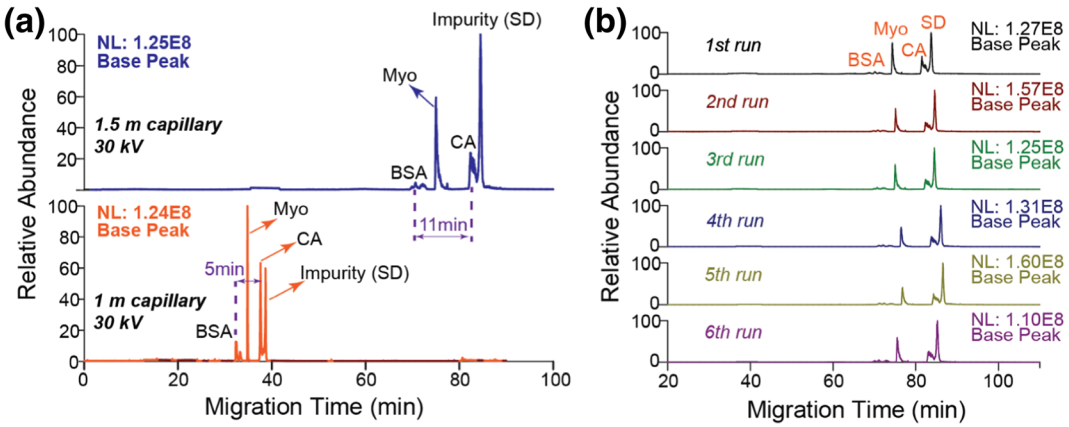


Figure 1. CZE-MS analyses of a standard protein mixture. (**a**) Electropherograms of the standard protein mixture after CZE-MS with a 1.5-m capillary and a 1-m capillary. (**b**) Electropherograms of the standard protein mixture after CZE-MS analyses in sextuplicate. A 1.5-m-long and LPA-coated capillary was used. Gaussian smoothing (5 points) was applied

Then, we tested three different sample loading volumes using the 1.5-m capillary. The sample loading volumes were 0.5 µL (17% of the total capillary volume), 1 µL (33% of the total capillary volume), and 2 µL (67% of the total capillary volume), corresponding to 8 ng, 16 ng, and 32 ng of total proteins, Figure 2. When the sample loading volume was increased from 0.5 to 2 µL, the separation window of the system for the proteins was boosted by over 100%, Figure 2 a. The migration time of proteins increased obviously with the increase of sample loading volume, Figure 2 b. The protein

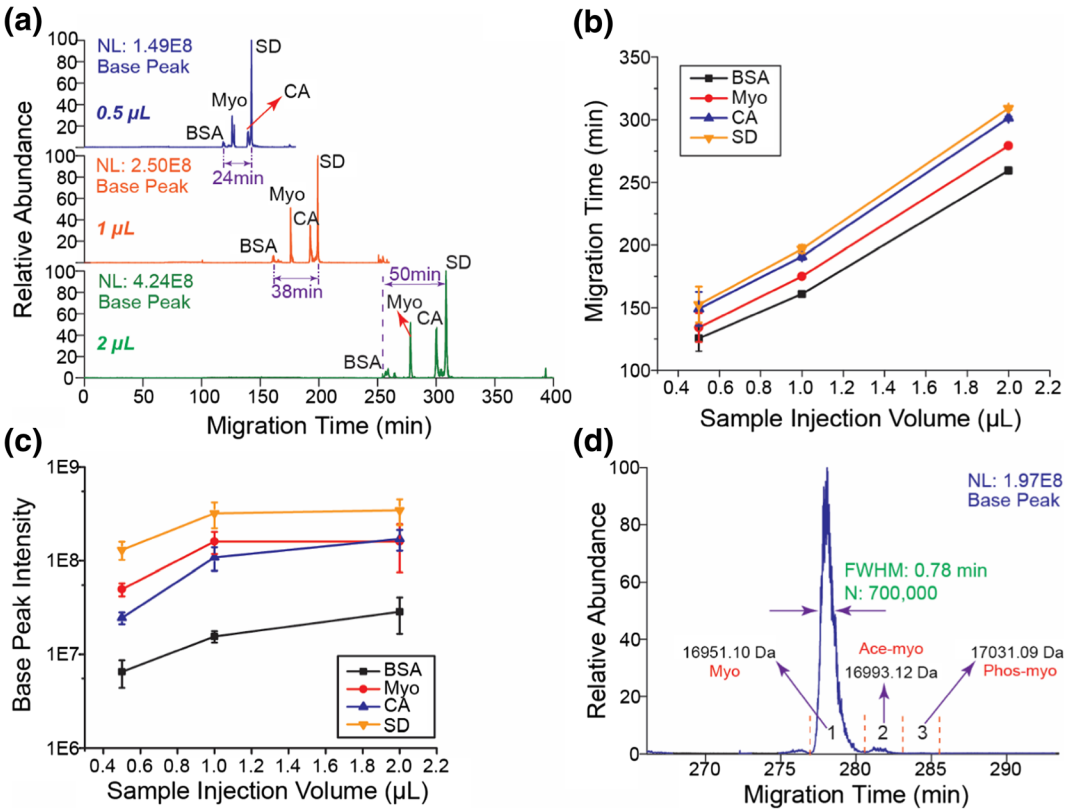


Figure 2. CZE-MS analyses of a standard protein mixture with three different sample loading volumes (0.5 µL, 1 µL, and 2 µL). A 1.5-m-long and LPA-coated separation capillary was used. Duplicate CZE-MS runs were performed with each sample loading volume. (**a**) Electropherograms of the standard protein mixture with the three different sample loading volumes. (**b**) Migration time of proteins as a function of sample loading volume. (**c**) Base peak intensity of proteins as a function of sample loading volume. (**d**) The zoomed-in peak of myoglobin from one CZE-MS run with a 2-µL sample loading volume. The full peak width at half maximum (FWHM) and the number of theoretical plates of the peak (N) are shown. Three different myoglobin peaks (1, 2, and 3) representing three different myoglobin proteoforms are highlighted. The error bars in (**b**) and (**c**) are standard deviations of migration time and intensity of proteins from the duplicate CZE-MS/MS runs

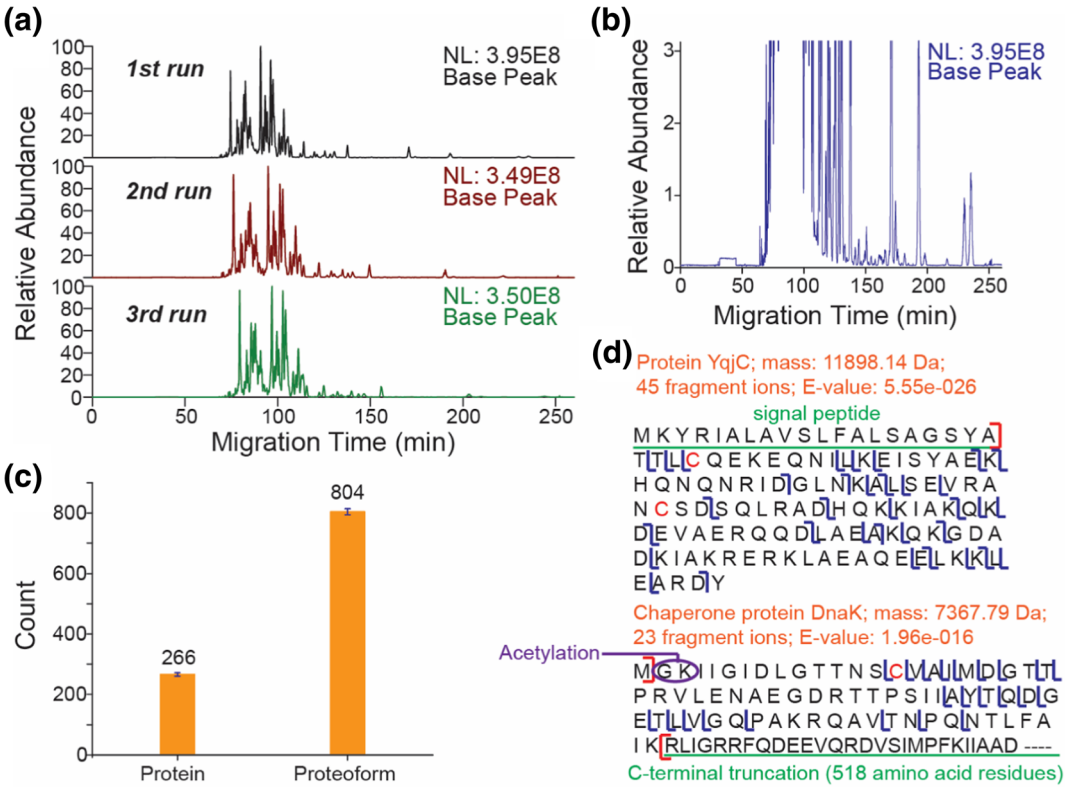


Figure 3. CZE-MS analyses of the *E. coli* proteome using a 1.5-m-long and LPA-coated separation capillary. One microgram of proteins was injected per CZE-MS/MS run. The CZE-MS/MS analyses were performed in triplicate. (**a**) Base peak electropherograms of the CZE-MS/MS runs. (**b**) A zoomed-in electropherogram of one CZE-MS/MS run. (**c**) Protein and proteoform identifications. The error bars represent the standard deviations of the number of identifications from the triplicate CZE-MS/MS runs. (**d**) Sequences and fragmentation patterns of protein YqjC and chaperone protein DnaK. The cysteine (C) residues marked in red have carbamidomethylation modification. There is one acetylation modification on the G or K residue for DnaK

sample was dissolved in 50 mM NH_4HCO_3 (pH 8.0) for CZE-MS, the BGE for CZE was 10% (*v*/*v*) acetic acid (pH ~ 2.2). At the beginning of CZE, the proteins were concentrated in the capillary based on a dynamic pH junction method [35, 36]. When the sample loading volume increased, the time required for sample zone titration and protein stacking was lengthened, leading to longer migration time of proteins and a wider separation window of the proteins. When the sample volume was increased from 0.5 to 1 µL, the protein intensity was boosted by, on average, threefold, Figure 2 c. The protein intensity was increased by, on average, only 30% when the sample volume was increased from 1 to 2 µL, Figure 2 c.

The CZE-MS system produced extremely high separation efficiency of myoglobin, Figure 2 d. The numbers of theoretical plates of myoglobin were estimated to be around 200,000, 700,000, and 1,000,000 with 0.5-µL, 2-µL, and 1-µL sample loading volumes. More importantly, three different myoglobin proteoforms with an over 100-fold concentration dynamic range were baseline separated and detected by the CZE-MS system with a 2-µL sample loading volume, Figure 2 d. The three different proteoforms were highlighted in the Figs. (1, 2, and 3). They were myoglobin without any post-translational modification (PTM) (1, average mass 16,951.10 Da), acetylated myoglobin (2, average mass 16,993.12 Da), and phosphorylated myoglobin (3, average mass 17,031.09 Da). We also performed bottom-up proteomics analysis of a myoglobin digest and identified acetylated and phosphorylated myoglobin peptides with high confidence, Figure S2 in supporting information I. The corresponding experimental details of the bottom-up proteomics experiment are shown in supporting information I. Based on the information of myoglobin (equine) in the UniProt database, no experimental evidence on the acetylated and phosphorylated myoglobin proteoforms was reported before.

Top-Down Proteomics of E. coli Cells Using Single-Shot CZE-MS/MS with a 1.5-M-Long Separation Capillary

As discussed before, a larger sample loading volume produced a wider separation window and higher protein intensity, Figure 2 a and c. However, the time required for one CZE-MS run was also drastically increased. For CZE-MS/MS analyses of *E. coli* samples and zebrafish brain samples in this work, the sample loading volume was 500 nL per run to control the instrument time. A 1.5-m-long LPA-coated separation capillary was used.

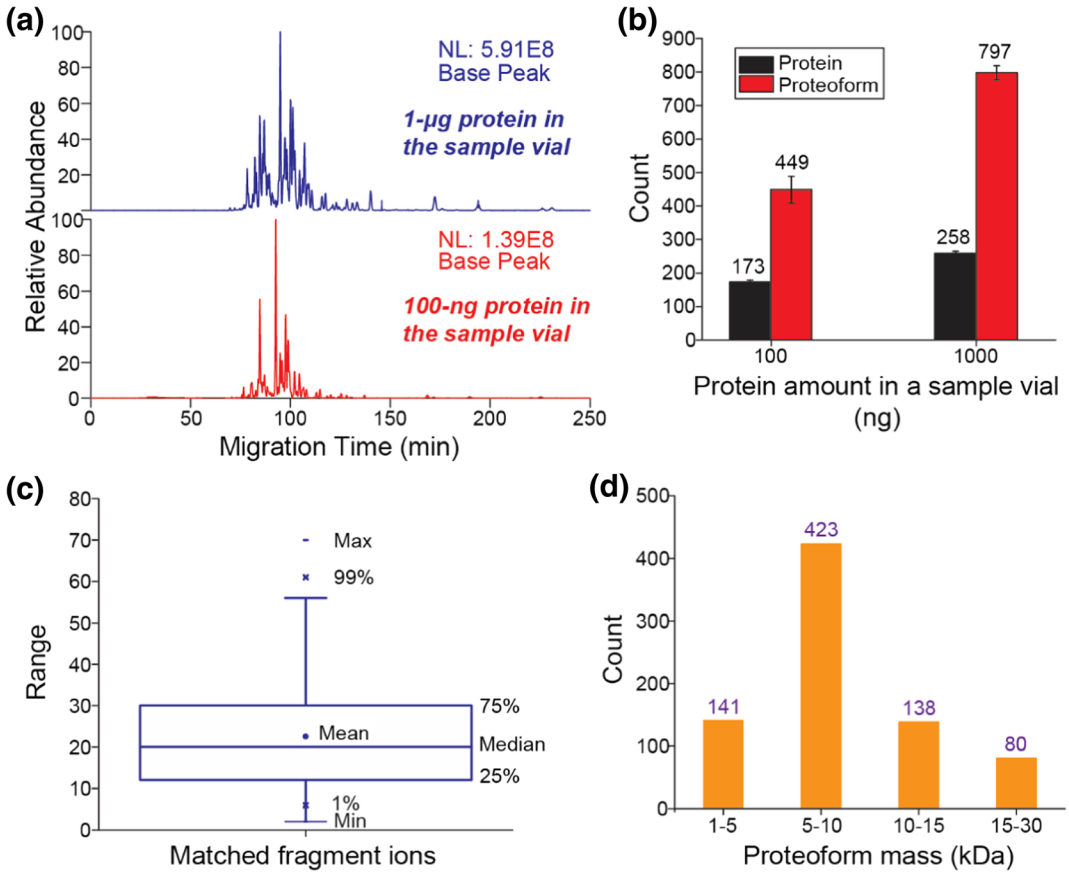


Figure 4. CZE-MS/MS analyses of mass-limited *E. coli* proteome samples. (**a**) Base peak electropherograms of the *E. coli* proteome sample with 1-µg and 100-ng proteins as the starting materials. (**b**) Proteoform and protein identifications from the analyses of 1-µg and 100-ng *E. coli* samples. The error bars represent the standard deviations of the number of identifications from duplicate CZE-MS/MS runs. (**c**) Box plot of the number of matched fragment ions of identified proteoforms from one CZE-MS/MS analysis of the 1-µg *E. coli* sample. (**d**) Mass distribution of the identified proteoforms from one CZE-MS/MS analysis of the 1-µg *E. coli* sample

The CZE-MS/MS system produced reproducible separation and detection of the *E. coli* proteome, Figure 3 a. The separation window was about 180 min, Figure 3 b, and it was 100% wider than that in our previous work using a 1-m-long and LPA-coated capillary [25]. Single-shot CZE-MS/MS identified 804 ± 10 proteoforms and 266 ± 6 proteins (n = 3) from the *E. coli* proteome with 1-µg proteins injected, Figure 3 c. The number of proteoform and protein identifications were boosted by over 30% compared to our previous work using a 1-m-long capillary [25]. N-terminal methionine excision, signal peptide cleavage, truncations, and various PTMs including acetylation, methylation, oxidation, and phosphorylation were detected. Figure 3 d shows the sequences and fragmentation patterns of protein YqjC and chaperone protein DnaK. These two proteins were identified with high numbers of fragment ions. For the protein YqjC, signal peptide cleavage was detected. For the chaperone protein DnaK, N-terminal methionine excision, acetylation, and C-terminal truncation were detected. The identified proteoforms from the *E. coli* samples are listed in supporting information II.

We then tested our CZE-MS/MS system for top-down proteomics of mass-limited *E. coli* proteome samples. One microgram of *E. coli* proteins dissolved in 2 µL of 50 mM NH_4HCO_3 (pH 8.0) was put into a sample vial for CZE-MS/MS analyses in duplicate. Five hundred nanoliters of the sample (25% of the total sample volume) corresponding to 250 ng of proteins was injected per CZE-MS/MS run. One base peak electropherogram is shown in Figure 4 a, and the separation profile and base peak intensity were comparable to that in Figure 3 a, although the sample loading amount was fourfold lower than that for Figure 3 a. Nearly 800 proteoforms and 260 proteins were identified using single-shot CZE-MS/MS with only 250-ng proteins injected, Figure 4 b. More importantly, the numbers of proteoform and protein identifications were almost the same as that from CZE-MS/MS analyses with 1-µg proteins loaded. We further tested the system with only 100-ng *E. coli* proteins as the starting material. One hundred nanograms of proteins dissolved in 2 µL of 50 mM NH_4HCO_3 (pH 8.0) was put into a sample vial for analyses in duplicate, and only 25 ng of proteins were injected per

run. The base peak intensity was roughly four times lower than that from the 250-ng protein run (1.39E8 vs. 5.91E8), Figure 4 a. The CZE-MS/MS system still identified 449 proteoforms and 173 proteins with only 100-ng *E. coli* proteins as the starting material, Figure 4 b. The data here highlight the power of CZE-MS/MS for top-down proteomics of mass-limited proteome samples. The identified proteoforms are listed in supporting information II.

Figure 4 c shows the distribution of the number of matched fragment ions of identified proteoforms from one CZE-MS/MS run with 250-ng *E. coli* proteins injected. The mean was 23, and the median was 20. One fourth of the proteoforms were identified with 12 or fewer matched fragment ions. Majority of the identified proteoforms from one CZE-MS/MS run (250-ng proteins loaded) had masses lower than 15 kDa; 80 proteoforms had masses in a range of 15–30 kDa, Figure 4 d. Detection and identification of large proteoforms (> 30 kDa) in a complex proteome sample are still difficult for top-down proteomics due to several reasons. First, signal-to-noise ratios of proteins drastically decrease with increasing protein molecular weight [37]. Second, co-migration or co-elution of small and large proteins in liquid-phase separation systems makes the MS detection of large proteins challenging. Third, the limited mass resolution of most mass analyzers makes it difficult to determine accurate masses of large proteins.

Quantitative Top-Down Proteomics of Zebrafish Brain Cb and Teo Regions Using CZE-MS/MS with a 1.5-M-Long Separation Capillary

Figure 5 a shows a diagram of a mature zebrafish brain. We collected the Cb and Teo regions from three fishes. We combined the three Cb samples and pooled the three Teo samples for protein extraction. Proteins extracted from the Cb and Teo were dissolved in 50 mM NH_4HCO_3 (pH 8.0) to reach a final concentration of ~ 1 mg/mL for CZE-MS/MS analyses in

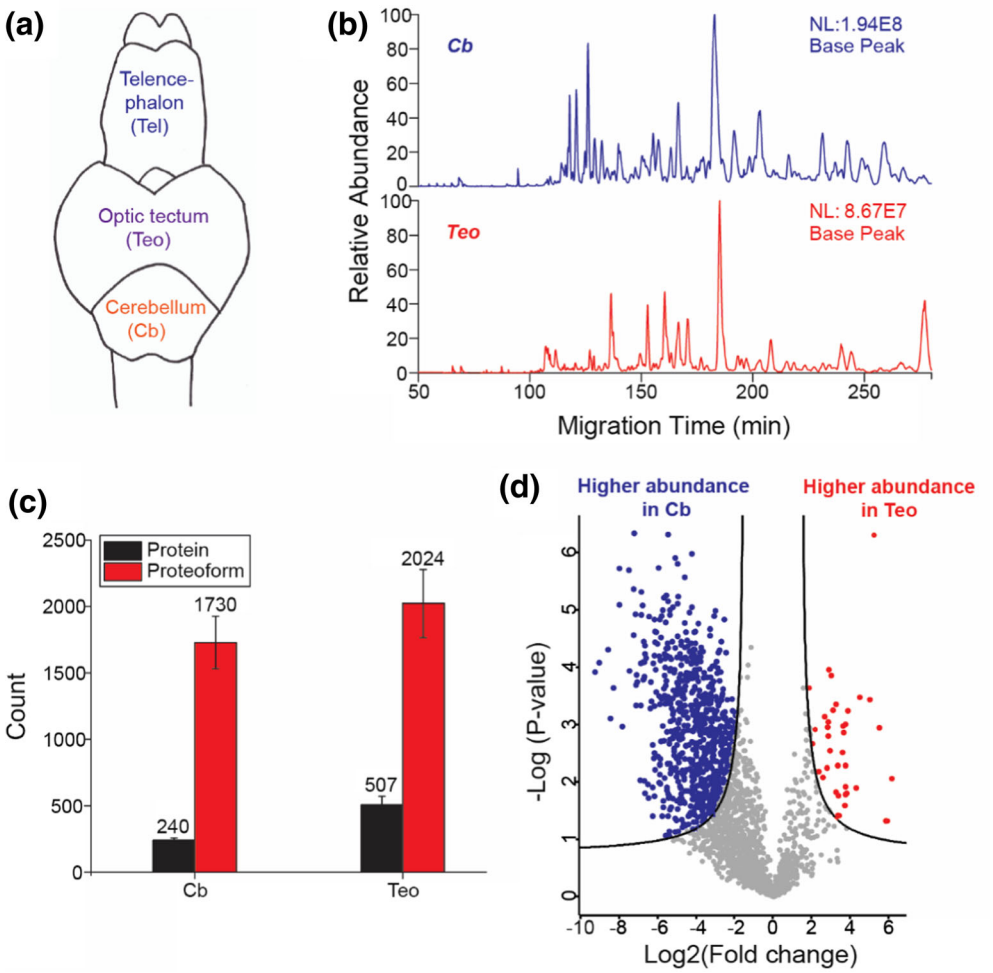


Figure 5. CZE-MS/MS analyses of zebrafish brain Cb and Teo regions. (**a**) A diagram of a mature zebrafish brain. (**b**) Base peak electropherograms of zebrafish brain Cb and Teo after CZE-MS/MS analyses. (**c**) Protein and proteoform identifications from the Cb and Teo samples. The error bars represent the standard deviations of the number of identifications from triplicate CZE-MS/MS runs. (**d**) Volcano plot of the quantified proteoforms. For each proteoform, the mean of proteoform intensities from triplicate CZE-MS/MS analyses of Teo or Cb was calculated, the ratio of means between Teo and Cb (fold change) was obtained, and log2 (fold change) is shown in the *x*-axis. The –log (*p* value) from *t* test is shown in the *y*-axis. The differentially expressed proteoforms after the *t* test (FDR ≤ 1% and S0 = 1) are marked in blue (higher abundance in Cb) and red (higher abundance in Teo)

triplicate. Each CZE-MS/MS run had an injection volume of 500 nL, corresponding to about 500 ng of proteins. A 1.5-m-long and LPA-coated separation capillary was used.

The CZE-MS/MS system reached a 180-min separation window for both Cb and Teo samples, Figure 5 b. The Cb and Teo samples showed drastically different separation profiles and significantly different base peak intensity (1.9E8 vs. 8.7E7), although the same amount of total proteins was loaded for the Cb and Teo samples, Figure 5 b. The CZE-MS/MS system identified 240 proteins and 1730 proteoforms from the Cb sample and identified 507 proteins and 2024 proteoforms from the Teo sample, Figure 5 c. The data suggest the significant difference between the Cb and Teo proteome samples. The data also clearly demonstrate the capability of our CZE-MS/MS system for thousands of proteoform identifications from complex proteomes with the consumption of nanograms of proteins. The identified proteoforms from Cb and Teo samples are listed in supporting information II.

We then quantitatively compared the Cb and Teo samples with a label-free approach based on the proteoform feature intensity. All the proteoform identifications from the six CZE-MS/MS runs were combined for proteoform quantifications, and about 4000 proteoforms were identified from the six runs. Only proteoforms having feature intensities in all the six CZE-MS/MS runs were considered as quantified proteoforms for comparison. We quantified about 2000 proteoforms. The feature intensities of each quantified proteoform were normalized to the proteoform feature intensity from the first CZE-MS/MS run of the Cb sample, and a Log2 transformation was applied. We employed the Perseus software [34] to perform a *t* test for validating the significance of proteoform intensity difference between Cb and Teo samples with FDR as 1% and S0 as 1. The S0 represents the artificial variance within groups (Cb and Teo) variance and controls the relative importance of *p* value and difference between means of proteoform intensity from triplicate CZE-MS/MS runs [38].

Figure 5 d shows the volcano plot of the quantified proteoforms. In total, 786 proteoforms showed significant differences in abundance between the Cb and Teo samples. Seven hundred forty-nine proteoforms from 131 proteins had significantly higher abundance in Cb; 37 proteoforms from 26 proteins showed higher abundance in Teo. The differentially expressed proteoforms are listed in supporting information II. We further performed biological process enrichment analysis for the 131 proteins that showed much higher abundance in Cb compared to Teo, and the DAVID Bioinformatics Resources 6.8 was used for the analysis [39]. These proteins were highly enriched in several biological processes, including muscle contraction (*p* value: 1E-4), glycolytic process (*p* value: 5E-16), and mesenchyme migration (*p* value: 0.01). We need to note that more experiments with multiple biological replicates are essential before solid biological conclusions can be made from the data. We consider this experiment as the first try of quantitative top-down proteomics of complex proteome samples using CZE-MS/MS. We also need to note that a big portion of the identified proteoforms from the Cb and Teo samples had masses lower than 5 kDa.

We further roughly estimated the dynamic range of proteoform abundance from single-shot CZE-MS/MS analysis of the Cb and Teo samples using the proteoform feature intensity, Figure 6. Single-shot CZE-MS/MS achieved up to six orders of magnitude dynamic range in proteoform abundance from the Cb and Teo samples, Figure 6 a. We noted that the obvious mass difference of proteoforms could influence the accuracy of the determined dynamic range. We then manually checked two proteoforms of parvalbumin 4 with very similar mass (11,558 Da vs.11,542 Da). These two proteoforms were both identified with high confidence, and

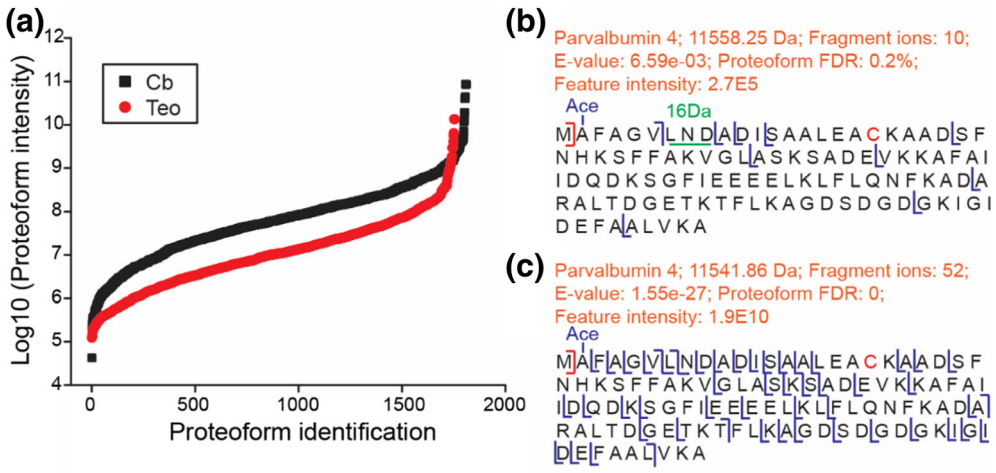


Figure 6. (**a**) Dynamic range of proteoform abundance from single-shot CZE-MS/MS analysis of Cb and Teo samples. The proteoform feature intensity was used to estimate the dynamic range roughly. (**b**) The sequence and fragmentation pattern of one proteoform of parvalbumin 4 with feature intensity as 2.7E5. (**c**) The sequence and fragmentation pattern of one proteoform of parvalbumin 4 with feature intensity as 1.9E10. N-terminal methionine removal, acetylation, and carbamidomethylation of cysteine residues (marked in red) were detected for both proteoforms. A 16-Da total mass shift was also detected for the proteoform in (**b**)

the feature intensity of one proteoform (Figure 6b) was nearly five orders of magnitude lower than that of the other one (Figure 6c). We drew two conclusions from the data. First, single-shot CZE-MS/MS can approach at least five orders of magnitude dynamic range in proteoform abundance. Second, different proteoforms from the same gene can have dramatically different abundance.

Conclusions

In this work, we present a CZE-MS/MS system with a 1.5-m-long and LPA-coated separation capillary for large-scale top-down proteomics of the *E. coli* cells and zebrafish brain Cb and Teo regions. The CZE-MS/MS system achieved a 180-min separation window and a 2-μL sample loading volume for analysis of protein mixtures. Single-shot CZE-MS/MS identified about 800 proteoforms and 450 proteoforms from the *E. coli* proteome with only 250 ng and 25 ng of proteins injected. Thousands of proteoforms were identified and quantified from the zebrafish brain Cb and Teo in one CZE-MS/MS run with the consumption of 500-ng proteins. The data highlight the power of our CZE-MS/MS system for large-scale top-down proteomics of mass-limited samples. We expect that the CZE-MS/MS system can be potentially useful for top-down characterization of tissue samples from laser capture microdissection, circulating tumor cells, and even single mammalian cells.

We need to point out that "complete identification and characterization" of proteoforms in top-down proteomics is still not straightforward because it requires accurate determination and localization of all modifications on the protein sequences. In this work, we identified proteoforms based on their precursor and fragment masses as well as a 5% proteoform-level FDR. Therefore, we did not accurately determine and localize the modifications for many identified proteoforms. We expect that more extensive gas-phase fragmentation of proteoforms will benefit the "complete identification and characterization" of proteoforms.

Acknowledgements

We thank Prof. Heedeok Hong's group at the Department of Chemistry of Michigan State University for kindly providing the *E. coli* cells for this project. We thank Prof. Jose Cibelli and Mr. Billy Poulos at the Department of Animal Science of Michigan State University for their help on collecting zebrafish brains for the project. We thank the support from the National Institute of General Medical Sciences, National Institutes of Health (NIH), through Grant R01GM118470 (X. Liu) and R01GM125991 (L. Sun and X. Liu).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no competing interest.

References

1. 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. **480**, 254–258 (2011)
2. Durbin, K.R., Fornelli, L., Fellers, R.T., Doubleday, P.F., Narita, M., Kelleher, N.L.: Quantitation and identification of thousands of human proteoforms below 30 kDa. J. Proteome Res. **15**, 976–982 (2016)
3. 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. Anal. Chem. **89**, 5467–5475 (2017)
4. 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. Proc. Natl. Acad. Sci. U. S. A. **110**, 10153–10158 (2013)
5. Shen, Y., Tolić, N., Piehowski, P.D., Shukla, A.K., Kim, S., Zhao, R., Qu, Y., Robinson, E., Smith, R.D., Paša-Tolić, L.: High-resolution ultrahigh-pressure long column reversed-phase liquid chromatography for top-down proteomics. J. Chromatogr. A. **1498**, 99–110 (2017)
6. Fornelli, L., Durbin, K.R., Fellers, R.T., Early, B.P., Greer, J.B., LeDuc, R.D., Compton, P.D., Kelleher, N.L.: Advancing top-down analysis of the human proteome using a benchtop quadrupole-orbitrap mass spectrometer. J. Proteome Res. **16**, 609–618 (2017)
7. 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. J. Proteome Res. **16**, 1087–1096 (2017)
8. Schaffer, L.V., Rensvold, J.W., Shortreed, M.R., Cesnik, A.J., Jochem, A., Scalf, M., Frey, B.L., Pagliarini, D.J., Smith, L.M.: Identification and quantification of murine mitochondrial proteoforms using an integrated top-down and intact-mass strategy. J. Proteome Res. **17**, 3526–3536 (2018)
9. Riley, N.M., Sikora, J.W., Seckler, H.S., Greer, J.B., Fellers, R.T., LeDuc, R.D., Westphall, M.S., Thomas, P.M., Kelleher, N.L., Coon, J.J.: The value of activated ion electron transfer dissociation for high-throughput top-down characterization of intact proteins. Anal. Chem. **90**, 8553–8560 (2018)
10. Jorgenson, J.W., Lukacs, K.D.: Capillary zone electrophoresis. Science. **222**, 266–272 (1983)
11. Valaskovic, G.A., Kelleher, N.L., McLafferty, F.W.: Attomole protein characterization by capillary electrophoresis-mass spectrometry. Science. **273**, 1199–1202 (1996)
12. Han, X., Wang, Y., Aslanian, A., Bern, M., Lavallée-Adam, M., Yates 3rd, J.R.: Sheathless capillary electrophoresis-tandem mass spectrometry for top-down characterization of Pyrococcus furiosus proteins on a proteome scale. Anal. Chem. **86**, 11006–11012 (2014)
13. Han, X., Wang, Y., Aslanian, A., Fonslow, B., Graczyk, B., Davis, T.N., Yates 3rd, J.R.: In-line separation by capillary electrophoresis prior to analysis by top-down mass spectrometry enables sensitive characterization of protein complexes. J. Proteome Res. **13**, 6078–6086 (2014)
14. Zhao, Y., Sun, L., Zhu, G., Dovichi, N.J.: Coupling capillary zone electrophoresis to a Q Exactive HF mass spectrometer for top-down proteomics: 580 proteoform identifications from yeast. J. Proteome Res. **15**, 3679–3685 (2016)
15. Li, Y., Compton, P.D., Tran, J.C., Ntai, I., Kelleher, N.L.: Optimizing capillary electrophoresis for top-down proteomics of 30-80 kDa proteins. Proteomics. **14**, 1158–1164 (2014)
16. Sun, L., Knierman, M.D., Zhu, G., Dovichi, N.J.: Fast top-down intact protein characterization with capillary zone electrophoresis-electrospray ionization tandem mass spectrometry. Anal. Chem. **85**, 5989–5995 (2013)
17. Haselberg, R., de Jong, G.J., Somsen, G.W.: Low-flow sheathless capillary electrophoresis-mass spectrometry for sensitive glycoform profiling of intact pharmaceutical proteins. Anal. Chem. **85**, 2289–2296 (2013)
18. Bush, D.R., Zang, L., Belov, A.M., Ivanov, A.R., Karger, B.L.: High resolution CZE-MS quantitative characterization of intact

biopharmaceutical proteins: proteoforms of interferon-β1. Anal. Chem. **88**, 1138–1146 (2016)
19. Sarg, B., Faserl, K., Kremser, L., Halfinger, B., Sebastiano, R., Lindner, H.H.: Comparing and combining capillary electrophoresis electrospray ionization mass spectrometry and nano-liquid chromatography electrospray ionization mass spectrometry for the characterization of post-translationally modified histones. Mol. Cell. Proteomics. **12**, 2640–2656 (2013)
20. Zhao, Y., Sun, L., Champion, M.M., Knierman, M.D., Dovichi, N.J.: Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry for top-down characterization of the Mycobacterium marinum secretome. Anal. Chem. **86**, 4873–4878 (2014)
21. 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 Commun. Mass Spectrom. **24**, 2554–2560 (2010)
22. 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. J. Proteome Res. **14**, 2312–2321 (2015)
23. Moini, M.: Simplifying CE-MS operation. 2. Interfacing low-flow separation techniques to mass spectrometry using a porous tip. Anal. Chem. **79**, 4241–4246 (2007)
24. Chen, D., Shen, X., Sun, L.: Capillary zone electrophoresis-mass spectrometry with microliter-scale loading capacity, 140 min separation window and high peak capacity for bottom-up proteomics. Analyst. **14**, 2118–2127 (2017)
25. 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. Anal. Chem. **89**, 12059–12067 (2017)
26. 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. J. Vis. Exp. **140**, e58644 (2018)
27. 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. Anal. Chem. **90**, 5529–5533 (2018)
28. 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. **146**, 839–843 (2016)
29. 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. Angew. Chem. Int. Ed. **52**, 13661–13664 (2013)
30. Kou, Q., Xun, L., Liu, X.: TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics. **32**, 3495–3497 (2016)
31. Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P.: ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. **24**, 2534–2536 (2008)
32. 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. Anal. Chem. **74**, 5383–5392 (2002)
33. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods. **4**, 207–214 (2007)
34. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., Cox, J.: The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods. **13**, 731–740 (2016)
35. Aebersold, R., Morrison, H.D.: Analysis of dilute peptide samples by capillary zone electrophoresis. J. Chromatogr. **516**, 79–88 (1990)
36. Britz-McKibbin, P., Chen, D.D.: Selective focusing of catecholamines and weakly acidic compounds by capillary electrophoresis using a dynamic pH junction. Anal. Chem. **72**, 1242–1252 (2000)
37. Compton, P.D., Zamdborg, L., Thomas, P.M., Kelleher, N.L.: On the scalability and requirements of whole protein mass spectrometry. Anal. Chem. **83**, 6868–6874 (2011)
38. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A. **98**, 5116–5124 (2001)
39. da Huang, W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. **4**, 44–57 (2009)