HIGHLY SENSITIVE QUALITATIVE AND QUANTITATIVE TOP-DOWN PROTEOMICS USING CAPILLARY-ZONE ELECTROPHORESIS-ELECTROSPRAY IONIZATION-TANDEM MASS SPECTROMETRY

By

Rachele Anne Lubeckyj

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry - Doctor of Philosophy

2021

ABSTRACT

HIGHLY SENSITIVE QUALITATIVE AND QUANTITATIVE TOP-DOWN PROTEOMICS USING CAPILLARY-ZONE ELECTROPHORESIS-ELECTROSPRAY IONIZATION-TANDEM MASS SPECTROMETRY

By

Rachele Anne Lubeckyj

Proteomic studies commonly utilize bottom-up proteomics due its high sensitivity, throughput, and robustness, bottom-up proteomics has issues with distinguishing proteoforms with high sequence similarity. Top-down proteomics (TDP) overcomes this issue by analyzing intact proteins and identifying proteoforms and their post translational modifications' (PTMs) with higher confidence providing opportunities to gain valuable insight into biological mechanisms. Reversed-phase liquid chromatography mass spectrometry (RPLC-MS) is the most widely used method for top-down analysis; there are still issues with sample loss facilitating a need to have micrograms of starting material. Capillary zone electrophoresis (CZE)-MS is a highly sensitive separation and detection technique that has emerged as an alternative to RPLC-MS for mass-limited samples, however applying CZE-MS for large-scale top-down proteomics has been impeded by the limited sample loading capacity and narrow separation window. This thesis will describe three projects improving CZE-ESI-MS for large-scale TDP of complex samples.

Chapter 2 and 3 focus on the improvement of single-shot CZE-MS/MS for TDP. First, the systemic evaluation of the sample stacking technique, dynamic pH junction, for the focusing of proteoforms during CZE-MS. The optimized dynamic junction-based CZE-MS/MS platform reached 1-μL sample loading volume, 90-min separation window with high peak capacity (~280)

for the identification and characterization of ~600 proteoforms from an *E. coli* proteome. The data in this work represents the largest loading capacity, separation window, peak capacity and proteomic identification of CZE for TDP of complex proteomes.

In chapter 3, the dynamic pH junction-based CZE-MS/MS using a 1.5-m separation platform achieved 180-min separation window with a 2-µL sample loading volume. This improved CZE-MS/MS platform produced high separation of myoglobin by baseline separating three proteoforms with over 100-fold concentration range and produced nearly 1,000,000 theoretical plates. The CZE-MS/MS platform also identified ~449 proteoforms from the *E. coli* sample using only 25 ng of proteins per run. Single-shot CZE-MS/MS identified over 1,500 and 2,000 proteoforms from two different regions of Zebrafish brain (cerebellum (Cb) and optic tectum (Teo)) utilizing nanograms of material. Label-free TDP of the two brain regions quantified thousands of proteoforms and revealed significant differences between the two regions.

In chapter 4, we present for the first time a highly sensitive modified sample preparation workflow for TDP using laser capture microdissected (LCM) tissue samples of Zebrafish brain tissue. This workflow utilized OG, a MS-compatible detergent, while using a freeze/thaw method for protein extraction eliminated the necessity of detergent removal resulting in a lower sample loss for mass limited samples using TDP. This modified workflow identified ~220 proteoforms of laser captured microdissected tissue sections (500-μm^2 tissue section) when <250 cells were injected, demonstrating the sensitivity of this platform for mass limited samples. This procedure facilitated quantitative top-down proteomics that produced protein expression profiles that can efficiently distinguish between different microdissected tissue sections even when the sample were isolated from the same brain region. This is the first attempt at utilizing LCM with CZE-ESI-MS/MS for highly sensitive TDP of Zebrafish brain tissues.

Copyright by
RACHELE ANNE LUBECKYJ
2021

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Liangliang Sun. I truly appreciate that he allowed me into his group right away and we started doing experiments together right off the bat allowing me to publish in my first year in grad school. Dr. Sun has helped to shape me into the scientist I am now and gave me the freedom to truly embrace my confidence. The amazing thing about Dr. Sun is that even when he is incredibly busy, he always had an open-door policy and helped you when you needed it; he is super patient and knowledgeable and I am incredibly lucky, fortunate and proud that I got to be in his group. I know that his group and his students will go far under his guidance.

I would also like to thank Dr. Xiaowen Professor Xiaowen Liu from IUPUI who developed the TopPIC program we use. He was always quick to respond to emails about issues with the TopPIC and I emailed him so many times that I think he was almost tired of seeing my name. I want to thank Mr. Billy Poulos. He takes care of the Zebrafish in Dr. Cibelli's lab and he was always there whenever I needed anything dealing with the zebrafish. He was always incredibly helpful; I even forgot my cell phone at his lab on a Friday evening and he came into work on Saturday to let me into the building so I could grab it. I always loved talking to him about grad school and he was a joy to be around.

Next, I would like to thank all my group members. Xiaojing Shen would also give me helpful suggestion on my experiments. I remember one time I was having an issue with the autosampler and I was telling her about, she stopped what she was doing and helped me fix the autosampler so I could go on with my experiment. I absolutely loved talking to her and was thankful that her

desk was next to mine so I could do it when I wanted. Daoyang Chen was also very helpful whenever I needed it on my experiments, and I knew I could always go to him whenever I needed help. Eli McCool always put up with me whenever I bothered him in the office, which I appreciate. Finally, I appreciate that they all welcomed me and even showed me many Chinese dishes and going to the Chinese New Year Festival in the international center!

Finally, I am lucky to have the support and love of my wonderful family. My parents always have always pushed for academic achievement and I wouldn't be here if it weren't for their encouragement. I know if it wasn't for the love and support of my entire family, this accomplishment wouldn't have been possible. I missed many opportunities of shared time with my parents and siblings, but their frequent encouragement helped me more than they could imagine, and I really appreciate their patience. Lastly, I want to thank my husband, Mason Ley. He was there during this entire achievement; listening to me complain and most importantly reassuring me that I could do this. I couldn't imagine what my life would have been like if you weren't here for this. I love you more than anything.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

KEY TO ABBREVIATIONS ..... **xvi**

CHAPTER 1. Introduction **1**
1.1 Proteomics 1
1.1.1 Introduction to Proteomics 1
1.1.2 Bottom-Up Proteomics (BUP) 3
1.1.3 Top-Down Proteomics (TDP) 5
1.2. Analytical Approaches for the Identification of Proteoforms 7
1.2.1 Mass Spectrometry 7
1.2.2 Tandem Mass Spectrometry 10
1.2.3. Capillary Zone Electrophoresis-Mass Spectrometry for Top-Down Proteomics 12
1.2.3.1 Electrospray Ionization Interface for the Coupling of Capillary Zone Electrophoresis to Mass Spectrometry 16
1.2.3.2 Preconcentration Methods to Increase Sample Loading Capacity 18
1.2.3.3 Capillary Coatings to Increase Separation Window for CZE 20
1.3. Quantitative Top-Down Proteomics 21
1.3.1 Label-Free Quantitation 21
1.3.2 Isotopic Labeling Quantification 23
1.3.2.1 Stable Isotope Labeling of Amino Acids in Cell Culture 23
1.3.2.2 Tandem Mass Tags 24
1.4. Current Challenges to Top-Down Proteomics 25
1.4.1 Separation 25
1.4.2 Fragmentation 26
1.4.3 Bioinformatics for Proteoform Identification 27
1.5. Summary 29
REFERENCES **31**

CHAPTER 2. Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 Escherichia coli Proteoforms 41
2.1 Introduction 42
2.2 Experimental 44
2.2.1 Materials and Reagents 44
2.2.2 Sample Preparation 44
2.2.3 CZE-ESI-MS/MS 45
2.2.4 Measurement of electroosmotic mobility 47
2.2.5 Data Analysis 48
2.3 Results and Discussion 49
2.3.1 Comparison of Dynamic pH Junction and FESS Methods 50
2.3.2 Optimization of the Dynamic pH Junction-Based CZE-MS 54
2.3.3 Single-Shot Top-down Proteomics with CZE-MS/MS 58
2.4 Conclusions 62
2.5 Acknowledgements 62
APPENDIX 64
REFERENCES 69

CHAPTER 3: Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples 75
3.1 Introduction 75
3.2 Experimental 77
3.2.1 Materials and Reagents 77
3.2.2 Sample Preparation 78
3.2.3 CZE-ESI-MS/MS 80
3.2.4 Data Analysis 81
3.3 Results and Discussion 83
3.3.1 Evaluation of the CZE-MS Platform with a 1.5-m Long Separation Capillary Using a Standard Protein Mixture 83
3.3.2 Top-Down Proteomics of *E. coli* Cells Using Single-Shot CZE-MS/MS with a 1.5-m Long Separation Capillary 87
3.3.3 Quantitative Top-Down Proteomics of Zebrafish Brain Cb and Teo Regions Using CZE-MS/MS with a 1.5-m Long Separation Capillary 92
3.4 Conclusions 96
3.5 Acknowledgements 97
APPENDIX 98
REFERENCES 101

CHAPTER 4. Development of a Highly Sensitive Top-Down Proteomic Workflow Using Capillary-Zone Electrophoresis-Electrospray Ionization Tandem Mass Spectrometry for Spatially Resolved Proteomics of Zebrafish Brain 106
4.1 Introduction 106
4.2 Experimental 109
4.2.1 Material and Reagents 109
4.2.2 Tissue Preparation 110
4.2.3 Laser Capture Microdissection.....110
4.2.4 Sample Preparation for CZE-MS/MS 110
4.2.5 CZE-ESI-MS/MS 113
4.2.6 Data Analysis.....114
4.3 Results and Discussion 115
4.3.1 Effectiveness of OG for Qualitative Top-Down Proteomics using Single-Shot CZE-ESI-MS/MS of Laser Capture Microdissection Tissue Samples 115
4.3.2 Spatially Resolved Quantitative Top-Down Proteomic Analysis of Microdissected Zebrafish Brain Tissue 122
4.4 Conclusions 128
4.5 Acknowledgements 129
REFERENCES 130

CHAPTER 5. The Future of Capillary-Zone Electrophoresis-Electrospray Ionization Tandem Mass Spectrometry for Top-Down Proteomics.....135

LIST OF TABLES

Table 2.1. Reproducibility data from the 11 consecutive CZE-MS runs using the standard protein mixture.....65

Table 4.1. Upregulated proteins of the microdissected tissue samples from the Teo brain region. Fold change values are log2 scale.....126

Table 4.2. Upregulated proteins of the microdissected tissue samples from the Tel brain region. Fold change values are log2 scale.....127

LIST OF FIGURES

Figure 1.1. The biological sources that cause alterations resulting in different proteoforms originating from a single gene. Reprinted with permission from reference [9]. Copyright (2021) John Wiley and Sons. 2

Figure 1.2. Divergent workflows of top-down and bottom-up proteomics. **(A**) Typical workflow of top-down proteomics. (**B**) Typical workflow of bottom-up proteomics 4

Figure 1.3. Mass spectrum illustrating the different charge states for (**A**) peptide using bottom up proteomics and (**B**) proteoform using top-down proteomics.....7

Figure 1.4. Electrospray Ionization. (**A**) The mechanism of electrospray ionization when operated in positive mode. (**B**) Desorption of a gaseous unfolded protein ion. Reprinted with permission from reference [28]. Copyright (2021). American Chemical Society.....9

Figure 1.5. Nomenclature for fragmentation of proteins**.** 11

Figure 1.6. Diagram of the CZE separation theory. An open tubular capillary is placed between two buffer vials, through which a high voltage is applied. This voltage causes analytes to migrate from the injection end to the detector according to their electrophoretic mobility. The EOF that results from the electrical double layers drives the separation of analytes.....14

Figure 1.7. (A) The design of the electrokinetically pumped sheath flow CE-MS interface. (**B**) Three different generations of the CZE-MS interface illustrating that a larger emitter opening will allow for a shorter distance between the capillary end and the emitter opening, increasing the sensitivity and robustness. Reprinted with permission from reference [68]. Copyright (2021) American Chemical Society. 17

Figure 2.1. Protein intensity change vs various sample injection volumes (50-, 100-, 200-, and 500-nL) for (**A**) control, (**B**) FESS, and (**C**) dynamic pH junction. The error bars represent the standard deviations of protein intensity from triplicate CZE-MS analyses. (**D**) EIEs of the mixture of standard proteins from CZE-MS under the three different conditions: top panel is the control, middle panel is FESS, and bottom panel is dynamic pH junction. The proteins labeled in the electropherograms are (a) lysozyme, (b) cyto.c, (c) myoglobin, (d) CA, and (e) β-casein. 51

Figure 2.2. EIEs of the standard protein mixture dissolved in 50 mM NH_4HCO_3 (pH 8.0) analyzed by the dynamic pH junction-based CZE-MS with (**A**) 500-nL sample injection and (**B**) 1-μL sample injection. The theoretical plate value (N) of each protein was calculated based on the peak width and migration time of each protein in the EIEs. BSA was not extracted in the figures due to its low signal-to-noise ratio. 56

Figure 2.3. Top-down proteomics of *E. coli* using CZE-MS/MS. (**A**) Sample loading volume vs the number of proteoform IDs and the number of proteoform-spectrum matches (PrSMs). (**B**) Electropherograms of the *E. coli* protein sample analyzed by top-down based CZE-MS/MS in

duplicate runs. For the CZE-MS experiments, 20 kV was applied at the injection end for separation. (**C**) The zoom-in electropherogram of the *E. coli* protein sample showing the separation window from the 1st run CZE-MS/MS in (**B**).....58

Figure 2.4. (**A**) Distribution number of identified proteoforms from each *E. coli* gene. (**B**) Detected mass error distribution from the identified proteoforms. (**C and D**) Sequence and fragmentation pattern of two identified proteins. Carbamidomethylation sites on cysteines are shown in red. The single-shot CZE-MS/MS data in figure 3 B was used for these analyses.....61

Figure 2.5. Calibration curve of the protein concentration and protein intensity for lysozyme, CA and myoglobin. The errors bars are the standard deviations of protein intensity from duplicate CZE-MS runs. A LTQ-XL mass spectrometer was used for the experiments. 66

Figure 2.6. Protein intensity from dynamic pH junction based CZE-MS utilizing various BGEs. BGEs. (**A**) Different formic acid concentrations (0.1-0.5% (v/v)). (**B**) Different acetic acid concentrations (5% and 10% (v/v)). The protein intensity from formic acid BGEs were normalized to 0.1% (v/v) formic acid; the intensity from acetic acid BGEs were normalized to 5% (v/v) acetic acid. (**C**) The acid concentrations that produced the highest protein intensities (0.1% (v/v) formic acid and 5% (v/v) acetic acid) were compared. LTQ-XL mass spectrometer was used for the experiments while the errors bars represent the standard deviations of protein intensity from duplicate CZE-MS runs.....67

Figure 2.7. Distribution of the identified proteoform mass from single-shot CZE-MS/MS 68

Figure 3.1. Comparison of the 1-m and 1.5-m separation capillary for CZE-MS analyses of a standard protein mixture. (**a**) Electropherograms of the standard protein mixture using CZE-MS with a 1.5-m capillary (top panel) and a 1-m capillary (bottom panel). (**b**) Electropherograms of the standard protein mixture using the 1.5-m separation capillary CZE-MS analyses with 1.5-m LPA-coated separation capillary in sextuplicate. 84

Figure 3.2. CZE-MS analyses of a standard protein mixture using different sample loading volumes (0.5-μL, 1-μL, and 2-μL). A 1.5-m LPA-coated separation capillary was used for all the CZE-MS runs in duplicate. (**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.....86

Figure 3.3. CZE-MS analyses of the *E. coli* proteome using a 1.5-m-long LPA-coated separation capillary. One microgram of proteins was injected per triplicate CZE-MS/MS run. (**a**) Base peak electropherograms of the triplicate CZE-MS/MS runs. (**b**) A zoomed-in electropherogram of one CZE-MS/MS run showing the separation window. (**c**) Protein and proteoform identifications of the triplicate CZE-MS/MS runs. 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. Carbamidomethylation modification are marked in red on the cysteine (C) residues. DnaK has one acetylation modification on either the G or K residue. 89

Figure 3.4. CZE-MS/MS analyses using the 1.5-m LPA-coated separation capillary of mass-limited *E. coli* proteome samples. (**a**) Base peak electropherograms of the *E. coli* proteome sample with 1-μg (top panel) and 100-ng proteins (bottom panel) as the starting materials. (**b**) Proteoform and protein identifications from the analyses of 1-μg and 100-ng *E. coli* samples. The error bars are the standard deviations 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 using the 1-μg *E. coli* sample. (**d**) Mass distribution of the identified proteoforms from one CZE-MS/MS analysis using the 1-μg *E. coli* sample. 91

Figure 3.5. CZE-MS/MS analyses using the 1.5-m LPA-based separation capillary of zebrafish brain Cb and Teo regions. (**a**) An illustration of a mature zebrafish brain. (**b**) Base peak electropherograms of zebrafish brain Cb (top panel) and Teo (bottom panel) after CZE-MS/MS analyses. (**c**) Protein and proteoform identifications from the Cb and Teo samples. The error bars represent the standard deviations from triplicate CZE-MS/MS runs. (**d**) Volcano plot of the quantified proteoforms. The proteoforms with higher abundance in Cb are marked in blue and the proteoforms with higher abundance in Teo are marked in red.. 93

Figure 3.6. (**a**) Proteoform abundance dynamic range from single-shot CZE-MS/MS analysis using the 1.5-m LPA-coated separation capillary of Cb and Teo samples. The proteoform feature intensity was used to approximation the dynamic range. (**b,c**) The sequence and fragmentation pattern of two proteoforms of parvalbumin 4. 96

Figure 3.7. Fragmentation and sequence pattern of superoxide dismutase using (UniProt ID: P00442). A N-terminal acetylation was identified and is labeled on the alanine (A) residue. Highlighted in grey is the carbamidomethylation on two cysteine residues. 99

Figure 3.8. (**a**) Base peak electropherogram of the tryptic digest of myoglobin analyzed by bottom-up based CZE-MS/MS. (**b**) Myoglobin peptide with a N-terminal acetylation with the annotated MS/MS spectrum using bottom-up proteomics. (**c**) A phosphopeptide myoglobin with the annotated MS/MS spectrum using bottom up proteomics. Confirmation of the proteoforms that were detected using top-down based CZE-MS with a 1.5-m LPA-coated separation capillary. 100

Figure 4.1. Schematic overview of the LCM workflow. Two workflows were used: the standard top-down proteomic workflow and a modified top-down proteomic workflow. The modified workflow used two different extraction methods: ultrasonication and freeze/thaw. 112

Figure 4.2. (**A**) A 20-μm-thick Zebrafish brain slice used in the study. Three separate regions of the Zebrafish brain, cerebellum (Cb), Optic Tectum (Teo), and Telencephalon (Tel), were microdissected with a spatial resolution of 500 μm^2. (**B**) The microscopic images of the three-

square regions of brain tissue regions after the microdissection. (**C**) Corresponding microdissected tissue section on the LCM cap. 117

Figure 4.3. Comparison between proteoforms and proteins identified from the LCM cap using the ten tissue samples originating from the same Zebrafish brain slice utilizing OG detergent for protein extraction with either (**A**) ultra-sonication and (**B**) freeze/thaw extraction method. The error bars are the standard deviations from duplicate CZE-ESI-MS/MS runs. 119

Figure 4.4. CZE-MS analyses of the LCM samples using the freeze/thaw method using a 75-cm-long and LPA-coated separation capillary. The CZE-MS/MS analyses were performed in triplicate. (**A**) Base peak electropherograms of the triplicate CZE-MS/MS runs of the Tel2 microdissected tissue sample. (**B**) Mass distribution of the identified proteoforms from one CZE-MS/MS analysis from each of the LCM samples using the freeze/thaw method. (**C**) Distribution of matched fragment ions of one CZE-MS/MS run. (**D**) Sequence and fragmentation pattern of protein Calmodium. There is one acetylation and a mass shift of +42 Da. 120

Figure 4.5. Principle component analysis of proteoform expression of the (**A**) the two different microdissected tissue samples from Teo brain region and (**C**) the two different microdissected tissue samples from Tel brain region. Pairwise correlation matrix using the log2-transformed LFQ intensities for (**B**) the two different microdissected tissue samples from Teo brain region and (**D**) the two different microdissected tissue samples from Tel. The numbers in the corner correspond to Pearson's correlation brain region. 123

Figure 4.6. Z-score hierarchical clustering based on squared Euclidean distance measure. Each row represents one proteoform and each column represents one sample. The color scale means the proteoform expression standard deviations from the mean, with blue for low expression and red for the high expression levels. (**A**) Hierarchical clustering analysis (HCA) of the two tissue samples and their technical replicates microdissected from the Teo brain region, and (**B**) HCA of the two tissue samples and their technical replicates microdissected from the Tel brain region. 125

KEY TO ABBREVIATIONS

AI-ETD	Activated ion on electron transfer dissociation
ABC	Ammonium bicarbonate
BUP	Bottom-up proteomics
CZE	Capillary zone electrophoresis
Da	Daltons
DDA	Data-dependent acquisition
E. coli	*Escherichia coli*
ETD	Electron transfer dissociation
EOF	Electroosmotic flow
ESI	Electrospray ionization
FESS	Field enhanced sample stacking
HCD	Higher-energy collision dissociation
ID	Identification
LC	Liquid chromatography
LFQ	Label-free quantification
m/z	Mass-to-charge ratio
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
PTM	Post-translational modification
PrSM	Proteoform spectral match

RPLC	Reversed-phase liquid chromatography
SEC	Size exclusion chromatography
SILAC	Stable Isotope Labeling of Amino Acids in Cell Culture
SPE	Solid-phase extraction
TDP	Top-down proteomics
TMT	Tandem mass tags

CHAPTER 1: Introduction

1.1 Proteomics

1.1.1 Introduction to Proteomics

Proteomics aims to identify and quantify the entire set of proteins (proteome) produced in an organism (i.e. cell line or tissue)[1-7]. Proteomics has gone through tremendous progress over the years and now encompasses not just the proteins in any given cell, but also isoforms and their post-translational modifications (PTMs), their structural information (such as their higher order complexes), and the interactions between them[6,7]. Proteins are responsible for almost all the biological processes in cells, because PTMs can influence proteins' functions and the majority of proteins form complexes to function in the cell, therefore it is extremely important to study the proteomes and their dynamics under various biological conditions to bridge genotypes and phenotypes of different organisms[6,7]. Proteomic studies are typically carried out in two fashions: bottom-up proteomics and top-down proteomics. In this section, relevant terminology will be discussed, as well as the advantages and disadvantages between the two methods that are used to carry out proteomic analysis.

Due to the evolution of proteomic terminology over the years, a brief section will explain the language that will be used in this thesis. For many years, there has been no specific term for the arrangement of amino acids and the PTMs that interact with them[8]. The terms: proteins, protein forms and protein isoforms were the terms that were commonly used to describe a proteome within a cell. Due to the advancement of top-down proteomics over the last 20 years, researchers

realized that these terms are not precise enough to describe proteomics within a cell (**figure 1.1**)[8-13].

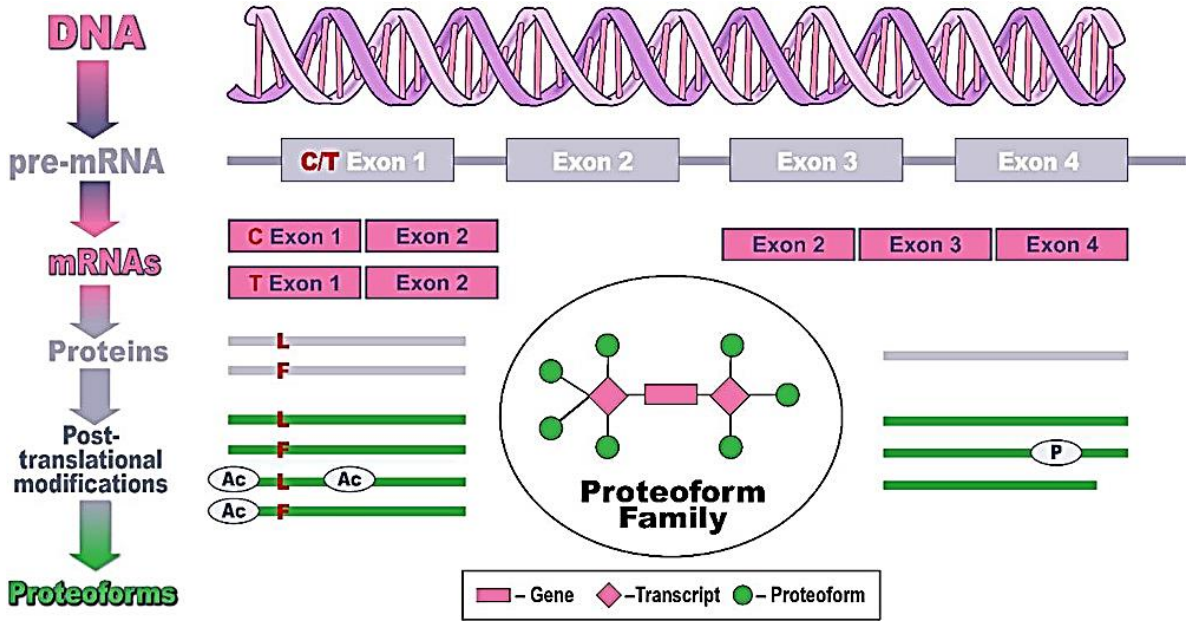


Figure 1.1. The biological sources that cause alterations resulting in different proteoforms originating from a single gene. **Reprinted with permission from reference [9]. Copyright (2021) John Wiley and Sons.**

The term *protein* is a general word that is used to describe a linear set of amino acids held together by peptide bonds. In many cases, one protein per gene convention is often used[12]. However, this is a very crude estimate; genes are transcribed into pre-mRNA strands, which are then translated into a unique protein. Multiple events can occur before the transcription process (pre-mRNA strand) such as introduction or combination of introns into the amino acid sequence, single polynucleotide polymorphism (SNPs), and/or endogenous proteolysis. In addition, after the transcription process (post-mRNA strand) modifications can also be added, such as 5' capping, poly A tail addition, and alternative splicing[8,9]. These events are numerous and can create unique protein molecules that can have very different roles that occur within an organism.

When considering the various events that can occur on the mRNA strand arising from a single gene, the term that is used is called *isoform* (**figure 1.1**) [8,9].

Isoform is a term that describes pre- and post-transcriptional regulation of a protein sequence. However, until recently there has been no term that also relates the PTMs that can occur on the protein sequences after translation[8-10,12]. The top-down community realized this lack of specificity of proteomic terminology and coined the term *proteoforms* (**figure 1.1**), which was led by the Kelleher group[8]. *Proteoform* is an all-encompassing term for a defined amino acid sequence (i.e. isoform) that also includes the localized PTMs[10,13]. When discussing all the proteoforms that come from a single gene is called a *proteoform family* (**figure 1.1**)[9,10,14]. A *proteoform family* contains all the proteoforms, therefore protein products from all pre- and post-mRNA events and their PTMs arising from one gene[9,10,14].

1.1.2 Bottom-Up Proteomics (BUP)

Bottom-up proteomics (BUP) is the most widely used and mature approach for protein identification and quantification. A typical workflow (**figure 1.2B**) for bottom up proteomics consists of multiple steps: (1) extraction of the protein mixture from a biological sample, (2) digestion by enzymes (usually the serine protease trypsin) of the proteins into peptides after reduction and alkylation, (3) fractionation by liquid chromatography (LC), and (4) electrospray ionization (ESI)-mass spectrometry (MS) and tandem mass spectrometry (MS/MS) analysis of the fractionated peptides, and (5) protein identification by a database search[15-17]. Protein identification is then done by a process know as ‘protein inference’, where the occurrence of a specific protein within the sample is ‘inferred’ from the peptides IDs it contains.

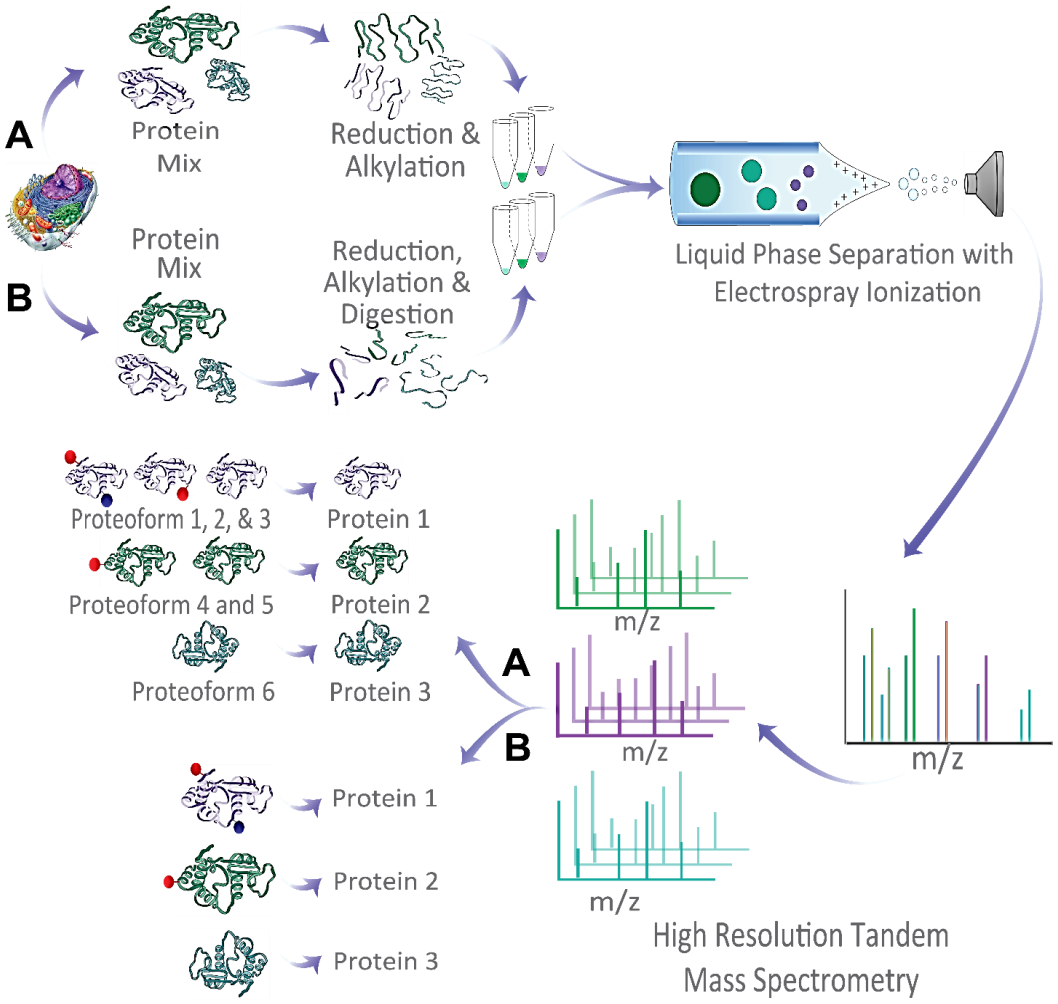


Figure 1.2. Divergent workflows of top-down and bottom-up proteomics. (A) Typical workflow of top-down proteomics. (B) Typical workflow of bottom-up proteomics

BUP exploits the advantages that peptides have over proteins during analysis. When the proteoforms are digested into peptides these large biomolecules become smaller with a more uniform molecular weight (**figure 1.3A**). For instance, trypsin can specifically cleave proteins into peptides with a molecular weight range of 600-1000 Da. Therefore, peptides are effortlessly separated by reverse-phase liquid chromatography (RPLC) causing a reduction in the coelution of peptides and allowing for large-scale characterization of proteins from complex mixtures. Peptides can also ionize easier through electrospray ionization (ESI) and produce a more predictable fragmentation pattern during MS//MS[15]. Due to these features, the mass spectra are

quite simple and eliminate the 'charge state dilution' and 'isotope dilution' effects that are common when analyzing proteoforms (**figure 1.3**)[4,22]. BUP typically utilize data dependent acquisition (DDA) workflows, where after the acquisition of one mass spectrum, the top N (N being the number of peptides that are selected) most intense peptide ions in the mass spectrum are isolated for fragmentation in sequence to produce tandem mass spectra[15-17]. These workflows are quite powerful and have the capability of achieving deep proteome coverages. For instance, Kulak *et al.* used nano-RPLC-ESI-MS and MS/MS in the DDA mode for the analyses of 12 different human cell lines[18]. This approach identified, on average, over 11,000 proteins from those cell lines using sub-microgram amounts of protein materials; this method was able to reveal differences between the cell lines arising from their developmental origin[18].

Even with results such as these, there are still challenges associated with BUP. The sequence coverage generated from BUP can be limiting; for large-scale bottom-up proteomics studies, the median sequence coverage is roughly 30%[15]. There is also the protein inference problem which arises from inferring protein identifications from a limited number of tryptic peptides which can cause a narrow and biased viewpoint of the proteome[15]. BUP uses a canonical protein database that can include multiple isoforms from a single gene creating higher sequence redundancy[9]. The number of identified peptides per protein is limited and a big portion of the identified proteins usually only have one or two matched peptides[15]. This issue makes it difficult to delineate proteoforms that have high sequence similarity [9,15].

1.1.3 Top-Down Proteomics

Top-down proteomics (TDP) has gained great attention in recently years for measuring proteoforms directly at the global scale. With the onset of technological advancement in sample preparation, liquid-phase separation, mass spectrometry and bioinformatics, TDP has enabled

high throughput analysis of proteoforms from complex proteomes[92-99]. TDP follows a more simplified workflow compared to BUP (**Figure 1.2A**); proteins do not need to be digested into peptides (no endoproteinase digestion) and has the option to be chemically modified (i.e. reduction and alkylation) resulting in a lower occurrence of experimental artifacts[19-22]. Compared to BUP, TDP can identify and quantify distinct proteoforms that can be lost by endoproteinase digestion[19-22].

Two main issues have impeded top-down proteomics; the first is its sensitivity issues and the second is the need for high-capacity and high-throughput liquid-phase separation of proteoforms in complex mixtures. The sensitivity issue arises from 'charge state dilution' and 'isotope dilution' effects[22]. During the ESI process, a single protein will acquire multiple charges causing the protein's total signal to be divided over all the different charge states (i.e., charge state dilution)[4]. The larger the protein, the more signal distribution will occur across multiple channels (**figure 1.3B**). Additionally, protein's large molecular weight causes a broad isotopic distribution (**figure 1.3B**) that lowers the S/N ratio (i.e., isotope dilution)[22]. The term proteoforms can capture the various sources of biological variations, during and after protein synthesis, that can change the composition at the protein level[9,19-22]. Considering all these variations (i.e., PTMs, SNP, alternative splicing, etc.) on the protein level, the estimates of the amount of proteoforms within the human proteome leads to an estimated size of 1,000,000 distinct proteoforms within a certain cell type[23]. Reflecting on this number, there is a need for a high-capacity liquid-phase separation of proteoforms from complex mixtures before mass spectrometry analysis[8-10,19-23].

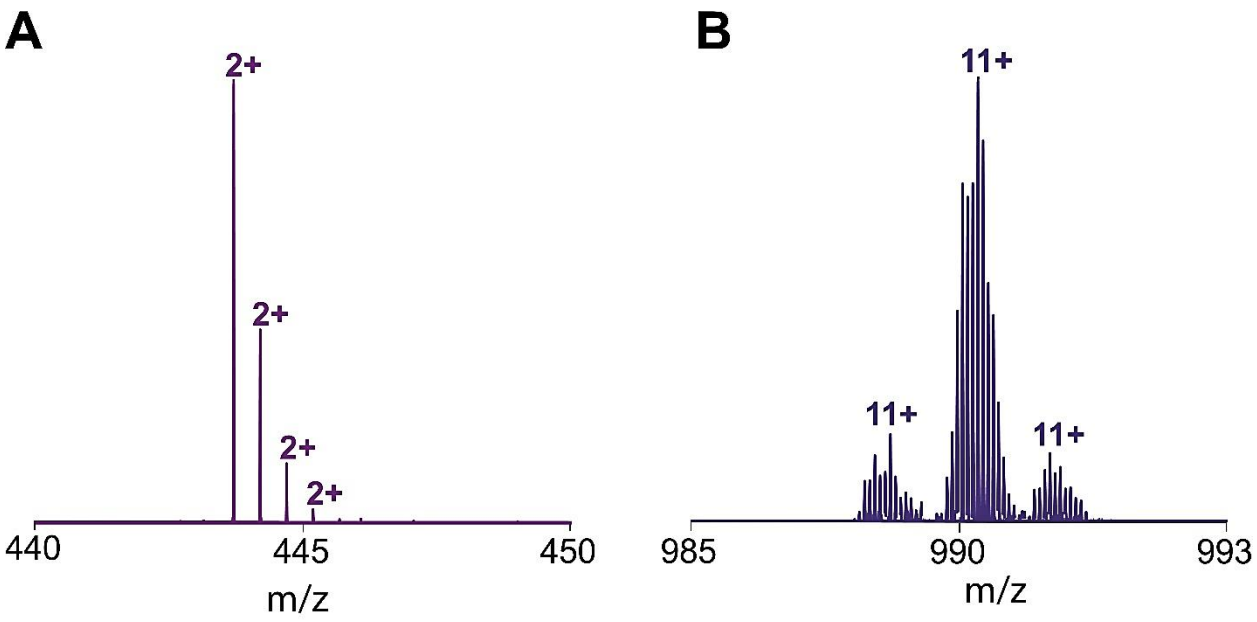


Figure 1.3. Mass spectrum illustrating the different charge states for (**A**) a peptide using bottom up proteomics and (**B**) a proteoform using top-down proteomics

1.2. Analytical Approaches for the Identification of Proteoforms

1.2.1 Mass Spectrometry

Mass spectrometry (MS) has been the preferred technique of choice for the analysis of complex protein mixtures, because MS allows for the detection and quantification of proteins with high speed, accuracy, and sensitivity[24-26]. A mass spectrometer is an analytical technique that separates and measures the mass to charge (*m/z*) ratio of gas phase ions. Mass spectrometers consist of an ion source that converts ions from the liquid phase to the gas phase, a mass analyzer(s) that measures the *m/z* ratio of the gas phase ions, and a detector that will record the sum of ions at specific *m/z* values[25,26]. Accurate *m/z* measurements can provide both the possible molecular formula and monoisotopic masses of the analytes, which can aid in the annotation and discovery of proteoforms[24-26]. Since MS-based proteomics is typically coupled to liquid phase separation techniques, an ionization method that can convert nonvolatile proteoform ions from

the liquid phase to the gas phase without degradation is crucial. ESI is the most used soft ionization method for proteomic studies.

For the analysis of large biomolecules, electrospray ionization (ESI) has been the gold standard for an ionization source. ESI is a soft ionization technique, providing little ion excitation to produce no fragmentation of the analyte. ESI also imparts positive charges onto the analyte, generating multiply charged species, therefore allowing the analysis of large molecular weight species. ESI has allowed for the ability to study large biomolecules, such as proteins, because typically mass spectrometers can only detect m/z values less than 3000 Da.

ESI goes through three main steps: (1) droplet formation, (2) droplet shrinkage, and (3) desorption of gaseous ions (**figure 1.4A**)[28]. During droplet formation, cations and anions arising from the sheath buffer will migrate to the outer emitter tip. Electrostatic forces will cause the cations and anions to flow away from each creating a Taylor cone[28]. Combination of coulombic repulsion of ions and surface tension will cause charged droplets (μm size) to break away from the tip[28]. Next, droplet shrinkage will occur due to solvent evaporation leading to increasingly smaller droplets until repulsive coulombic forces will exceed the surface tension causing fission of the main droplet into smaller highly charged doplets[28]. Lastly, desorption of gaseous ions is the generation of gas phase positive ions, and different analytes follow different models. For proteins, it follows a model called Chain Ejection Model (**figure 1.4B**)[28]. Proteins contain both hydrophobic and hydrophilic amino acids. Since the hydrophobic amino acid chain isn't energetically favorable with the solvent, the chain will migrate to the surface to minimize solvent interaction[28]. Once the chain has migrated to the surface, the chain is sequentially ejected from the droplet starting with one end of the termini[28].

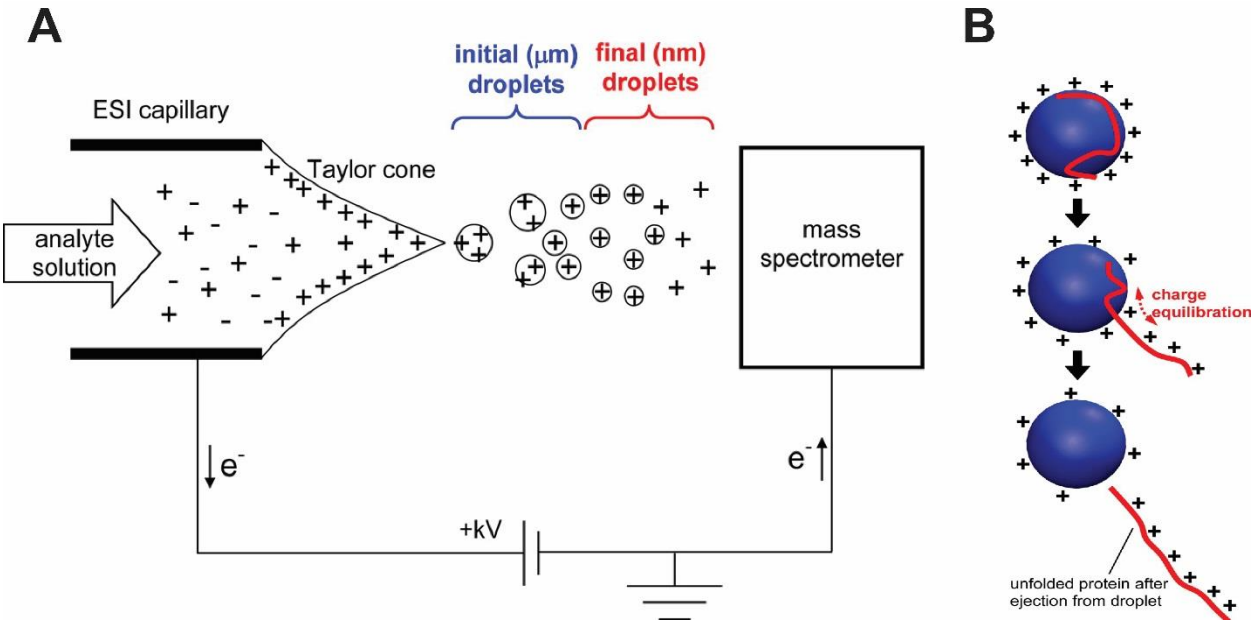


Figure 1.4. Electrospray Ionization. (**A**) The mechanism of electrospray ionization when operated in positive mode. (**B**) Desorption of a gaseous unfolded protein ion. **Reprinted with permission from reference [28]. Copyright (2021). American Chemical Society.**

After ionization, the proteins will be measured within the mass analyzer of the mass spectrometer. The orbitrap mass analyzer was introduced ~15 years ago by Makarov[101] and commercial by Thermo Fisher in 2005 in a hybrid instrument[102]. Hybrid orbitrap instruments have gain popularity within the proteomic field due to the high resolution (1,000,000 at m/z 200) and mass accuracy (sub-1 ppm)[102]. The orbitrap is a small spindle-shaped electrostatic device which packets of ions are tangentially injected onto. The ion packets will orbit around the central electrode using three different types of motion: rotational, radial and axial. The ion's axial motions induce an image current in which the ion's signal will be Fourier transformed to yield the high-resolution mass spectra.

1.2.2 Tandem Mass Spectrometry

Tandem mass spectrometry (MS/MS) allows for multiple analysis within a mass analyzer with fragmentation between the mass analyzers. In the two stage MS/MS, the first mass analyzer

(most often a filtering quadruple) will select product ions of a specific *m/z* for fragmentation within a HCD collision cell, then the second mass analyzer (for example, a high resolution orbitrap) will scan through an array of *m/z* values to determine the masses of the fragment ions.

Identification of proteoforms by tandem mass spectrometry is enabled by assigning a monoisotopic mass and charge state to a specific fragment ion. To facilitate this process, fragment ions need to be obtained using data-dependent higher-energy collisional dissociation (HCD) that subjects the top N (N being the number of precursor ions selected for fragmentation at a specific time period) ions to collisional activation with a target gas[29,30]. HCD is a collisional induced dissociation-like technique; this fragmentation technique takes places in an external octupole collision cell that utilizes a higher radiofrequency enabling the capture of low mass m/z ions that is adjacent to the C-trap[29]. HCD works by colliding a high translational energy ion (<100 eV) with an inert gas (such as argon or nitrogen) causing a portion of the ion's translational energy to be transformed into internal vibrational energy. This vibrational energy will be quickly transferred to the rest of the ion causing an increased rate of unimolecular ion dissociation; this dissociation will form fragment ions with masses and isotopomer envelopes that will aid in the identification of proteins.

To identify different types of protein fragment ions, systemic fragmentation nomenclature was developed by Roepstorff et al. to categorize these fragment ions[31]. Two major representations have been established: (1) alphabet letters, and (2) numbers. The alphabet letters describe the fragment ions based on where the bond was cleaved on the peptide backbone while the numbers illustrate the location of the cleavage position in relation to the C- or N- terminus **(Figure 1.5**). Generally, when a fragment ion contains the N-terminus of the peptide it is denoted as either an *a*, *b,* and *c*-ion while the fragment ions that contain the C-terminus are designated as a *x*-, *y*-, and

z-ion. Corresponding fragment ion pairs are labelled as *a/x*, *b/y*, and *c/z* ions and these complementary pairs are generated by the cleavage of C-C, and C-N and N-C bonds, respectively[29,32,33]. During fragmentation, there can also be internal ions that are formed due to the cleavage of multiple bonds within the peptide backbone or from loss of the N- or C-terminus resulting from secondary fragmentation of the primary fragment ions[32,33]. In addition to the internal ion fragmentation, Immonium ions can also be lost which are single amino acid residues of low mass [32,33]. For HCD fragmentation, studies have shown that there is extensive clusters of y-ions and shorter, less frequent b-ions present in HCD spectra of proteins[32,33]. Furthermore, internal ions, side-chain fragments and immonium ions are also common throughout the low-mass range of the spectra[32,33].

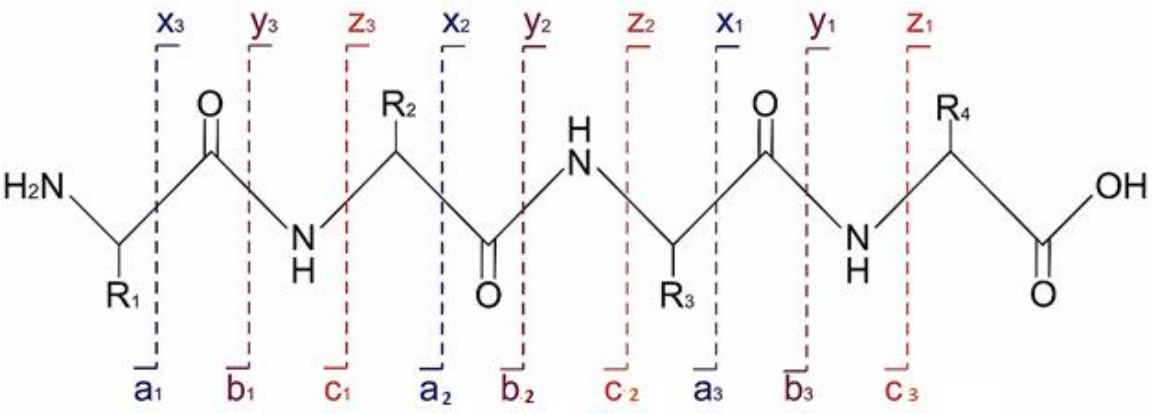


Figure 1.5. Nomenclature for fragmentation of proteins.

1.2.3 Capillary Zone Electrophoresis-Mass Spectrometry for Top-Down Proteomics

The size of the proteome is expansive; it has a very high protein concentration dynamic range, somewhere around 6 order of magnitude in concentration within cells[23]. Right now, researchers believe that there could be more than 1 million proteoforms in the human body[23]. To fully study,

proteoforms from complex mixtures, there is a need for high-capacity separation to reduce sample complexity. Ways we can do this is by utilizing the differences in proteoforms, such as size by using size exclusion chromatography (SEC) or by hydrophobicity using RPLC[34-37]. The typical separation technique for top-down proteomics is RPLC[92-99]. A single run using ultrahigh-pressure long column RPLC-MS/MS was able to identify ~900 proteoforms with high confidence (1% FDR) using only a few μg of simple microbial lysate. However, weaknesses have come apparent when applying this technique to mass-limited proteome samples using top-down proteomics First, sample loss is significant due to the sample injection valves and the analyte's high affinity for the stationary and mobile phase resulting in peak broadening and poor peak capacity. Second, there is sample bias for RPLC due to the loss of hydrophilic proteoforms. Third, high capacity separation using LC techniques require fraction collection in the μg/mL range for high identification rate. However, to analyze mass limited top-down proteomic samples, we need a technique that has a highly sensitive separation and a highly sensitive detection method.

CZE, an on-line liquid-phase separation that can be coupled to mass spectrometry, is one method that can improve the sensitivity for mass-limited samples. CZE-MS has been well-documented as an important platform for the characterization of proteoforms, due to its high separation efficiency and sensitive detection of proteins[39-48]. Valaskovic et al. achieved attomole level sensitivity using CZE-MS for the characterization of carbonic anhydrase (28,780.4 Da) from human blood[39]. In 2013, Sun et al. baseline separated four model proteins and their impurities using CZE-MS with limits of detection (LOD) ranging from 20-800 amol[44]. Zhao et al. identified over 500 proteoforms corresponding to 180 proteins from a fractionated yeast proteome using CZE-MS/MS. Then in 2016, Bush et al. identified 138 proteoforms of recombinant human

interferon β-1 and quantified 55 proteoforms using CZE-MS/MS; triantennary isomers of this interferon were separated and identified[46].

CZE separates based on an analyte's mass to charge ratio under the influence of an electric field in a buffer-filled separation capillary (**figure 1.6**)[49-51]. CZE separations are performed in 10- to 100-cm long fused silica open tubular capillaries that have an inner diameter of 10- to 100-µm under an electric potential of 10-30 kV[6-8]. Migration time (analyte's velocity within the capillary) depends on two properties: (1) the analytes electrophoretic mobility under the electric field, and (2) the electroosmotic flow, also called EOF (the solvent flow due to the electric double layer at the capillary surface)[49-51]. Due to these properties, when applying an electric field across the capillary cations will migrate towards the cathode and the anions will migrate towards the anode, while neutral molecules that have no charge will produce no movement. This technique is an innovative approach for proteomic analysis, because CZE produces better separation efficiency than liquid chromatographic separations and can be explained using the van-deemter equation[52,53].

$$H = A + B/\mu + C\mu$$

This analytical equation relates plate height (H) to the various thermodynamic, physical and kinetic parameters that can cause peak broadening on a chromatographic column. The peak broadening terms include: A the eddy diffusion term that describes the analytes path through the stationary phase, B which is the longitudinal diffusion term that explains how the analyte moves in the longitudinal directions due to the differences in the analyte concentration within the mobile phase, the C term is the resistance to mass transfer that illustrates how the analyte interacts with both the stationary and mobile phase, and lastly the µ term is linear velocity. Applying the Van deemter equation to a CZE separation would cause the A and C terms to drop

out due to the lack of a stationary phase in this technique, leaving only the B term. Now applying this to proteome samples, proteoforms have a low diffusion coefficient meaning that the effect of the B term is small demonstrating a high separation efficiency. In 2014, Han et. al. compared CZE-MS and RPLC-MS in the analysis of the Dam1 complex[41]. CZE -MS was able to baseline separate and detect nine subunits of this complex with similar signal-to-noise ratios as RPLC-MS, but with 100-fold less sample consumption (2.5 ng vs. 250 ng) illustrating the sensitivity of this technique[41]. Consequently, CZE is a promising separation technique for the highly sensitive separation and detection of proteoforms.

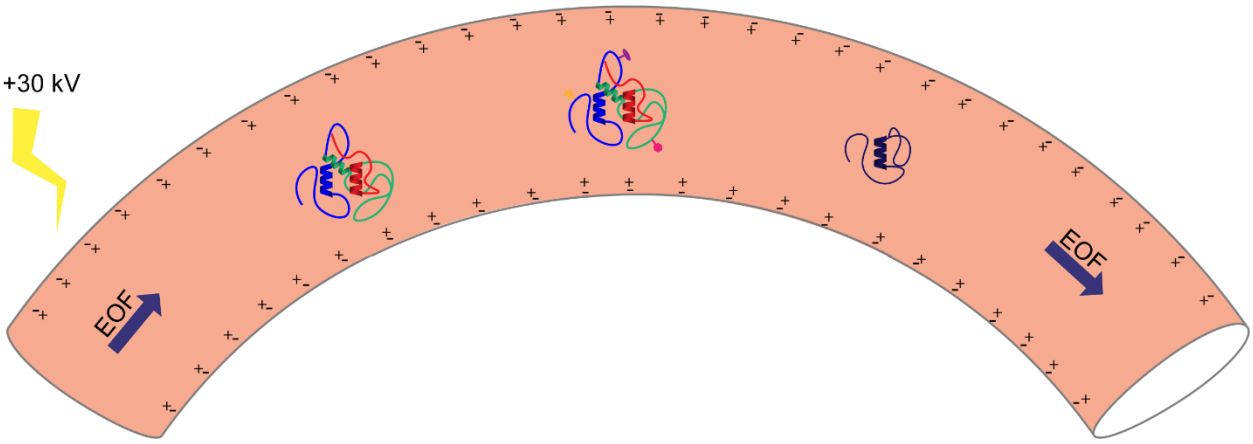


Figure 1.6. Diagram of the CZE separation theory. An open tubular capillary is placed between two buffer vials, through which a high voltage is applied. This voltage causes analytes to migrate from the injection end to the detector according to their electrophoretic mobility. The EOF that results from the electrical double layers drives the separation of analytes.

Though CZE is a hopeful technique for the characterization and identification of proteoforms, this technique has been limited to the practice due to three main disadvantages: (1) coupling CZE to the mass spectrometer, (2) small sample loading capacity, and (3) small separation window. CZE has mostly been combined with fluorescence detection, however this has limited CZE's application to discovery proteomics. Interfacing CZE to mass spectrometry has been an issue

over the years due to the low sensitivity, limited robustness limiting this application for top down proteomics.

Historically, the sample volume has been limited to 1% of the capillary volume, therefore a 100-cm long, 50-µm id capillary has only a total volume of 2 µL and the injection volume can only be roughly 20-40 nL[55]. This limited injection volume arises from the small inner diameter and the theory of the separation itself. CZE utilizes an open tubular capillary with no stationary phase, meaning that when sample is injected onto the capillary there is no "trapping" mechanism at the beginning like there is in RPLC separations and since the diffusion of the analyte is felt the most within a CZE separation, the analyte will diffuse across the capillary. In addition, CZE is considered a "fast" separation due to the EOF; when a large sample plug is injected within the capillary, the analytes will migrate out of the capillary fast. This small loading capacity has limited CZE, because the sample material cannot be fully utilized compared to LC methods. This "fast" feature of CZE also results in a small separation window leading to a small peak capacity, thus the detector that is used with CZE must also be sufficiently fast to capture these small electrophoretic peaks and even mass spectrometer have an issue responding to these subsecond peaks[46]. The generation of tandem mass spectra will be small and will impact the proteoform identification number as well. Applying this technique to TDP can have issues such as poor identification of low abundant proteoforms from complex samples and the small acquisition number of tandem mass spectra can also impede identification numbers as well.

1.2.3.1 Electrospray Ionization Interface for the Coupling of Capillary Zone Electrophoresis to Mass Spectrometry

The coupling of CZE to a mass spectrometer was made possible by specific interface designs that allows for an electrical contact to the distal end of the capillary so that an electrical circuit

can be made to drive electrophoresis and support electrospray[55-57]. Initial electrosprays were similar in design to those that are used for LC-MS and employed a high flow rate of the sheath liquid that allowed for a greater flexibility in the sheath buffer composition, however this increased sample dilution and reduced sensitivity. Currently, there are two types of electrospray interfaces that are used for coupling CZE to the mass spectrometer for top-down proteomics: (1) sheathless CZE-MS interface, and (2) sheath flow CZE-MS interface.

The sheathless CZE-MS interface functions by applying a direct voltage to the background electrolyte (BGE)[59,60]. One advantage of the sheathless interface is the decreased sample dilution and background noise resulting in a higher S/N compared to the sheath flow interfaces[59,60]. The main challenge is to maintain a closed electrical circuit during CZE separation and the ESI source[59]; this can be completed in two ways: by coating the ESI in a conductive metal or by including a porous region near the tip to have direct voltage connection to the BGE[59-62]. In 2007 the Moini Group developed a porous ESI tip for a sheathless ESI interface that produced low sensitivity and high robustness[61]. The sheathless interface was also applied to the characterization of histones using CZE-MS and identified a variety of citrullinated proteoforms on histone H4[62].

The Dovichi Group in 2010 developed the electrokinectically pumped sheath flow CZE-MS interface; this interface displayed high sensitivity and robustness (**figure 1.7**)[63]. This CZE-MS interface was later applied by Zhao et al. to the identification of *Mycobacterium marinum* secretome and yeast proteoforms using top-down proteomics[64,65]. The electrokinectically pumped sheath flow interface works by threading the separation capillary through a PEK tubing Tee into the borosilicate glass electrospray emitter that has been pulled to a 30-40 µm outer diameter. A plastic tube connects a side arm of the tee to a sheath reservoir, which is connected

through a platinum electrode that is held at roughly ~1000 V. The plastic side arm acts as a sort of salt bridge that separates the sheath buffer and the electrolysis products produced at the electrode[66-68]. The pka of borosilicate glass is ~3.5 while the sheath buffer is a mixture of methanol and acetic acid with a pH that is higher than this pka[67]. At this pH, the cation will be attracted to the negatively charged silanol groups on the borosilicate glass electrospray emitter creating an electric double layer. The potential that is applied to the sheath buffer will then drive electroosmosis in the emitter to pump the sheath liquid at a nL/min flow rate out of the outer borosilicate electrospray emitter[66-68]. The EOF produced will surround the analyte with water molecules produced from the sheath buffer and exit the separation capillary[66-68].

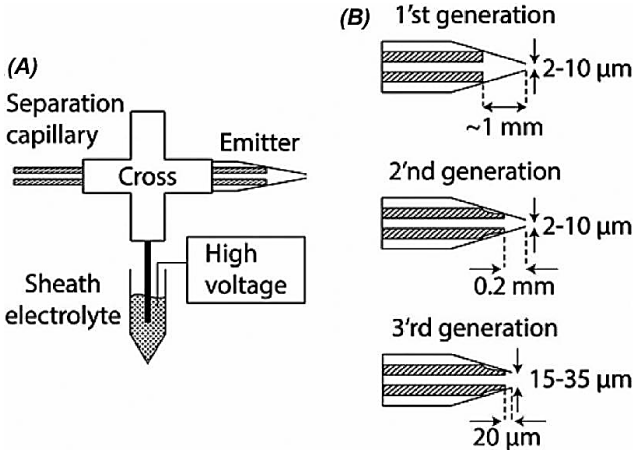


Figure 1.7. (**A**) The design of the electrokinetically pumped sheath flow CE-MS interface. (**B**) Three different generations of the CZE-MS interface illustrating that a larger emitter opening will allow for a shorter distance between the capillary end and the emitter opening, increasing the sensitivity and robustness. **Reprinted with permission from reference [68]. Copyright (2021) American Chemical Society.**

1.2.3.2 Preconcentration Methods to Increase Sample Loading Capacity

A major limitation of CZE is the loading capacity (typical loading capacities are around1%, or 20-nL, of the total capillary length), which affects both the sensitivity of the separation and the identification of low abundant proteoforms from complex proteome samples[51]. One advantage of

CZE is that there are sample stacking methods that can preconcentration samples by just changing the solvent of the separation[51,69,72]. Sample stacking can enhance the analyte detection by two factors: (1) analyte bands within the capillary will narrow causing a smaller peak width and a larger peak height leading to a greater S/N, improving LODs, and (2) the sample volume can be increased due to the smaller peak widths without loss in separation efficiency, because more analyte can reach the detector [51,69,72]. There are multiple preconcentration methods, such as field enhanced sample stacking (FESS), solid phase extraction (SPE), and dynamic pH junction.

The simplest stacking method is called FESS; the addition of organic solvents, such as isopropanol alcohol (IPA), to the sample buffer will decrease the conductivity of the sample buffer zone in comparison to the higher conductivity BGE buffer zone[,69,72]. FESS works on the principle that analytes will experience a higher mobility when migrating through a lower conductivity buffer zone (sample zone) when under the influence of an electric field[51,69,72]. When the analyte reaches a higher conductivity zone (BGE zone), it will dramatically slow down at the interface between the two zones, essentially stacking the analytes at the interface between them[51,69,72]. Zhao et al. used FESS (sample was dissolved in 35% (v/v) acetic acid with 50% (v/v) ACN and separated in a 5% (v/v) acetic acid BGE) with CZE-MS for the analysis of reduced monoclonal antibodies; the sample loading capacity was roughly 7% of the totally capillary length and both the heavy and light chains of the antibody were baseline resolved with 3-30 µg/mL detection limits[70].

SPE-CZE is an online preconcentration method that can concentration dilute samples on the capillary directly[72]. SPE-CZE employs covalently anchored monoliths or nanoparticles that are prepared directly in the capillary, essentially eliminating frits[72]. Lin et al. preconcentrated acidic and basic proteins using both centrifugal ultrafiltration and then loading the proteins onto a

nanoparticle filled CZE capillary[73]. This method was able to achieve a 110-nL loading volume with nanomolar LODs[73].

A widely used sample stacking method created by the Chen group in 2000 used in BUP is called dynamic pH junction[74]. The stacking method works by utilizing the pH differences between the sample zone and the BGE zone within the capillary. The sample is typically dissolved in a basic buffer, such as ammonium bicarbonate (pH ~8) and the BGE buffer is acidic, such as 5%(v/v) acetic acid (pH ~2.4). First, the capillary is filled with the acidic BGE buffer and then a long plug of the basic buffer is injected into capillary. Both ends of the capillary are immersed in the acidic BGE, essentially creating two different pH boundaries within the capillary: pH boundary 1 at the injection end of the capillary and pH boundary 2 somewhere within the capillary. After applying a high voltage to the injection end of the capillary, protons from the acidic BGE will migrate into the basic sample zone moving pH boundary 1 towards pH boundary 2, slowly titrating the basic sample zone. In the meantime, the anions from the basic sample zones will migrate backwards towards pH boundary 1. pH boundary 1 will move towards pH boundary 2 until the two boundaries meet, essentially stacking the sample between the two boundaries, at which point the sample undergoes normal CZE separation. Chen et al. systemically evaluated dynamic pH junction for bottom-up proteomics that produced a 140-min separation window with a μL loading capacity for complex proteome digests[75]. Zhao et al. utilized dynamic pH junction (5 mM ammonium bicarbonate (pH 8) as the sample buffer and 5% (v/v) acetic acid (pH 2.4) as the BGE) for the analysis of a yeast lysate[42]. This study produced a 240-nL sample loading capacity leading to a separation window of 30 mins and a peak capacity of 100[42].

1.2.3.3 Capillary Coatings to Increase Separation Window for CZE

An open tubular fused silica capillary is used for CZE separation; therefore, the inner wall of the capillary has negatively charged silanol groups. When the capillary is filled with BGE and a high voltage is applied across the capillary, an EOF will be produced allowing for the migration of the analytes out of the capillary for detection. This EOF process causes a rapid separation, resulting in a separation window of 1-30 mins, making this an appealing feature for high throughput analysis of simple samples[51]. However, if CZE-MS/MS is to be applied to complex large-scale top-down proteomic samples, this fast separation isn't appealing as the acquisition number of MS/MS spectra is low limiting the number of identifications that can be produced. Therefore, increasing the separation window is of importance for large-scale proteoform identification rate; one critical component of CZE is that modification of the inner wall can be easily accomplished and can have an impact on the separation window by eliminating the EOF. There have been numerous types of neutral, hydrophilic and cationic coatings on the inner wall that have been used to eliminate the EOF within the capillary to increase the separation window, as well as to reduce protein adsorption to the inner wall.

The most used neutral coating for top-down proteomic based CZE-MS analysis is the linear polyacrylamide (LPA) coatings. Belov et al. found that neutral coatings produced higher separation efficiency for native and intact proteoforms compared to the bare fused capillaries as a result of lower protein adsorption to the inner wall of the capillary[76]. Reports have also shown that neutral coatings produce almost zero EOF for intact proteoforms[77]. The Dovichi group was the first to fully review LPA coatings for proteomic studies[42,44,48,50,54,64-67,70,71]. These studies have shown that the LPA coatings have high reproducibility and less than 1% relative standard deviation (RSD) for migration time for the separation window. Zhao et al. utilized an LPA-

coated capillary using dynamic pH junction based CZE-MS/MS for the analysis of a yeast lysate that produced a separation window of 30 mins and a peak capacity of 100[42].

1.3. Quantitative Top-Down Proteomics

Quantitative proteomics is used for both discovery and targeted proteomics and is important for gaining critical insights into global proteomic dynamics within organisms[9,69]. Quantitative information can be used to increase our understanding of biological disorders for disease pathology and hopefully therapeutic medicine. Many disease pathways can be regulated by abnormal PTM's on a variety of proteins, therefor discovering proteoform-level abundant changes are important[69]. Due to the sensitivity issues (i.e. lower S/N ratio) and separation reproducibility issues that are inherent to top-down proteomics, quantitating proteoforms is challenging[9]. However, BUP quantification still has the protein inference problem, meaning that protein quantification is inferred from its quantified constituent peptides[9]. In proteoform quantification there is no need to infer proteoforms from ambiguous peptide sequences, because direct intact proteoforms are quantified[9]. Quantitative top-down proteomics can be accomplished using two strategies: label-free quantification (LFQ) and isotopic labeling quantification.

1.3.1 Label-Free Quantification

The main difference between LFQ and isotopic labeling quantification is that there is no need for isotopic labeling for the samples, therefore LFQ can be used for samples that were not grown in cell cultures. LFQ is completed by comparing the MS1 spectra of two or more biological samples to determine relative proteoform abundance[9,69]. Generally, when using a separation method before mass spectrometry analysis, the extracted peak areas from the chromatogram/electropherogram for specific ions can be used for relative quantification of

proteoforms between samples[9,69,76]. However, two important factors need to be considered: (1) normalization need to be applied for the measurement variations between the peak intensities of the proteoforms from technical replicates, and (2) variations of migration time and m/z values of identical proteoforms across different measurements need to be minimized[9,76]. LFQ is simple, inexpensive, and consistently reliable when put under statistical validation. However, LFQ is the least accurate of the quantification techniques, and there needs high reproducibility between technical runs to guarantee that the observed changes in intensity are due to biological fluctuations and not artificial.

The first instance of applying LFQ to top down proteomics was accomplished by the Kelleher Lab, which quantified over 800 proteoforms from yeast mutant vs a wild type strain[77]. This technique was accomplished by utilizing a hierarchical linear model allowing for variations to be accounted for when quantifying proteoforms with significant statistical changes across the two biological conditions[77]. The Smith Group developed Proteoform Suite and created a label-free quantification strategy[78]. This strategy utilizes the abundance differences in observed masses across different biological conditions[78]. The Smith Group used this strategy to quantify mouse mitochondrial proteoform within myoblasts and myotubes; it was found that over 100 proteoforms had statistically significant proteoform abundance changes between the two biological conditions[78].

1.3.2 Isotopic Labeling Quantification

Isotopic labeling is a quantification strategy where the proteoforms are chemically labeled and proteoforms in different samples are pooled after labeling for MS and MS/MS analysis with higher throughput compared to the label free approach. There are two major isotopic labeling

strategies that have been used for top-down proteomics: stable isotope labeling of amino acids in cell culture (SILAC) and Tandem Mass Tag (TMT).

1.3.2.1 Stable Isotope Labeling of Amino Acids in Cell Culture

In SILAC[100], two cell populations in cell cultures are cultivated where one population is fed with growth medium that has normal amino acids and the other population is fed with growth medium that contains amino acids labeled with non-radioactive heavy/light isotopes. The population using the heavy/light isotope growth medium will incorporate the heavy/light isotopes into all their proteins. The two populations will then be combined and analyzed using mass spectrometry. The mass difference between proteoforms relate to the number of heavy/light isotopes that were incorporated in the proteoform sequence. An advantage of SILAC is that the two populations can be combined immediately after cell lysis and before sample preparation. reducing the measurement variations from sample preparation, fractionation and MS analysis. Therefore, SILAC usually has lower quantitative variation and higher accuracy compared to LFQ due to the uniform sample handling.

Rhoads *et al.* developed a variant of SILAC called NeuCode, which incorporates closely spaced heavy isotope-labeled amino acids quantification of proteoforms[79]. The study incorporated $^{13}C_6{}^{15}N_2$-lysine or $^{2}H_8$-lysine, which are isotopologues of lysine that are set apart 36 mDa[79]. Shortreed et al. then used NeuCode to differentiate proteoforms by their intact mass using the number of lysines incorporated[10].

The major challenges associated with this strategy arise during the incorporation of the heavy isotopes within the cell culture populations; as the mass of the proteoform increase, there is a decreased probability of labeling the entire proteoform with the isotopic label. Furthermore, the

incorporation of the isotopes causes a complicated spectrum due to the presence of two sets of isotopic envelopes making the data analysis complicated.

1.3.2.2 Tandem Mass Tags

TMT utilizes isobaric chemical tags that can be used to label proteoforms on the N-terminus and on side-chain amino groups to provide relative quantification. This method was first introduced by Thompson et al. in 2003 for bottom-up proteomic quantification[100] and has since been applied to top-down proteomics[80,101]. The tags consist of four distinct regions: a mass reporter region, a mass normalization region, a cleavable linker region, and a protein reactive group. The protein reactive group will bind to either to the N-terminus or a side-chain amino group. TMT has multiplexing abilities and can provide relative quantification of up to 16 different samples. TMT works by using different isobaric tags to label different biological sample conditions. The samples will then be mixed, separated, and analyzed by mass spectrometry. Since the tags all contain the same mass, the proteoforms will co-elute during the separation and cannot be distinguished by MS, but during MS/MS fragmentation, the mass reporter region of the tag will break away from the tag and create a unique reporter ion within the low m/z area of the MS/MS spectra. Comparison of the reporter ion intensities will provide the proteoform abundance difference.

TMT labeling on intact proteins have only been used on standard protein mixtures as shown by Hung and Tholey[80]. Yu et al. developed a TMT-labeling platform for intact proteoform quantification from E. coli cell lysate[9]. This study was used in conjunction with TopPIC were the TMT modifications on lysine and N-termini were set as fixed PTM. It was found that 303 proteoforms were labeled at all lysines and N-termini residues and 64 proteoforms were labeled at all lysine residues, but with a missing N-termini label[9]. These results show that TMT-labeling

can be applied for top-down proteomic quantification for intact proteoforms from complex samples.

1.4. Current Challenges to Top-Down Proteomics

Top-down proteomics offer distinct advantages over bottom-up proteomics for its ability to identify proteoforms, there are still separation, fragmentation and bioinformatic technologies that are not as established as BUP and still lacking in robustness. This section will cover the remaining limitations in this field.

1.4.1 Separation

High-capacity and highly efficient separation of intact proteoforms still remains a challenge due to solubility issues and co-elution of proteoforms. Solubility causes an issue, because the larger proteoforms contain more hydrophobic amino acids that will span the proteoform and many of these insoluble proteoforms are also insoluble in MS-compatible buffers. Proteoform separation is important, because of the complexity of the proteome. As mentioned earlier, estimates say there are over 1,000,000 proteoforms in the human body, therefore there is a high chance of co-elution. Co-elution is disadvantageous, because mass spectrometers have a limited capability to distinguish different proteoforms at the same time due to a finite charge capacity therefore making low abundant proteoform identification difficult. Furthermore, tandem mass spectra of proteoforms are difficult to resolve because of the overlapping fragments, which complicates data analysis. To improve the S/N for high resolving power to reduce the complexity of the spectra, there is a need to average multiple numbers of spectra. While there have been improvements for offline intact proteoform separation to reduce complexity, the resolution of online proteoform separation still isn’t enough to prevent proteoform co-elution.

1.4.2 Fragmentation

Even with coupling multiple dimensions for fractionation to complex proteoform samples to enable the identification of thousands of proteoforms, the complete characterization of proteoforms still remain a challenge due to the lack of comprehensive gas-phase fragmentation techniques. Comprehensive gas-phase fragmentation techniques are crucial for accurate localization of PTMs on the proteoforms, therefore better fragmentation of proteoforms is critical.

For proteomics, collision-based dissociation methods (i.e. CID and HCD) are the universal gold standard. HCD has shown promise for small proteoforms for enhancing sequence coverage, however HCD favors the cleavage of the most labile bonds. Catherman et al. used LC-MS with HCD for the large-scale top-down proteomics of the human proteome resulting in over 5,000 proteoforms, however over 50% of the proteoforms were below 30 kDa and could not accurately localize the PTMs on the proteoforms even though they were detected[81].

Alternative gas-phase fragmentation techniques based on electron-based activation methods (electron transfer dissociation (ETD) and activated ion ETD (AI-ETD)) are an appealing substitute due to HCD and CID[82-84]. Riley et al. used AI-ETD to extend the m/z range for fragmentation of intact proteoform[83]. The study also show that AI-ETD increased the number of -c and -z type ions for all charge states and low charge density precursors were boosted by 4-fold for product ion yield; AI-ETD also outperformed HCD for generating fragment ions with greater sequence coverage[83].

1.4.3 Bioinformatics for Proteoform Identification

Currently, there is no single criteria for proteoform identification. Proteomic labs utilize various characterization levels, false discovery rates, and software programs. Routinely, a database search is used for proteoform identification by top-down tandem mass spectrometry (MS/MS)[1]. Proteoforms will be identified by its intact mass and the experimental MS/MS fragment peaks. The experimental MS/MS spectra will be searched against a theoretical protein database using a software program; the theoretical protein database will help to generate theoretical MS/MS spectra of each protein containing the theoretical masses of fragment ions. Comparing the experimental and theoretical MS/MS spectra is used to identify a specific proteoform from each MS/MS spectrum (a proteoform spectrum match, PrSM). The quality of the match (i.e., the number of matched fragment ions) will determine the score of the PrSM, and eventually will define the confidence of the proteoform identification. Borrowing the idea from bottom-up proteomics, the overall false discovery rate (FDR) of the proteoform identifications can be estimated by the target-decoy database search strategy[9,86,87]. The "identification" of a proteoform in top-down proteomics ultimately means the full characterization of a proteoform, including the localization of various PTMs on the sequence. However, it is very challenging because only small numbers of fragment ions can be matched during the database search for most of the identified proteoforms, leading to low backbone cleavage coverage. One of the major issues for proteoform identifications using bioinformatics tools is the high complexity of the MS/MS spectra of proteoforms since proteoforms are much larger and have much more possible PTMs than peptides in bottom-up proteomics.

There is many software developed for top-down proteomics, for example, TopPIC, Proteoform Suite, and ProSight PTM[85-88]. In this thesis, I mainly used TopPIC for proteoform identification

and label-free quantification. TopPIC uses spectral alignment for MS/MS search to determine unknown mass shifts and therefore unknown proteoforms and is useful for discovery top-down proteomics[86,87]. TopPIC is an open-source database searching software tool for high-throughput top-down MS for proteome-wide proteoform identification and characterization[86,87]. First, top-down MS/MS spectra must be pre-processed due to the highly charged, broad isotopic peaks. The pre-processing is done in two steps: generation of an XML format file containing centroided peaks from raw data files using a file format conversion tool, and a deconvolution method (i.e. MS-Deconv) to create a monoisotopic mass list that is extracted from one MS/MS spectrum[87]. At this point, the deconvoluted data will be transferred to the TopPIC software to generate a list of PrSMs, which will contain characterized proteoforms including unknown cSNPs and PTMs in the form of mass shifts.

The TopPIC software consists of 2 elements: (1) identification of proteoforms, and (2) characterization of unknown mass shifts[86,87]. The identification of proteoforms is completed by first searching the deconvoluted mass spectrum against an annotated protein sequence database which will produce a best scoring PrSM. A PrSM is produced based on three steps: first, a potential protein identification will be reduced to a few dozen options using a filtering algorithm; second, each potential identification will be aligned with each spectrum to find the best alignment using a spectral alignment algorithm; third, the PrSM with the best Expectation Value (E-value) will be reported by utilizing a generation function method that computes E values for each potential PrSMs[87.] The second element, characterization of unknown mass shifts, is based on Bayesian models to produce a MIScore method; this method uses a common modification list that is provided by the user to identify and localize modifications that are found within the PrSMs[87]. The MIScore method will assess explanations for each reported unknown mass shift

and reports the best fitting modification for that mass shift, as well as a confidence score[87]. Li et al., found that TopPIC had highly sensitive proteoform identification and high mass shift localization accuracy when no more than two mutations were on homologous protein sequences[86].

1.5. Summary

In this chapter, mass spectrometry-based proteomics was presented. While most proteomic studies utilize BUP due its high sensitivity, throughput, and robustness, this technique introduces a protein inference problem; often inferring proteins based on a small number of peptides that cover limited regions of the full-length protein. BUP has issues with distinguishing protein isoforms and proteoforms that have high sequence similarity and can lose information on the combinations of PTMs on the protein sequence after digestion. Top-down proteomics can overcome this issue by analyzing intact proteins and identifying proteoforms and their PTM's with higher confidence than BUP. Presently, TDP provides opportunities to gain valuable insight into biological mechanisms by analyzing proteoform abundances. RPLC-ESI-MS/MS has been shown to identify more than >1,000 proteoforms for large-scale top-down proteomics of complex protein samples. While the proteoform identification numbers of RPLC is excellent, there are still issues with sample loss on the stationary phase facilitating a need to have micrograms of starting material to reach these number. What is needed is a highly sensitive separation and detection technique that can enable the identification and quantification of mass limited proteome samples. CZE-ESI-MS/MS is a highly sensitive separation and detection technique that has made extraordinary progress in the proteomic field. The advances of CZE-ESI-MS interfaces, incorporating capillary coatings, and implementing sample stacking method to improve separation capacity and sensitivity has allowed for analysis of simple proteomic

samples, however applying CZE-ESI-MS/MS for large-scale top-down proteomics of complex samples has been impeded by the limited sample loading capacity and narrow separation window. The following chapters within this thesis will describe three projects improving CZE-ESI-MS/MS for the large-scale top-down proteomics of complex samples.

REFERENCES

REFERENCES

1. Banks, R. E.; Dunn, M. J.; Hochstrasser, D. F.; Sanchez, J. C.; Blackstock, W.; Pappin, D. J.; Selby, P. J. Proteomics: New Perspectives, New Biomedical Opportunities. *Lancet* **2000**, *356* (9243), 1749–1756.

2. Adams, J. The Proteome: Discovering the Structure and Function of Proteins. *Nature Education* **2008**, *1* (3), 6.

3. Garrels, J. I. Encyclopedia of Genetics. In *Proteome*; West Academic Press: Saint Paul, MN, 2001; pp 1575–1578.

4. Righetti, P. G. Proteome. In *Brenner's Encyclopedia of Genetics (Second Edition)*; Hughes, S. M. K., Ed.; West Academic Press: Saint Paul, MN, 2013; pp 504–507.

5. Proteomics. In *Encyclopedia of Food Microbiology*; West Academic Press: Saint Paul, MN, 2014; pp 793-802.

6. Aslam, B.; Basit, M.; Nisar, M. A.; Khurshid, M.; Rasool, M. H. Proteomics: Technologies and Their Applications. *J. Chromatogr. Sci.* **2017**, *55* (2), 182–196.

7. Yates, J. R., 3rd. Recent Technical Advances in Proteomics. *F1000Res.* **2019**, *8*, 351.

8. Smith, L. M.; Kelleher, N. L.; Consortium for Top Down Proteomics. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10* (3), 186–187.

9. Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; Sun, L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z.; Yu, D.; Shortreed, M. R.; Smith, L. M. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19* (10), e1800361.

10. Shortreed, M. R.; Frey, B. L.; Scalf, M.; Knoener, R. A.; Cesnik, A. J.; Smith, L. M. Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. *J. Proteome Res.* **2016**, *15* (4), 1213–1221.

11. Nesvizhskii, A. I.; Aebersold, R. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.

12. Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlén, M.; Wu, C. H.; Yamamoto, T.; Paik, Y.-K.; Omenn, G. S. The Human

Proteome Project: Current State and Future Direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111.009993.

13. LeDuc, R. D.; Schwämmle, V.; Shortreed, M. R.; Cesnik, A. J.; Solntsev, S. K.; Shaw, J. B.; Martin, M. J.; Vizcaino, J. A.; Alpi, E.; Danis, P.; Kelleher, N. L.; Smith, L. M.; Ge, Y.; Agar, J. N.; Chamot-Rooke, J.; Loo, J. A.; Pasa-Tolic, L.; Tsybin, Y. O. ProForma: A Standard Proteoform Notation. *J. Proteome Res.* **2018**, *17* (3), 1321–1325.

14. Dai, Y.; Buxton, K. E.; Schaffer, L. V.; Miller, R. M.; Millikin, R. J.; Scalf, M.; Frey, B. L.; Shortreed, M. R.; Smith, L. M. Constructing Human Proteoform Families Using Intact-Mass and Top-down Proteomics with a Multi-Protease Global Post-Translational Modification Discovery Database. *J. Proteome Res.* **2019**, *18* (10), 3671–3680.

15. Dupree, E. J.; Jayathirtha, M.; Yorkey, H.; Mihasan, M.; Petre, B. A.; Darie, C. C. A Critical Review of Bottom-up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes* **2020**, *8* (3), 14.

16. Gillet, L. C.; Leitner, A.; Aebersold, R. Mass Spectrometry Applied to Bottom-up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **2016**, *9* (1), 449–472.

17. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R., 3rd. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394

18. Kulak, N. A.; Geyer, P. E.; Mann, M. Loss-Less Nano-Fractionator for High Sensitivity, High Coverage Proteomics. *Mol. Cell. Proteomics* **2017**, *16* (4), 694–705.

19. Armirotti, A.; Damonte, G. Achievements and Perspectives of Top-down Proteomics. *Proteomics* **2010**, *10* (20), 3566–3576.

20. Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **2016**, *9* (1), 499–519.

21. Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445* (4), 683–693.

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.; Paša-Tolić, 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. *Nat. Methods* **2019**, *16* (7), 587–594.

23. Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher,

N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How Many Human Proteoforms Are There? *Nat. Chem. Biol.* **2018**, *14* (3), 206–214.

24. Pandey, A.; Mann, M. Proteomics to Study Genes and Genomes. *Nature* **2000**, *405* (6788), 837–846.

25. Han, X.; Aslanian, A.; Yates, J. R., 3rd. Mass Spectrometry for Proteomics. *Curr. Opin. Chem. Biol.* **2008**, *12* (5), 483–490.

26. Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003**, *422* (6928), 198–207.

27. The 2002 Nobel Prize in Chemistry - Popular information - NobelPrize.org https://www.nobelprize.org/prizes/chemistry/2002/popular-information/ (accessed Jan 24, 2021).

28. Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. Unraveling the Mechanism of Electrospray Ionization. *Anal. Chem.* **2013**, *85* (1), 2–9.

29. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. *Nat. Methods* **2007**, *4* (9), 709–712.

30. Goldfarb, D.; Wang, W.; Major, M. B. MSAcquisitionSimulator: Data-Dependent Acquisition Simulator for LC-MS Shotgun Proteomics. *Bioinformatics* **2016**, *32* (8), 1269–1271.

31. Roepstorff, P.; Fohlman, J. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomed. Mass Spectrom.* **1984**, *11* (11), 601.

32. Michalski, A.; Neuhauser, N.; Cox, J.; Mann, M. A Systematic Investigation into the Nature of Tryptic HCD Spectra. *J. Proteome Res.* **2012**, *11* (11), 5479–5491.

33. Macias, L. A.; Santos, I. C.; Brodbelt, J. S. Ion Activation Methods for Peptides and Proteins. *Anal. Chem.* **2020**, *92* (1), 227–251.

34. Righetti, P. G.; Castagna, A.; Antonioli, P.; Boschetti, E. Prefractionation Techniques in Proteome Analysis: The Mining Tools of the ThirdMillennium. *Electrophoresis* **2005**, *26* (2), 297–319.

35. Simpson, D. C.; Ahn, S.; Pasa-Tolic, L.; Bogdanov, B.; Mottaz, H. M.; Vilkov, A. N.; Anderson, G. A.; Lipton, M. S.; Smith, R. D. Using SEC-RPLC and RPLC-CIEF as Two-

Dimensional Separation Strategies for Protein Profiling. *Electrophoresis.* **2006**, 27 (13), 2722–2733.

36. Qian, W.-J.; Jacobs, J.; Liu, T.; Camp, D. G.; Smith, R. D. Advances and Challenges in Liquid Chromatography-Mass Spectrometry Based Proteomic Profiling for Clinical Applications. *Mol Cell Proteomics.* **2006**, *5* (10), 1727–1744.

37. Y., S.; R., Z.; Moore, R. J.; J., K.; Metz, T. O.; R., Z.; Livesay, E. A.; Udseth, H. R.; Smith, R. D. Automated 20 Kpsi RPLC-MS and MS/MS with Chromatographic Peak Capacities of 1000−1500 and Capabilities in Proteomics and Metabolomics. *Anal. Chem.* **2007**, *5* (10), 3090–3100.

38. 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.* **2018**, *90* (9), 5529–5533.

39. Valaskovic, G.A., Kelleher, N.L., McLafferty, F.W.: Attomole protein characterization by capillary electrophoresis-mass spectrometry. *Science*. **1996,** 273, 1199–1202 (1996)

40. 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.* **2014**, 86, 11006–11012.

41. 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.* **2014**, 13, 6078–6086.

42. 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.* **2016**, 15, 3679–3685.

43. 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.* 2014, 14, 1158–1164.

44. 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.* **2013**, 85, 5989–5995.

45. 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.* **2013**, 85, 2289–2296.

46. 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.* **2016**, 88, 1138–1146.

47. 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.* **2013**, 12, 2640–2656.

48. 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.* **2014**, 86, 4873–4878.

49. Jorgenson, J. W.; Lukacs, K. D. Zone Electrophoresis in Open-Tubular Glass Capillaries. *Anal. Chem.* **1981**, *53* (8), 1298–130.

50. Zhang, Z.; Peuchen, E. H.; Dovichi, N. J. Surface-Confined Aqueous Reversible Addition-Fragmentation Chain Transfer (SCARAFT) Polymerization Method for Preparation of Coated Capillary Leads to over 10 000 Peptides Identified from 25 Ng HeLa Digest by Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry. *Anal. Chem.* **2017**, *89* (12), 6774–6780.

51. Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. Capillary Zone Electrophoresis-Mass Spectrometry for Top-down Proteomics. *Trends Analyt. Chem.* **2019**, *120* (115644), 115644.

52. Faserl, K.; Sarg, B.; Kremser, L.; Lindner, H. Optimization and Evaluation of a Sheathless Capillary Electrophoresis-Electrospray Ionization Mass Spectrometry Platform for Peptide Analysis: Comparison to Liquid Chromatography-Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **2011**, *83* (19), 7297–7305.

53. Whatley, H. Basic Principles and Modes of Capillary Electrophoresis. *In Clinical and Forensic Applications of Capillary Electrophoresis*; Humana Press: Totowa, NJ, 2001; pp 21–58.

54. Cheng, Y. F.; Wu, S.; Chen, D. Y.; Dovichi, N. J. Interaction of Capillary Zone Electrophoresis with a Sheath Flow Cuvette Detector. *Anal. Chem.* **1990**, *62* (5), 496–503.

55. Maxwell, E. J.; Chen, D. D. Y. Twenty Years of Interface Development for Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry. Anal. Chim. Acta 2008, 627 (1), 25–33.

56. Krenkova, J.; Foret, F. On-Line CE/ESI/MS Interfacing: Recent Developments and Applications in Proteomics. *Proteomics* **2012**, *12* (19–20), 2978–2990.

57. Klepárník, K. Recent Advances in the Combination of Capillary Electrophoresis with Mass Spectrometry: From Element to Single-Cell Analysis: CE and CEC. *Electrophoresis* **2013**, *34* (1), 70–85.

58. Fonslow BR, Yates JR, 3rd. 2009. Capillary electrophoresis applied to proteomic analysis. *J. Sep. Sci* **32**(8): 1175– 1188.

59. Fonslow, B. R.; Yates, J. R., 3rd. Capillary Electrophoresis Applied to Proteomic Analysis. *J. Sep. Sci.* **2009**, *32* (8), 1175–1188.

60. Yin, Y.; Li, G.; Guan, Y.; Huang, G. Sheathless Interface to Match Flow Rate of Capillary Electrophoresis with Electrospray Mass Spectrometry Using Regular-Sized Capillary: Sheathless Interface for CE-MS Coupling. *Rapid Commun. Mass Spectrom.* **2016**, *30*, 68–72.

61. Moini, M. Simplifying CE−MS Operation. 2. Interfacing Low-Flow Separation Techniques to Mass Spectrometry Using a Porous Tip. *Anal. Chem.* **2007**, 79, 4241−4246

62. Faserl, K.; Sarg, B.; Maurer, V.; Lindner, H. H. Exploiting Charge Differences for the Analysis of Challenging Post-Translational Modifications by Capillary Electrophoresis-Mass Spectrometry. *J. Chromatogr. A.* **2017**, 1498, 215–223.

63. Faserl, K.; Sarg, B.; Maurer, V.; Lindner, H. H. J. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Chromatogr. A.* **2017**, 1498, 215−223

64. 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.* **2014**, *86* (10), 4873–4878.

65. Sun, L.; Zhu, G.; Yan, X.; Zhang, Z.; Wojcik, R.; Champion, M. M.; Dovichi, N. J. Capillary Zone Electrophoresis for Bottom-up Analysis of Complex Proteomes. *Proteomics* **2016**, *16* (2), 188–196.

66. Peuchen, E. H.; Zhu, G.; Sun, L.; Dovichi, N. J. Evaluation of a Commercial Electro-Kinetically Pumped Sheath-Flow Nanospray Interface Coupled to an Automated Capillary Zone Electrophoresis System. *Anal. Bioanal. Chem.* **2017**, *409* (7), 1789–1795.

67. Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. A third-generation electro-kinetically 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.* **2015**, 14 (5), 2312–2321.

68. Chen, Y.; Lü, W.; Chen, X.; Teng, M. Review of Recent Developments of On-Line Sample Stacking Techniques and Their Application in Capillary Electrophoresis. *Open Chem.* **2012**, *10* (3), 611–638.

69. Gomes, F. P.; Yates, J. R., 3rd. Recent Trends of Capillary Electrophoresis-Mass Spectrometry in Proteomics Research. *Mass Spectrom. Rev.* **2019**, 38 (6), 445–460.

70. Zhao, Y.; Sun, L.; Knierman, M. D.; Dovichi, N. J. Fast Separation and Analysis of Reduced Monoclonal Antibodies with Capillary Zone Electrophoresis Coupled to Mass Spectrometry. *Talanta* **2016**, *148,* 529–533.

71. Zhang ZB, Qu YY, Dovichi NJ, Capillary zone electrophoresis-mass spectrometry for bottom-up proteomics, *Trac-Trends in Analytical Chemistry.* **2018,** 108, 23–37.

72. Chen, Y.; Lü, W.; Chen, X.; Teng, M. Review of Recent Developments of On-Line Sample Stacking Techniques and Their Application in Capillary Electrophoresis. *Open Chem.* **2012**, *10* (3), 611–638.

73. Britz-McKibbin P.; Chen D.D. Selective focusing of catecholamines and weakly acidic compounds by capillary electrophoresis using a dynamic pH junction. Anal Chem. 2000 Mar 15; 72(6):1242-52.

74. Chen, D.; Shen, X.; Sun, L. Strong Cation Exchange-Reversed Phase Liquid Chromatography-Capillary Zone Electrophoresis-Tandem Mass Spectrometry Platform with High Peak Capacity for Deep Bottom-up Proteomics. *Anal. Chim. Acta* **2018**, *1012*, 1–9.

75. Belov, A.M.; Viner, R.; Santos, M.R.; Horn, D.M.; Bern, M.; Karger, B.L.; Ivanov, A.R. Analysis of proteins, protein complexes, and organellar proteomes using sheathless capillary zone electrophoresis–native mass spectrometry. *J Am Soc Mass Spectrom* **2017** 28(12): 2614– 2634.

76. Nikolov, M.; Schmidt, C.; Urlaub, H. Quantitative Mass Spectrometry-Based Proteomics: An Overview. *Methods Mol. Biol.* **2012**, *893,* 85–100.

77. Ntai, I.; Kim, K.; Fellers, R. T.; Skinner, O. S.; Smith, A. D., 4th; Early, B. P.; Savaryn, J. P.; LeDuc, R. D.; Thomas, P. M.; Kelleher, N. L. Applying Label-Free Quantitation to Top down Proteomics. *Anal. Chem.* **2014**, *86* (10), 4961–4968.

78. 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.* **2018**, *17* (10), 3526–3536.

79. Rhoads, T. W.; Rose, C. M.; Bailey, D. J.; Riley, N. M.; Molden, R. C.; Nestler, A. J.; Merrill, A. E.; Smith, L. M.; Hebert, A. S.; Westphall, M. S.; Pagliarini, D. J.; Garcia, B. A.; Coon, J. J. Neutron-Encoded Mass Signatures for Quantitative Top-down Proteomics. *Anal. Chem.* **2014**, *86* (5), 2314–2319.

80. C.-W. Hung and A. Tholey, Tandem Mass Tag Protein Labeling for Top-Down Identification and Quantification, Anal. Chem., 2012, 84(1), 161–170.

81. 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. *Mol. Cell. Proteomics* **2013**, *12* (12), 3465–3473.

82. Riley, N. M.; Coon, J. J. The Role of Electron Transfer Dissociation in Modern Proteomics. *Anal. Chem.* **2018**, *90* (1), 40–64.

83. Riley, N. M.; Westphall, M. S.; Coon, J. J. Activated Ion Electron Transfer Dissociation for Improved Fragmentation of Intact Proteins. *Anal. Chem.* **2015**, *87* (14), 7109–7116.

84. Rush, M. J. P.; Riley, N. M.; Westphall, M. S.; Coon, J. J. Top-down Characterization of Proteins with Intact Disulfide Bonds Using Activated-Ion Electron Transfer Dissociation. *Anal. Chem.* **2018**, *90* (15), 8946–8953.

85. Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y.-B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: Improved Protein Identification and Characterization for Top down Mass Spectrometry. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W701-6.

86. Li, Z.; He, B.; Kou, Q.; Wang, Z.; Wu, S.; Liu, Y.; Feng, W.; Liu, X. Evaluation of Top-down Mass Spectral Identification with Homologous Protein Sequences. *BMC Bioinformatics* **2018**, *19* (Suppl 17), 494.

87. 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.

88. Cesnik, A. J.; Shortreed, M. R.; Schaffer, L. V.; Knoener, R. A.; Frey, B. L.; Scalf, M.; Solntsev, S. K.; Dai, Y.; Gasch, A. P.; Smith, L. M. Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families. *J. Proteome Res.* **2018**, *17* (1), 568–578.

89. Durbin, K. R.; Tran, J. C.; Zamdborg, L.; Sweet, S. M. M.; Catherman, A. D.; Lee, J. E.; Li, M.; Kellie, J. F.; Kelleher, N. L. Intact Mass Detection, Interpretation, and Visualization to Automate Top-Down Proteomics on a Large Scale. *Proteomics* **2010**, *10* (20), 3589–3597.

90. Liu, X.; Segar, M. W.; Li, S. C.; Kim, S. Spectral Probabilities of Top-down Tandem Mass Spectra. *BMC Genomics* **2014**, *15 Suppl 1* (S1), S9.

91. Zambborg, L. The Complexity of Bioinformatics: Techniques for Addressing the Combinatorial Explosion in Proteomics and Genomics in Department of Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, **2010**.

92. 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. 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), 254–258.

93. 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.* **2016**, 15 (3), 976–982.

94. 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.* **2017**, 89 (10), 5467–5475.

95. 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.* **2013**, 110 (25), 10153–10158.

96. 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. 2017*, 1498, 99–110.

97. 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.* **2017**, 16 (2), 609–618 .

98. 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.* **2017**, 16 (2), 1087–1096.

99. 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.* **2018**, 17 (10), 3526–3536.

100. Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **2002**, *1* (5), 376–386.

101. Eliuk, S.; Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **2015**, *8* (1), 61–80.

[1]CHAPTER 2. Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 Escherichia coli Proteoforms

2.1 Introduction

Capillary zone electrophoresis-electrospray ionization-mass spectrometry (CZE-ESI-MS) has been well acknowledged for the characterization of proteoforms[1-4]. RPLC-ESI-MS/MS is the most widely used method for the characterization of proteoforms. However, CZE-ESI-MS/MS has been recommended as an alternative to this widely used method[5-15].

Currently, top-down proteomics can be accomplished using CZE-MS, because the current ESI interfaces are sensitive and robust resulting in hundreds of proteoform identifications. However, there are two challenges that are impeding CZE-MS/MS for top-down proteomics: (1) the sample loading capacity, and (2) narrow separation window[1-15]. Currently, the largest reported sample capacity for CZE-MS/MS based top-down proteomics is 200-nL[8,10]. This small sample loading capacity cannot fully utilize the sample material resulting in low identification numbers, as well as hinders the identification of low abundant proteoforms from complex proteome samples. The narrow separation window impedes the number of tandem mass spectra that can be acquired in one experiment, limiting the proteoform identification number; the longest reported separation window for top-down proteomic based CZE-MS/MS is ~30-mins.

[1] This chapter was adapted with permission from 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.* **2017**, 89 (22), 12059–12067.

One advantage of CZE is that samples can be preconcentrated directly onto the capillary by sample stacking methods that can both improve the sample loading capacity and separation window. Chen et al. systemically evaluated dynamic pH junction-based CZE-MS/MS for bottom-up proteomics resulting in a 140-min separation window and a uL loading capacity for the analysis of complex proteome digests[20]. One stacking method is called dynamic pH junction; it works by simply utilizing the difference in pH to stack the sample directly onto the capillary[21,22]. The sample will be dissolved within a basic buffer and the background electrolyte (BGE) is acidic. First, the capillary will be filled with the acidic BGE and a long plug of sample will be pressure injected onto the capillary essentially creating two pH boundaries: pH boundary I is at the injected end of the capillary and pH boundary II is within the capillary. Next, both ends of the capillary will be emerged in acidic BGE vials and a voltage will be applied to the capillary. Protons from the BGE will migrate into the sample plug, moving pH bound I towards pH boundary II. Meanwhile, the anions from the basic buffer will move back towards the injection end, focusing at the moving pH boundary I. The sample zone will be slowly be titrating and moving pH boundary I to H boundary II when the two boundaries meet, essentially stacking the sample at the interface. After which a normal CZE separation will continue[23-26].

There has been no systemic evaluation of dynamic pH junction for the concentration of proteoforms for top-down proteomics, even though it has been used for the centration of peptides and small molecules. Zhao et al. is the only published paper that has used dynamic pH junction-based CZE-MS/MS for the large-scale top-down proteomics of a yeast lysate. This study used 5 mM ammonium bicarbonate (pH 8) as the sample buffer and 5% (v/v) acetic acid (pH 2.4) as the BGE. In terms of the sample loading capacity and separation window, the study injected 100 to 240-nL onto the capillary for CZE-MS/MS analysis resulting in a 30-min separation window

with a 100-peak capacity. For the first time, this study systemically evaluated dynamic pH junction-based CZE-MS/MS for large-scale top-down proteomics of proteoforms. The optimized CZE-MS/MS platform enabled a uL scale loading capacity, 90-min separation window, ~280 peak capacity resulting in identification of 600 proteoforms from a *E. coli* proteomics using single-shot CZE-MS/MS.

2.2 Experimental

2.2.1 Materials and Reagents

Acrylamide was purchased from Acros Organics (NJ, USA). Standard proteins, ammonium bicarbonate (NH4HCO3), 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). Aqueous mixtures were filtered with Nalgene Rapid-Flow Filter units (Thermo Scientific) with 0.2 µm CN membrane and 50 mm diameter. Fused silica capillaries (50 µm i.d./360 µm o.d.) were obtained from Polymicro Technologies (Phoenix, AZ).

2.2.2 Sample Preparation

A mixture of standard proteins consisting of lysozyme (14.3 kDa, pI 11.0, 0.1 mg/mL), cytochrome c (Cyto.c, 12 kDa, pI 10.0, 0.1 mg/mL), myoglobin (16.9 kDa, pI 7.0, 0.1 mg/mL), β-casein (24 kDa, pI 4.5, 0.4 mg/mL), carbonic anhydrase (CA, 29 kDa, pI 5.1, 0.5 mg/mL), 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 appropriately with different buffers for various experiments.

Escherichia coli (*E. coli*, strain K-12 substrain MG1655) was cultured in LB medium at 37 oC with 225 rpm shaking until OD600 reached 0.7. *E. coli* cells were harvested by centrifuge at 4,000 rpm for 10 min. Then the *E. coli* cells were washed with PBS three times. The *E. coli* cells were then lysed in a lysis buffer containing 8 M urea, 100 mM Tris-HCl (pH 8.0) and protease inhibitors. The cell lysis was assisted by sonication with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 10 minutes. After centrifugation (18,000 x g for 10 min), the supernatant containing the extracted proteins was collected. A small aliquot of the extracted proteins was used for BCA assay to determine the protein concentration. The leftover protein extracts were stored at -80 ˚C before use. 1 mg of *E. coli* proteins in 8 M urea and 100 mM Tris-HCl (pH 8.0) were denatured at 37 oC, reduced with dithiothreitol (DTT) and alkylated with iodoacetamide (IAA). Then, the proteins were desalted with a C4-trap column (Bio-C4, 3 µm, 300Å, 4.0 mm i.d., 10 mm long) from Sepax Technologies, Inc. (Newark, DE). A HPLC system (Agilent Technologies, 1260 Infinity II) was used. The HPLC eluate from the trap column was collected and further lyophilized with a vacuum concentrator (Thermo Fisher Scientific). The dried protein sample was reconstituted in 50 mM NH4HCO3 (pH 8.0) to get about 2 mg/mL protein concentration (theoretical concentration based on 100% recovery from the whole sample preparation process) for CZE-MS/MS analysis.

2.2.3 CZE-ESI-MS/MS

An automated CZE-ESI-MS system was used in the experiments. The system contained an ECE-001 CE autosampler and a commercialized electro-kinetically pumped sheath flow CE–MS

interface from CMP Scientific (Brooklyn, NY).(16, 27) The CE system was coupled to a LTQ-XL or a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific).

A fused silica capillary (50-μm i.d., 360-μm o.d., 1 m long) was used for CZE separation. The inner wall of the capillary was coated with linear polyacrylamide (LPA) based on refs 20 and 28. One end of the capillary was etched with hydrofluoric acid based on ref 29 to reduce the outer diameter of the capillary. (Caution: use appropriate safety procedures while handling hydrofluoric acid solutions.) Different BGEs were used for CZE, including 5–10% (v/v) acetic acid and 0.1–0.5% (v/v) formic acid. 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 volume. High voltage (30 or 20 kV) was applied at the injection end of the separation capillary for separation, and 2–2.2 kV was applied for ESI. At the end of each CZE-MS run, we flushed the capillary with BGE by applying 5 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.

For all the LTQ-XL experiments, only MS1 spectra were acquired using positive ion mode, and no protein fragmentation was performed. The scan range was m/z 600-2,000 using three microscans. The maximum injection time was 50-ms and the AGC target value was 3.0E4. For all the standard-protein-mixture experiments on the Q-Exactive HF mass spectrometer, only MS1 spectra were acquired and no protein fragmentation was performed. "Intact protein mode" was used for all experiments with a trapping pressure of 0.2. The temperature of the ion transfer capillary was 320 ˚C and the s-lens RF level was 55. Full MS scans were acquired with the

number of microscans as three, the resolution as 240,000 (at m/z 200), the AGC target value as 1E6, the maximum injection time as 50-ms and the scan range as 600-2000 m/z. Data-dependent acquisition (DDA) methods were used for analysis of the E.coli sample on the Q-Exactive HF mass spectrometer. The MS/MS spectra were acquired with the number of microscans as one, the resolution as 120,000 (at m/z 200), the AGC target value as 1E5 and the maximum injection time as 200 ms. The three or eight most intense ions (Top 3 or Top 8 DDA) in the full MS spectrum were sequentially isolated with 4 m/z isolation window and further sequentially fragmented in the HCD fragmentation cell with NCE as 20%. The intensity threshold for triggering fragmentation was 1.0E5. Charge exclusion and exclude isotopes were turned on. Only protein ions with charge state higher than five can be isolated for fragmentation. The dynamic exclusion was turned on, and the setting was 30 s. The other parameters were the same as those used for the standard-protein-mixture experiments.

2.2.4 Measurement of electroosmotic mobility

The protocol used here for measuring the electroosmotic mobility in the LPA coated capillary was based on references [1] and [2]. Benzyl alcohol (neutral marker, M.W. 108.14 g/mol) was dissolved in the background electrolyte (BGE) and used as the sample. The LPA coated capillary (50-μm i.d., 360-μm o.d., 1 meter long) was flushed and filled with the BGE. First, the neutral marker (N1) was injected by applying 5-psi pressure for tinj (2s). Then, a plug of BGE was injected into the separation capillary by applying 5-psi pressure for time tr (40s). After that, a second short plug of neutral marker (N2) was injected into the capillary for tinj (2s). Subsequently, another plug of BGE was injected into the capillary by applying pressure for tr. The separation voltage (30 kV) was then applied at the injection end of the capillary for tmig (50 min). During this period, the two neutral markers (N1 and N2) moved toward the cathode end

with mobilities that were equal to the electroosmotic mobility (μeof). After tmig has been completed, a third short plug of neutral marker (N3) was injected into the capillary for tinj (2s). Finally, 5-psi pressure was applied at the injection end of the capillary, which was immersed in the BGE, to push the three plugs of neutral marker out of the separation capillary, and the MS data acquisition was simultaneously started to record the signal of the neutral marker. The μeof was calculated by:

$$\frac{[(t_{N3} - t_{N1}) - (t_{N2} - t_{N1})] * L^2}{V_{separation} * t_{mig} * (t_{N3} + \frac{t_{inj}}{2})}$$

Where t_{N1}, t_{N2}, t_{N3} are the observed migration time for neutral marker N1, N2, and N3. L corresponds to the length of the capillary, $V_{separation}$ is the separation voltage applied and t_{inj} is the injection time.

2.2.5 Data Analysis

The standard protein data was 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* RAW files were analyzed with the TopFD[30] (TOP-down mass spectrometry feature detection) and TopPIC (TOP-down mass spectrometry based proteoform identification and characterization) pipeline[31]. TopFD is an improved version of MS-Deconv[32]. It converts precursor and fragment isotope clusters into monoisotopic masses and finds possible proteoform features in CZE-MS data by combining precursor isotope clusters with similar monoisotopic masses and close migration times (the isotopic clusters may have different charge states). The

RAW files were first transferred into mzXML files with Msconvert tool[33]. Then, spectral deconvolution was performed with TopFD to generate msalign files. Finally, TopPIC (version 1.1.3) was used for database searching with msalign files as input. *E. coli* (strain K12) UniProt database (UP000000625, 4307 entries, version June 7, 2017) was used for database search. The spectrum-level false discovery rate (FDR) was estimated using the target-decoy approach[34]. Cysteine carbamidomethylation was set as a fixed modification, and the maximum number of unexpected modifications was 2. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da, and the identified proteoform-spectrum matches (PrSMs) were filtered with a 1% FDR at the spectrum level. In order to reduce the redundancy of proteoform identifications, we considered the proteoforms identified by multiple spectra as one proteoform ID if those spectra correspond to the same proteoform feature reported by TopFD or those proteoforms are from the same protein and have similar precursor masses (within 1.2 Da).

2.3 Results and Discussion

If CZE-MS is to be applied to large-scale top-down proteomics, then the sample loading capacity and the separation window needs to be improved for the characterization of complex proteomes. Our group recently demonstrated, using complex peptide mixtures, that dynamic pH junction-based CZE-MS can reach up to uL scale loading capacity with a 140-min separation window. We theorize that we can use dynamic pH junction-based CZE-MS for the characterization of proteoforms from complex proteomes. To test the theory, we first evaluated the performance of dynamic pH junction by comparing different sample loading injections (50 to 500-nL) to another sample stacking method, field enhanced sample stacking (FESS), using a variety of six standard proteins. After the comparison of the stacking methods, we then optimize the dynamic pH

junction-based CZE-MS platform and applied the optimized platform to a top-down proteomics of an *E. coli* proteome.

2.3.1 Comparison of Dynamic pH Junction and FESS Methods

First, we compared the concentration performance of dynamic pH junction to FESS for proteoforms during CZE-MS analysis across four different sample injection volumes. The volumes were: 50-nL (2.5% of the total capillary volume), 100-nL (5% of the total capillary volume), 200-nL (10% of the total capillary volume), and 500-nL (25% of the total capillary volume). The stock solution (2.2 mg/mL) was diluted 2x (~1 mg/mL) for the comparison experiments. **Figure 2.1** shows the protein intensity as a function of sample injection volume for control (A), FESS (B), and dynamic pH junction (C) methods. Protein intensities were obtained using the extracted ion electropherograms (EIE) using the highest charge state for each protein within the mixture. **Figure 2.1** legend has the m/z that were used for the EIEs. **Figure 2.1D** illustrates the EIEs of the standard protein mixture from the control, FESS, and dynamic pH junction, respectively. A 500-nL sample injection volume was used for each EIE in **figure 2.1D**. Even though we were able to detect BSA for all the experiments, BSA has a large molecular weight (66.5 kDa) causing a low S/N making it hard to extract protein intensity. Therefore, BSA is not used for protein intensity comparison (**figure 2.1A-C**).

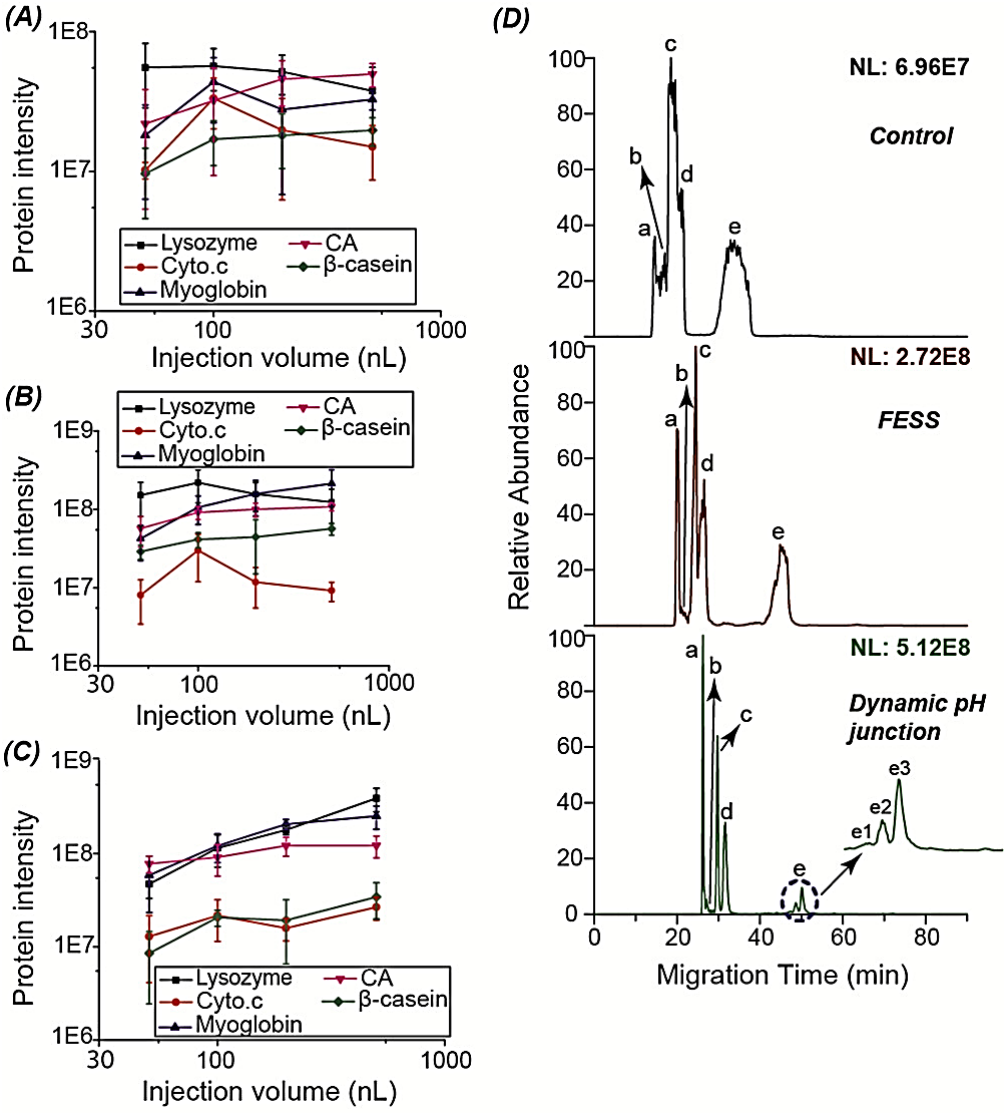


Figure 2.1. Protein intensity change vs various sample injection volumes (50-, 100-, 200-, and 500-nL) for (**A**) control, (**B**) FESS, and (**C**) dynamic pH junction. The error bars represent the standard deviations of protein intensity from triplicate CZE-MS analyses. (**D**) EIEs of the mixture of standard proteins from CZE-MS under the three different conditions: top panel is the control, middle panel is FESS, and bottom panel is dynamic pH junction. The proteins labeled in the electropherograms are (a) lysozyme, (b) cyto.c, (c) myoglobin, (d) CA, and (e) β-casein.

For the control experiments, the sample was dissolved in 5% acetic acid with 5% acetic acid as the BGE. **Figure 2.1A** shows the protein intensity change across 50-500 nL injection volumes for the control experiments. Four (except lysozyme) of the five proteins shows reasonably protein intensity increases from 50 to 100-nL. The average protein intensity increased by a factor

of 2. Four proteins (except Cyt. C) illustrated consistent protein intensity when the injection volume increased from 100 to 500-nL. There was <10% change in average protein intensity of the five proteins. Cyt. C. illustrated a unique trend in that the protein intensity was lower in the 500-nL injection volume compared to the 100-nL injection volume. One explanation for this is ionization suppression from BSA. Both BSA and Cyt. C comigrated out of the separation capillary together when the injection volume became 500-nL causing partial separation by CZE for the two proteins.

The FESS experiments were completed by dissolving the sample protein mixture in 20% acetonitrile (v/v) with 5% acetic acid (v/v) as the BGE. There was an average protein intensity increase of 2x when the injection volume increased from 50 to 100-nL. Four (except Cyt. C.) of the five protein illustrated a steady protein intensity when the injection volume increased from 100 to 500-nL. The average protein intensities for the sample increased roughly 20%. The data indicates that FESS can effectively concentrate protein molecules at low sample injection volumes (<100-nL, 5% of the total capillary volume), however when the injection volume increased from 100 to 500 -nL, there protein molecules could not concentrate directly on the capillary. Once again, the partial separation of BSA from Cyt. C caused ionization suppression issues resulting in Cyt. C. having a lower protein intensity at the higher injection volumes. The protein intensity for the FESS method was 2-3 x higher compared to the control at the same injection volume (**figure 2.1D**, 6.96E7 vs 2.72E8); the reason for this is that FESS does have the ability to stack protein molecules on the capillary more efficiently than the control. The FESS method also produced a better protein separation than the control due to its stacking performance. For the control experiments using a 500-nL injection volume, lysozyme, myoglobin, Cyt. C. and CA illustrated no separation from each other. In contrast to FESS

method using a 500-nL injection volume, the proteins produced a reasonable separation from each other with higher separation efficiency. For example, the control experiments produced only a 400 theoretical plate value for myoglobin, while the FESS method produced a theoretical late value of ~6600.

Dynamic pH junction method is executed by dissolving the sample within 50 mM ammonium bicarbonate (ABC, pH 8) with 5% acetic acid [(v/v) pH ~2.8]. **Figure 2.1C** shows the protein intensity change across the different injection volumes. There was a significant increase in protein intensity for all 5 proteins when the sample injection volume increased from 50 to 100-nL. There was an average protein intensity increase of roughly 2x. In addition, there was a significant protein intensity increase when the sample injection volume increased from 100 to 500-nL. The average protein intensity was 2x higher for the 500-nL injection volume compared to the 100-nL injection volume. The data here demonstrates the power of the dynamic pH junction based CZE-MS for increasing the loading capacity. Even at 500-nL (25% of the total capillary volume), the method could effectively concentration the protein directly on the capillary without loss in protein intensity. From 50 to 200-nL sample injection volumes, the average protein intensity for the dynamic pH junction method was comparable to the FESS method. However, when the sample injection volume increased to 500-nL the dynamic pH junction method produced an average protein intensity that was 80% better compared to FESS. The dynamic pH junction method also produced a better CZE separation than FESS (**figure 2.1D**). The FESS method could only partially separate myoglobin and CA (R=1), but with the dynamic pH junction method both myoglobin and CA could be baseline separated from each other (R=1.6). Dynamic pH junction even baseline separated three isoforms of B-casein with different masses. The masses that were used for the EIEs were 23983 Da (e3), 24022 Da (e2),

and 24092 Da (e1), at charge +23. Lastly, dynamic pH junction outperformed the FESS method in terms of separation efficiency. For example, FESS method produced a 6600 theoretical plate value for myoglobin and the dynamic pH junction method produced a 23,000 theoretical plate value. This data demonstrates dynamic pH junction surpassed the FESS method in both sample loading capacity and separation efficiency for the characterization of the six proteins within the standard protein mixture.

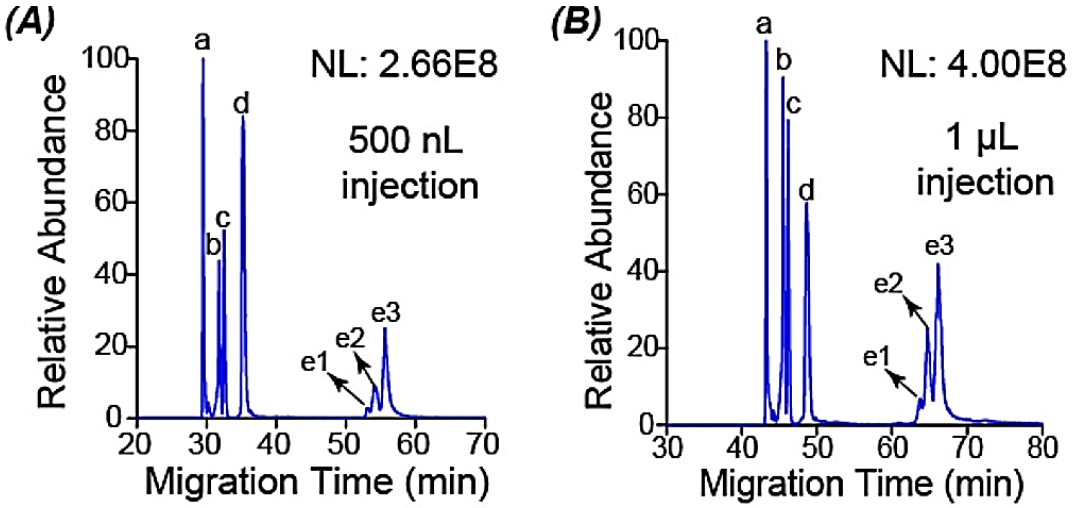
A calibration-curve was performed on the dynamic pH junction-based CZE-MS (**appendix figure 2.5**). The standard protein mixture stock solution (2.2 mg/mL) was diluted with 10 mM ABC (pH 8.0) by factors of 2, 6, 18, and 54 (1.1, 0.36, 0.12, 0.04 mg/mL), respectively. The sample injection volume was 500-nL per CZE-MS run. Three proteins, lysozyme, CA, and myoglobin were chosen for the calibration curve. The three proteins were detected, and baseline separated from each other in all CZE-MS. Correlations of 0.96-0.99 were produced for protein concentration and intensity, spanning a 30x dynamic concentration range illustrating the sensitivity of this technique. The data here demonstrates the potential of CZE-MS for quantitative top-down proteomics.

2.3.2 Optimization of the Dynamic pH Junction-Based CZE-MS

During the evaluation of FESS and dynamic pH junction, 10 mM ABC was used as the sample buffer for dynamic pH junction for all the experiments[20,38]. However, Imami et al. systematically evaluated the effect of the ABC salt concentration within the sample buffer on dynamic pH junctions stacking performance using peptide mixtures[25]. The authors increased the salt concentration from 10 to 200 mM, there was a consistent increase in peptide intensity until 100 mM[25]. Therefore, we evaluated the effect of increasing the salt concentration (5 to 20 mM ABC)

for the stacking performance of dynamic pH junction for intact proteins and noticed an increase of the protein intensity. Multiple studies have shown that when using CZE-MS for biomolecules that the salt concentration was 50 mM ABC[39-41]. We then used 50 mM ABC with 5% acetic acid for dynamic pH junction-based CZE-MS for top-down proteomics (**figure 2.2**). The standard protein mixture stock solution was diluted by a factor of 10 with 50 mM ABC. This sample was used for the remaining experiments.

The six proteins from the standard protein stock solution were baseline separated using a 500-nL sample injection and 1 uL sample injection (50% of the total capillary volume) (**figure 2.2A and 2.2B**). **Figure 2.2A** shows the standard protein mixture using a 500-nL sample injection volume, while **figure 2.2B** shows the standard protein mixture using a 1-uL sample injection volume. Both the 500-nL and 1-uL sample loading volume produced high separation efficiency using dynamic pH junction-based CZE-MS platform with 50 mM ABC as the sample buffer (**figure 2.2C**). For instance, the theoretical plate values ranged in 21,000 to 206,000 for the 500-nL sample loading volume and the theoretical plate values ranged in 30,000 to 292,000 for the 1-uL sample loading volumes. Myoglobin produced the highest theoretical plate value at 292,000 and 206,000 for 500-nL and 1-uL sample loading volumes, respectively. Based on the EIEs, the average protein intensity was 2.5x higher for the 1-uL sample loading capacity compared to the 500-nL loading capacity. The data indicates that using 50 mM as the sample buffer for dynamic pH junction-based CZE-MS system can effectively concentration proteins when even 50% of the total capillary volume has been filled.



(C)

Peak	Protein	N	
		500 nL	*1 µL*
a	Lysozyme	206,000	292,000
b	Cyto.c	101,000	187,000
c	Myoglobin	95,000	213,000
d	CA	31,000	41,000
e1	β-casein, peak e1	36,000	45,000
e2	β-casein, peak e2	21,000	46,000
e3	β-casein, peak e3	41,000	30,000

Figure 2.2. EIEs of the standard protein mixture dissolved in 50 mM NH_4HCO_3 (pH 8.0) analyzed by the dynamic pH junction-based CZE-MS with (**A**) 500 nL sample injection and (**B**) 1-µL sample injection. The theoretical plate value (N) of each protein was calculated based on the peak width and migration time of each protein in the EIEs. BSA was not extracted in the figures due to its low signal-to-noise ratio.

Comparing the average protein intensities from **figure 2.1D** (10 mM ABC as the sample buffer) and **figure 2.2A** (50 mM ABC as the sample buffer), it was found that 50 mM ABC as the sample buffer produced comparably average protein intensity as 10 mM ABC, but with 5x lower protein concentration. The data shows that 50 mM ABC as the sample buffer can produced sensitive separation of intact proteins using a lower concentration. Therefore, we used 50 mM ABC (pH 8) as the sample buffer for the remaining experiments.

Next, we investigated different BGEs (0.1–0.5% (v/v) formic acid (FA) and 5–10% (v/v) acetic acid). The sample injection was 500-nL per CZE-MS/MS run. 0.1 % (v/v) FA (pH 2.8) produced the highest protein intensity compared to 0.3% and 0.5% (v/v) FA (pH 2.3 and 2.1) (**appendix figure 2.6A**), while there was comparable intensity compared to 5 and 10% (v/v) acetic acid (pH ~2.4 and 2.2) as shown in **appendix figure 2.6B and 2.6C**. However, both 5% and 10% (v/v) acetic acid produced a longer separation window and protein migration time compared to 0.1% (v/v) FA as the BGE for the standard protein mixture. Two potential reasons can explain this phenomenon: (1) 5% and 10% (v/v) acetic acid have lower pHs and viscosity than FA reducing the EOF even more within the LPA-coated capillary, and (2) the lower acidic conditions lead to more protein unfolding causing a larger hydrodynamic radius of the proteins, resulting in a slower protein migration within the separation capillary under the electric field. We tested the electroosmotic mobility of 10% (v/v) of acetic acid and 0.1% (v/v) FA on the LPA-coated capillary based on methods reported in literature[42,43]. The electroosmotic mobility in 10% (v/v) acetic acid BGE was lower than that in 0.1% (v/v) formic acid BGE (6.8E–6 $cm^2\ V^{-1}\ s^{-1}$ vs 1.1E–5 $cm^2\ V^{-1}\ s^{-1}$).

Based on the data obtained from the electroosmotic mobility testing, we chose 5-10% (v/v) and 50 mM ABC as the optimized BGE and sample buffer. We investigated the reproducibility of the optimized dynamic pH junction-based CZE-MS platform using the standard protein mixture. Each CZE-MS run had a 500-nL sample injection volume. The system produced reproducible separation profiles and detection of the standard protein mixture during a 16-hr continuous run (11 CZE-MS runs). The relative standard deviations for migration time and protein intensity were >7% and >10%, respectively (**Appendix Table 2.1**). Based on our experiments, one LPA-

coated capillary can be used for 1 week of continuous runs for consistent separation and detection of intact proteins.

2.3.3 Single-Shot Top-down Proteomics with CZE-MS/MS

We then applied the optimized CZE-MS/MS platform for top-down proteomic analysis is of an *E. coli* proteome. We used 50 mM ABC as the sample buffer and 10% (v/v) acetic acid as the BGE for the remaining experiments. A 2 mg/mL *E. coli* protein sample was used. The injection volume was either 500-nL or 1-uL for sample injection volumes. A Q-Exactive HF mass spectrometer was used with either a top 3 or top 8 DDA.

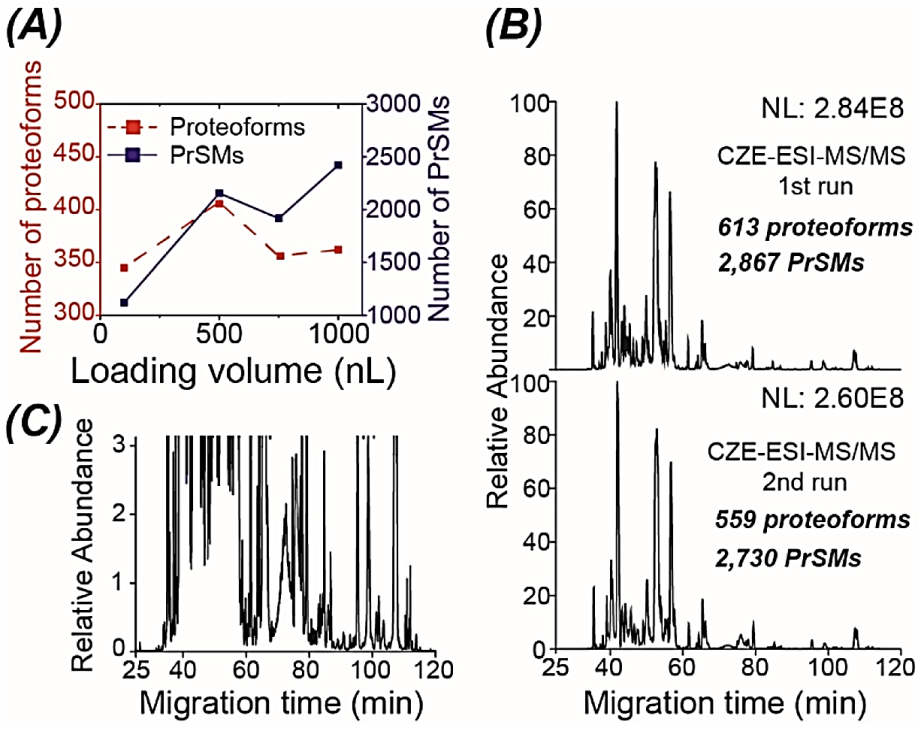


Figure 2.3. Top-down proteomics of *E. coli* using CZE-MS/MS. (**A**) Sample loading volume vs the number of proteoform IDs and the number of proteoform-spectrum matches (PrSMs). (**B**) Electropherograms of the *E. coli* protein sample analyzed by top-down based CZE-MS/MS in duplicate runs. For the CZE-MS experiments, 20 kV was applied at the injection end for separation. (**C**) The zoom-in electropherogram of the *E. coli* protein sample showing the separation window from the 1st run CZE-MS/MS in (**B**).

We investigated the effect of the sample loading volume on the identification (ID) and PrSMs (**figure 2.3A**). The top 3 DDA with a 500-nL sample loading capacity produced the highest number of proteoform IDs (407) at the 1% spectrum level FDR. When the same loading capacity was increased to 1-uL, the number of proteoform IDs decreased but remained relatively consistent with the 500-nL sampling loading capacity. The separation voltage was then lower from 30 to 20 kV to cause a lower electric field to be felt across the capillary, and therefore the electrophoretic mobility of the analyte should be lower resulting in a wider separation window. 468 proteoforms were identified with the 20 kV setting, resulting in a 15% increase in proteoform identifications compared to the 30 kV setting.

We further investigated the 20 kV setting and performed CZE-MS/MS analysis of the *E. coli* sample in duplicates using a 500-nL sample loading capacity (**figure 2.3B and 2.3C**). We then evaluated using a higher N (3 vs 8) for the DDA method. The top 8 DDA method identified 586 ± 38 proteoforms (n = 2) and 2798 ± 97 PrSMs (n = 2) with single shot CZE-MS/MS after filtering with a 1 % spectrum-level FDR. This proteoform number is 3x higher than that reported in literature for single-shot CZE-MS/MS (586 vs 140-180)[8,10]. The data here demonstrates the potential of CZE-MS/MS for large-scale top-down proteomics. Next, we evaluated the molecular weight (MW) distribution of the proteoforms resulting from the single-shot CZE-MS/MS platform using the top 8 DDA method (**appendix figure 2.7**). The MW distribution ranged from 2 kDa to 24 kDa, where ~70% of the proteoforms had MW smaller than 10 kDa. Top-down proteomics still has issues with identifying proteoforms larger than 30 kDa due to two reason: (1) the coelution of smaller proteoforms impacts the identification of the larger proteoforms, and (2) the S/N ratio of the orbitrap decreases with increasing molecular weight.

We further investigated the nearly 600 identified proteoforms from the *E. coli* proteome using the optimized single-shot CZE-MS/MS platform. The nearly 600 proteoforms that were identified corresponded to ~200 genes, and on average there were 3 proteoforms that were identified per gene (**figure 2.4A**). There was one proteoform/gene for about *100 E. coli* genes, 2–5 proteoforms/gene for about 80 genes, and 6–44 proteoforms/gene for about 20 genes. *E. coli* genes hdeA, acpP, and ybgS had the most proteoforms/gene identified at 44, 30, and 21, respectively. According to PaxDb, these three genes have the most abundant proteoforms for the *E. coli* proteome. There was an average of ~72% of proteoforms that produced a mass shift from the duplicate CZE-MS/MS runs. There was a total of 870 mass shifts that were detected, **figure 2.4B** shows the distribution of mass shifts. These mass shift events correspond to PTMs, such as oxidation (+16 Da) and acetylation (+42 Da). The CZE-MS/MS system also identified N-terminal methionine excision and signal peptide removal of proteins. **Figure 2.4C** shows the sequence pattern of the detected uncharacterized protein YggL with an N-terminal truncation and **figure 2.4D** shows the sequence pattern for the 30S ribosomal protein S17 that had a N-terminal methionone excisions. Both proteoforms show fragmentation within the termini and middle parts leading to 40 fragment ions.

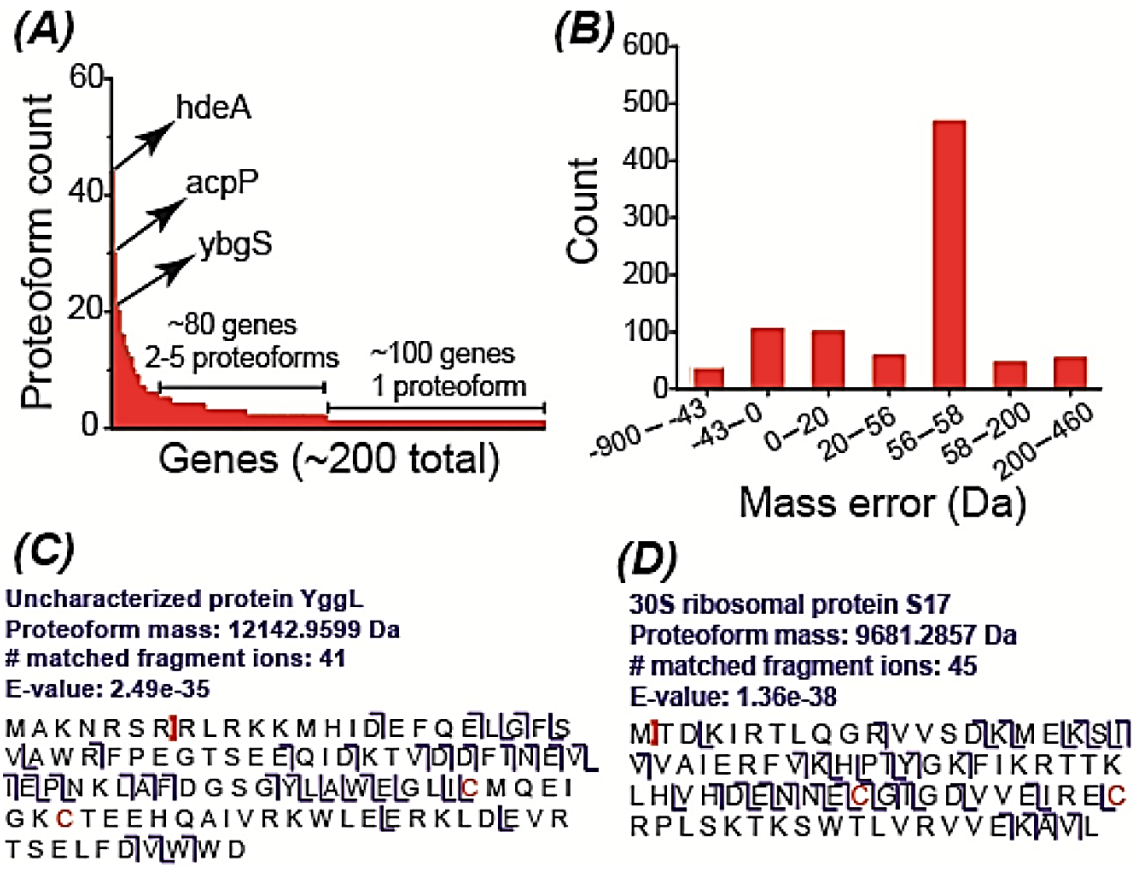


Figure 2.4. (**A**) Distribution number of identified proteoforms from each *E. coli* gene. (**B**) Detected mass error distribution from the identified proteoforms. (**C and D**) Sequence and fragmentation pattern of two identified proteins. Carbamidomethylation sites on cysteines are shown in red. The single-shot CZE-MS/MS data in figure 3 B was used for these analyses.

There are three reasons that can be attributed to the proteoform identification improvement from single-shot CZE-MS/MS: (1) the large sample loading capacity (0.5 to 1uL) of the optimized dynamic pH junction-based CZE-MS/MS platform allowed for more sample material to be used to acquire a large number of proteoform identification, (2) the longer separation window (90-mins) of the optimized dynamic pH junction-based CZE-MS/MS platform allowed for the acquisition of more tandem mass spectra resulting in higher proteoform identification, resulting in the a separation window 3x longer than previously reported[8,10], and (3) the high peak capacity (280 based on **figure 2.3B**) of the optimized dynamic pH junction-based CZE-MS/MS platform, resulting in a peak capacity that is 2-3x higher than in previous studies [8,10].

2.4 Conclusions

This study systemically evaluated CZE-MS for large-scale top-down proteomics. The optimized dynamic pH junction-based single-shot CZE-MS /MS platform for large-scale top-down proteomics produced a uL sample loading capacity, 90-min separation window and a 280-peak capacity, resulting in ~600 proteoform identifications from an *E. coli* proteome. The proteoform ID number is 3x higher than what was reported in literature and is equivalent to the ID number that was reported for single-shot RPLC-MS/MS when a 21 T FT-ICR mass spectrometer was used[45].

Even with the significant improvements provided in this study for single-shot CZE-MS/MS for top-down proteomics, it is still in the early development stage. Thus far, the 600 proteoforms that have been identified represent the largest proteoform identification for top-down proteomics using CZE-MS/MS. However, the state-of-the-art RPLC-MS/MS platforms are still ahead and can identify thousands of proteoforms from mammalian cell lines[45,48-52]. Further improvement on the CZE-MS/MS platform, for both the sample loading capacity and separation window, is still needed to reach RPLC-MS/MS identification numbers for large-scale top-down proteomics. Two potential improvements for the CZE-MS/MS platforms would be to increase the separation voltage to 60 kV or to increase the length of the separation capillary so more sample material can be used to significantly improve the CZE-MS/MS platform.

2.5 Acknowledgements

We thank Prof. Heedeok Hong's group at Department of Chemistry, Michigan State University, for kindly providing the Escherichia coli cells for our experiments. This research was funded by

Michigan State University. Q.K. and X.L. were supported by the National Institute of General Medical Sciences, National Institutes of Health (NIH), through Grant R01GM118470.

APPENDIX

Table 2.1. Reproducibility data from the 11 consecutive CZE-MS runs using the standard protein mixture.

	Relative standard deviations (%)	
	Migration time	Intensity
Lysozyme	6.6	15.4
BSA	6.0	9.2
Cyto.c	5.9	15.4
Myoglobin	5.4	8.1
CA	4.4	8.7
β-casein	5.4	12.2

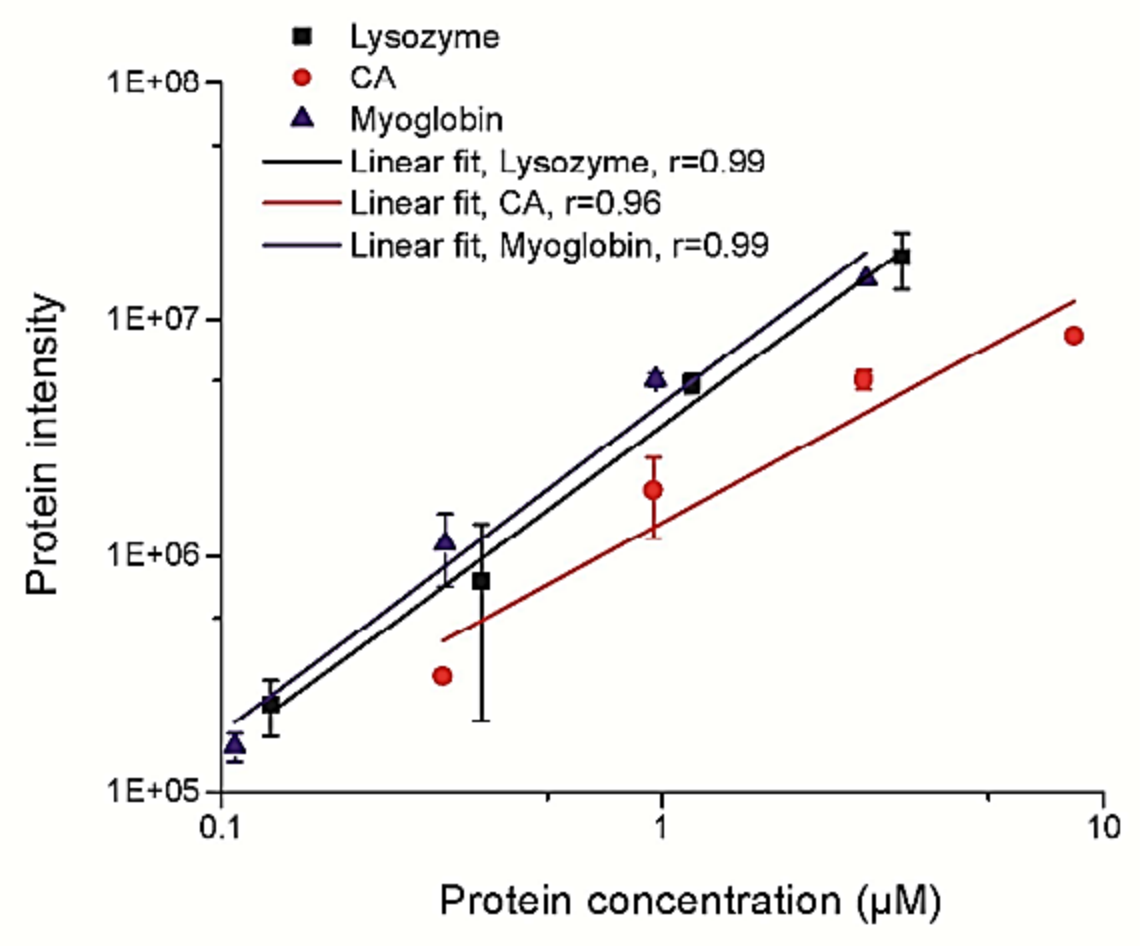


Figure 2.5. Calibration curve of the protein concentration and protein intensity for lysozyme, CA and myoglobin. The errors bars are the standard deviations of protein intensity from duplicate CZE-MS runs. An LTQ-XL mass spectrometer was used for the experiments.

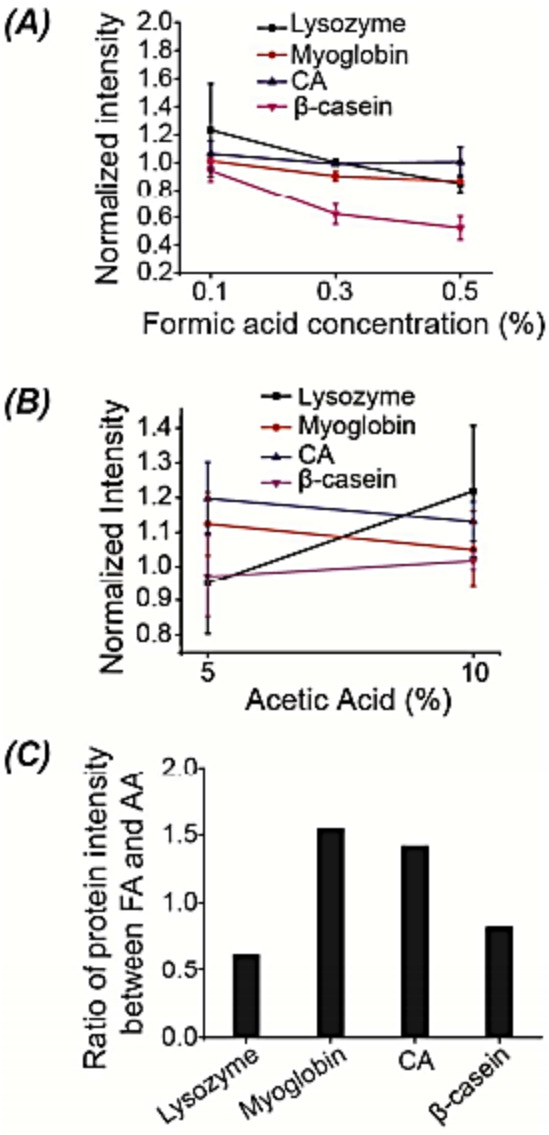


Figure 2.6. Protein intensity from dynamic pH junction based CZE-MS utilizing various BGEs. BGEs. (**A**) Different formic acid concentrations (0.1-0.5% (v/v)). (**B**) Different acetic acid concentrations (5% and 10% (v/v)). The protein intensity from formic acid BGEs were normalized to 0.1% (v/v) formic acid; the intensity from acetic acid BGEs were normalized to 5% (v/v) acetic acid. (**C**) The acid concentrations that produced the highest protein intensities (0.1% (v/v) formic acid and 5% (v/v) acetic acid) were compared. LTQ-XL mass spectrometer was used for the experiments while the errors bars represent the standard deviations of protein intensity from duplicate CZE-MS runs.

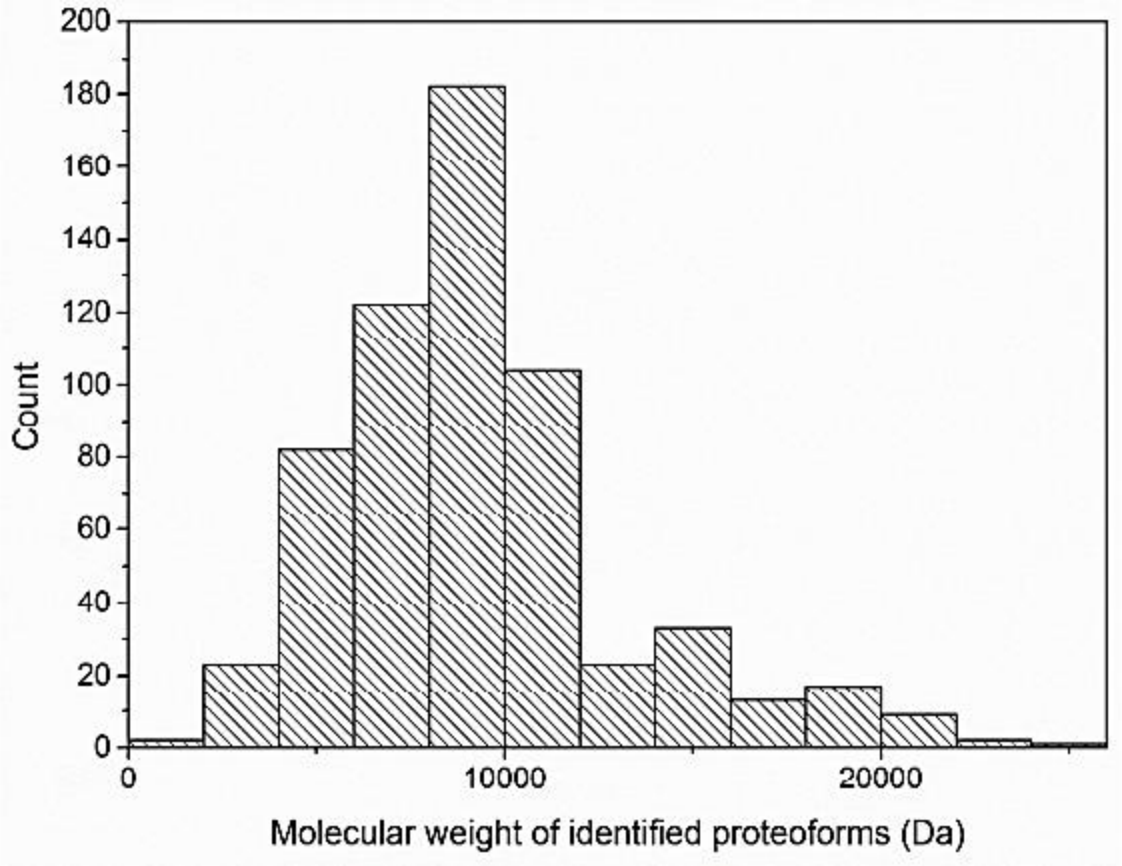


Figure 2.7. Distribution of the identified proteoform mass from single-shot CZE-MS/MS

REFERENCES

REFERENCES

1. Domínguez-Vega, E.; Haselberg, R.; Somsen, G. W. Capillary Zone Electrophoresis-Mass Spectrometry of Intact Proteins. Methods Mol. Biol. 2016, 1466, 25–41.

2. Haselberg, R.; de Jong, G. J.; Somsen, G. W. CE-MS for the Analysis of Intact Proteins 2010-2012: CE and CEC. *Electrophoresis* **2013**, 34 (1), 99–112.

3. Jorgenson, J. W.; Lukacs, K. D. Capillary Zone Electrophoresis. *Science* **1983**, 222 (4621), 266–272.

4. Harstad, R. K.; Johnson, A. C.; Weisenberger, M. M.; Bowser, M. T. Capillary Electrophoresis. *Anal. Chem*. **2016**, 88 (1), 299–319.

5. Valaskovic, G. A.; Kelleher, N. L.; McLafferty, F. W. Attomole Protein Characterization by Capillary Electrophoresis-Mass Spectrometry. *Science* **1996**, 273 (5279), 1199–1202.

6. 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*. **2013**, 85 (12), 5989–5995.

7. 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.* **2014**, 86 (10), 4873–4878.

8. 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.* **2016**, 15 (10), 3679–3685.

9. 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* **2014**, 14 (10), 1158–1164.

10. Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavallée-Adam, M.; Yates, J. R., 3rd. Sheathless Capillary Electrophoresis-Tandem Mass Spectrometry for Top-down Characterization of Pyrococcus Furiosus Proteins on a Proteome Scale. *Anal. Chem.* **2014,** 86 (22), 11006–11012.

11. Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates, J. R., 3rd. In-Line Separation by Capillary Electrophoresis Prior to Analysis by Top-down Mass Spectrometry Enables Sensitive Characterization of Protein Complexes. *J. Proteome Res.* **2014**, 13 (12), 6078–6086.

12. 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.* **2013**, 85 (4), 2289–2296.

13. 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-B1. *Anal. Chem.* **2016**, 88 (2), 1138–1146.

14. Faserl, K.; Sarg, B.; Maurer, V.; Lindner, H. H. Exploiting Charge Differences for the Analysis of Challenging Post-Translational Modifications by Capillary Electrophoresis-Mass Spectrometry. *J. Chromatogr. A.* **2017**, 1498, 215–223.

15. 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* **2013,** 12 (9), 2640–2656.

16. Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. Simplified Capillary Electrophoresis Nanospray Sheath-Flow Interface for High Efficiency and Sensitive Peptide Analysis: Capillary Electrophoresis Electrospray Interface. *Rapid Commun. Mass Spectrom.* **2010**, 24 (17), 2554–2560.

17. Moini, M. Simplifying CE-MS Operation. 2. Interfacing Low-Flow Separation Techniques to Mass Spectrometry Using a Porous Tip. *Anal. Chem*. **2007**, 79 (11), 4241–4246.

18. Yang, L.; Lee, C. S.; Hofstadler, S. A.; Pasa-Tolic, L.; Smith, R. D. Capillary Isoelectric Focusing-Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Protein Characterization. *Anal. Chem.* **1998**, 70 (15), 3235–3241.

    Jensen, P. K.; Pasa-Tolić, L.; Anderson, G. A.; Horner, J. A.; Lipton, M. S.; Bruce, J. E.; Smith, R. D. Probing Proteomes Using Capillary Isoelectric Focusing-Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **1999**, 71 (11), 2076–2084.

19. Jensen, P. K.; Pasa-Tolić, L.; Anderson, G. A.; Horner, J. A.; Lipton, M. S.; Bruce, J. E.; Smith, R. D. Probing Proteomes Using Capillary Isoelectric Focusing-Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem*. **1999**, 71 (11), 2076–2084.

20. 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* **2017**, 142 (12), 2118–2127.

21. Aebersold, R.; Morrison, H. D. Analysis of Dilute Peptide Samples by Capillary Zone Electrophoresis. J. Chromatogr. 1990, 516 (1), 79–88.

22. Britz-McKibbin, P.; Chen, D. D. Selective Focusing of Catecholamines and Weakly Acidic Compounds by Capillary Electrophoresis Using a Dynamic PH Junction. *Anal. Chem.* **2000**, 72 (6), 1242–1252.

23. Wang, L.; MacDonald, D.; Huang, X.; Chen, D. D. Y. Capture Efficiency of Dynamic PH Junction Focusing in Capillary Electrophoresis: CE and CEC. *Electrophoresis* **2016**, 37 (9), 1143–1150.

24. Cao, C.-X.; Fan, L.-Y.; Zhang, W. Review on the Theory of Moving Reaction Boundary, Electromigration Reaction Methods and Applications in Isoelectric Focusing and Sample Pre-Concentration. *Analyst* **2008**, 133 (9), 1139–1157.

25. Imami, K.; Monton, M. R. N.; Ishihama, Y.; Terabe, S. Simple On-Line Sample Preconcentration Technique for Peptides Based on Dynamic PH Junction in Capillary Electrophoresis-Mass Spectrometry. *J. Chromatogr. A* **2007**, 1148 (2), 250–255.

26. Ptolemy, A. S.; Britz-McKibbin, P. New Advances in On-Line Sample Preconcentration by Capillary Electrophoresis Using Dynamic PH Junction. *Analyst* **2008**, 133 (12), 1643–1648.

27. 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.* **2015**, 14 (5), 2312–2321.

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* **2016**, 146, 839–843.

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 Engl. 2013, 52 (51), 13661–13664.

30. qkou. TopFD http://proteomics.informatics.iupui.edu/software/topfd/ (accessed Feb 2, 2021).

31. 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.

32. Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins: A Combinatorial Approach: A Combinatorial Approach. *Mol. Cell. Proteomics* **2010**, 9 (12), 2772–2782.

33. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, 24 (21), 2534–2536.

34. Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods.* **2007**, 4 (3), 207–214.

35. Sun, L.; Hebert, A. S.; Yan, X.; Zhao, Y.; Westphall, M. S.; Rush, M. J. P.; Zhu, G.; Champion, M. M.; Coon, J. J.; Dovichi, N. J. Over 10,000 Peptide Identifications from the HeLa Proteome by Using Single-Shot Capillary Zone Electrophoresis Combined with Tandem Mass Spectrometry. *Angew. Chem. Int. Ed Engl.* **2014**, 53 (50), 13931–13933.

36. Simpson, S. L., Jr; Quirino, J. P.; Terabe, S. On-Line Sample Preconcentration in Capillary Electrophoresis. Fundamentals and Applications. *J. Chromatogr. A.* **2008**, 1184 (1–2), 504–541.

37. Wu, S.; Lourette, N. M.; Tolić, N.; Zhao, R.; Robinson, E. W.; Tolmachev, A. V.; Smith, R. D.; Pasa-Tolić, L. An Integrated Top-down and Bottom-up Strategy for Broadly Characterizing Protein Isoforms and Modifications. *J. Proteome Res.* **2009**, 8 (3), 1347–1357.

38. Zhu, G.; Sun, L.; Heidbrink-Thompson, J.; Kuntumalla, S.; Lin, H.-Y.; Larkin, C. J.; McGivney, J. B., 4th; Dovichi, N. J. Capillary Zone Electrophoresis Tandem Mass Spectrometry Detects Low Concentration Host Cell Impurities in Monoclonal Antibodies: Proteomics and 2-DE. *Electrophoresis* **2016**, 37 (4), 616–622.

39. Busnel, J.-M.; Schoenmaker, B.; Ramautar, R.; Carrasco-Pancorbo, A.; Ratnayake, C.; Feitelson, J. S.; Chapman, J. D.; Deelder, A. M.; Mayboroda, O. A. High Capacity Capillary Electrophoresis-Electrospray Ionization Mass Spectrometry: Coupling a Porous Sheathless Interface with Transient-Isotachophoresis. *Anal. Chem.* **2010**, 82 (22), 9476–9483.

40. Faserl, K.; Kremser, L.; Müller, M.; Teis, D.; Lindner, H. H. Quantitative Proteomics Using Ultralow Flow Capillary Electrophoresis-Mass Spectrometry. *Anal. Chem.* **2015,** 87 (9), 4633–4640.

41. Wang, Y.; Fonslow, B. R.; Wong, C. C. L.; Nakorchevsky, A.; Yates, J. R., 3rd. Improving the Comprehensiveness and Sensitivity of Sheathless Capillary Electrophoresis-Tandem Mass Spectrometry for Proteomic Analysis. *Anal. Chem.* **2012,** 84 (20), 8505–8513.

42. Williams, B. A.; Vigh, G. Fast, Accurate Mobility Determination Method for Capillary Electrophoresis. *Anal. Chem*. **1996**, 68 (7), 1174–1180.

43. Zhang, Z.; Peuchen, E. H.; Dovichi, N. J. Surface-Confined Aqueous Reversible Addition-Fragmentation Chain Transfer (SCARAFT) Polymerization Method for Preparation of Coated Capillary Leads to over 10 000 Peptides Identified from 25 Ng HeLa Digest by Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry. *Anal. Chem.* **2017**, 89 (12), 6774–6780.

44. Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H. 2016 Update of the PRIDE Database and Its Related Tools. *Nucleic Acids Res.* **2016**, 44 (D1), D447-56.

45. 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.* **2017**, 16 (2), 1087–1096.

46. Nguyen, A.; Moini, M. Analysis of Major Protein-Protein and Protein-Metal Complexes of Erythrocytes Directly from Cell Lysate Utilizing Capillary Electrophoresis Mass Spectrometry. *Anal. Chem*. **2008**, 80 (18), 7169–7173.

47. Belov, A. M.; Viner, R.; Santos, M. R.; Horn, D. M.; Bern, M.; Karger, B. L.; Ivanov, A. R. Analysis of Proteins, Protein Complexes, and Organellar Proteomes Using Sheathless Capillary Zone Electrophoresis - Native Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2017**, 28 (12), 2614–2634.

48. 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. 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), 254–258.

49. 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.* **2017**, 16 (2), 609–618.

50. 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.* **2016**, 15 (3), 976–982.

51. 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.* **2017,** 89 (10), 5467–5475.

52. Valeja, S. G.; Xiu, L.; Gregorich, Z. R.; Guner, H.; Jin, S.; Ge, Y. Three Dimensional Liquid Chromatography Coupling Ion Exchange Chromatography/Hydrophobic Interaction Chromatography/Reverse Phase Chromatography for Effective Protein Separation in Top-down Proteomics. *Anal. Chem*. **2015**, 87 (10), 5363–5371.

[2]CHAPTER 3: Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples

3.1 Introduction

For large-scale TDP, RPLC-ESI-MS/MS is commonly used. Using RPLC-ESI-MS/MS for the characterization of proteoforms from complex proteomes, extreme progress has been made for the identification and quantification of thousands of proteoforms. However, there are issues that remain for using RPLC-ESI-MS/MS for TDP, such as high-capacity separation of complex proteome mixtures and large-scale top-down characterization of proteoforms from mass limited samples[1-9]. Currently, to achieve thousands of proteoform identifications from complex proteomes using RPLC-ESI-MS/MS, hundreds of micrograms of protein materials are required, therefore applying this technique to mass limited proteomes remain a challenge[1-9]. Therefore, alternative TDP systems with better sensitivity for mass-limited proteome samples.

Capillary zone electrophoresis (CZE) is a simple and highly efficient analytical separation technique that separates based on their electrophoretic mobility when under the influence of an electric field[10]. Over the last 20 years, CZE-MS has been gaining traction as a powerful alternative to RPLC-MS for the characterization of intact proteins due to its highly efficient

[2] This chapter was adapted with permission from 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. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (8), 1435–1445.

separation and detection[11-20]. For example, 250 proteoforms were identified from recombinant human erythropoietin using CZE-MS; the system produced high separation efficiency and detection. Han et al. compared RPLC-MS to CZE-Ms for the characterization of the Dam1 complex, it was reported that CZE-MS produced similar signal-to-noise (S/N) ratios as RPLC-MS but consumption 100x less sample (250-ng vs 2.5-ng)[13]. The first reported analysis of intact proteins using CZE-MS was in 1996 where the authors produced attomole detection[11].

Two major issues have impeded CZE-MS/MS for large-scale top-down proteomics of complex proteome samples: (1) small sample loading capacity, and (2) narrow separation window. We have recently drastically improved the sample loading capacity and the separation window for CZE-MS/MS using a one-meter long LPA-coated separation capillary and a highly effective sample stacking method for preconcentration of sample directly onto the capillary[24-26]. Using the optimized dynamic pH junction based CZE-MS/MS platform, we increased the sample loading capacity to 500-nL with a 90-min separation window for the top-down proteomics of an *E. coli* proteome. This platform resulted in the identification of ~600 proteoforms in a single run. McCool et al. couple offline fractionation of size exclusion chromatography (SEC)-RPLC and utilized the optimized dynamic pH junction based CZE-MS/MS platform for the identification of over 6,000 proteoforms from an *E. coli* proteome using 1 mg/mL as the sample starting material.

We thought to boast the sample loading capacity and separation window of the CZE-MS/MS platform by building upon our previous and employing a much longer separation capillary compared to before (1.5-m vs 1-m). Employing a longer capillary will lead to a lower electric field that is felt across the capillary, resulting in a considerably lower electrophoretic velocity of the proteoforms. This will result in a longer separation window, producing more time for the mass spectrometer to acquire more tandem mass spectra of the proteoforms for a larger

identification number. Besides the advantage of a longer separation window, a larger sample loading capacity would also follow allowing for more sample to be injected onto the separation capillary without the loss in separation efficiency. In this study, we investigated three items. First, how the separation performance of CZE-MS/MS is affected by comparing the 1-m long separation capillary to the 1.5-m long separation capillary. The reproducibility and sample loading volume were also investigated using the 1.5-m long separation capillary for a standard protein mixture. Next, large-scale top-down proteomics using the CZE-MS/MS platform with a 1.5-m long capillary utilizing 1 and 0.100 ug as the starting material was evaluated. Third, quantitative top-down proteomics of two regions (cerebellum and optic tectum) of Zebrafish brain was done using the CZE-MS/MS platform using a 1.5-m long separation capillary.

3.2 Experimental

3.2.1 Materials and Reagents

Acrylamide was purchased from Acros Organics (NJ, USA). Standard proteins, ammonium bicarbonawasH4HCO3), urea, dithiothreitol (DTT), iodoacetamide (IAA), and 3-(Trimethoxysilyl)propyl methacrylate was 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.

3.2.2 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 NH4HCO3 (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 18000g 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 reconstituted in 50 mM NH4HCO3 (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 18000g 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 NH4HCO3 (pH 8.0) three times to remove urea. Finally, the proteins from the Cb and Teo regions were reconstituted in 50 mM NH4HCO3

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.

3.2.3 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.

3.2.4 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].

3.3 Results and Discussions

3.3.1 Evaluation of the CZE-MS system with a 1.5-m long separation capillary using a standard protein mixture

First, we compared a 1-m and 1.5-m separation capillary using a standard protein mixture concerning the separation window and protein intensity (**figure 3.1a**). The standard protein mixture consisted of myoglobin (myo, 16.9 kDa), carbonic anhydrase (CA, 29 kDa), and bovine serum albumin (BSA, 66.5 kDa). Each CZE-MS run had a sample injection volume of 500-nL. Both the 1-m and 1.5-m separation capillary baseline separated the three proteins; however, the 1.5-m separation capillary produced a 2x longer separation window compared to the 1-m separation window (11-mins vs 5-mins). However, the proteins within the 1.5-m separation capillary migrated slower, which is indicated by their longer migration time. It took ~70-mins for

the first protein (BSA) peak to migrate out of the 1.5-m separation capillary compared to ~30-mins for the 1-m separation capillary. The 1.5-m separation capillary produced a base peak intensity that was 2-fold lower compared to the 1-m separation capillary, because the protein diffused 2x longer within the 1.5-m separation capillary. After the CA peak is a strong impurity peak that was identified based on MS/MS data as superoxide dimutase (SD) (**appendix figure 3.7**). In the 1-m separation capillary, the intensity of SD was lower than that of CA and myo, which was different in the 1.5-m separation capillary. The replicated runs using the 1-m separation capillary produced inconsistent protein intensities, while the 1.5-m separation capillary produce a more constant protein intensity (**figure 3.1b**). One explanation of this phenomenon is that the smaller peak widths produced in the 1-m separation capillary data caused a fewer number of data points across the peak resulting in the inconsistent protein intensities.

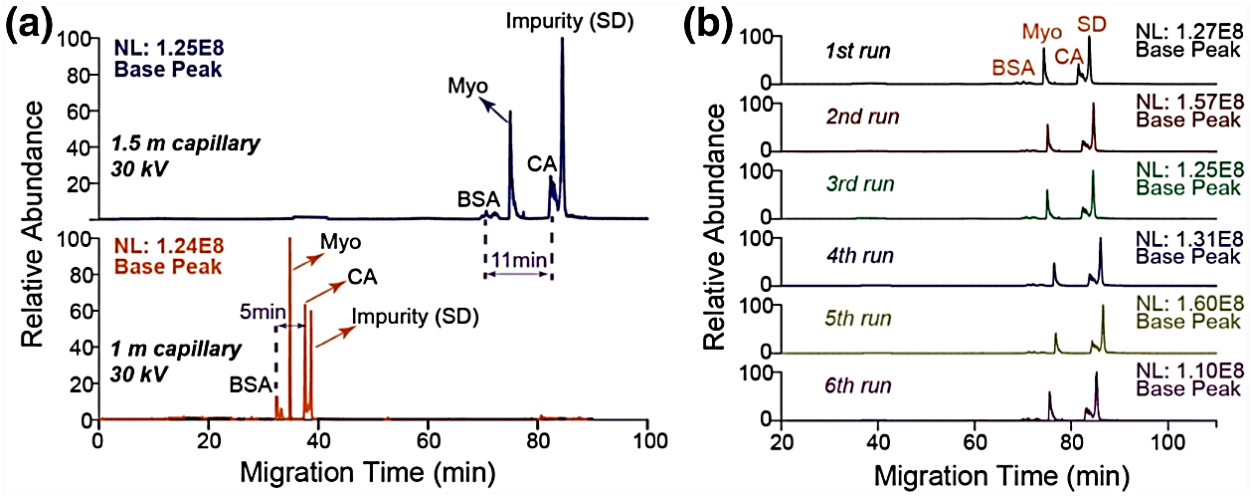


Figure 3.1. Comparison of the 1-m and 1.5-m separation capillary for CZE-MS analyses of a standard protein mixture. (**a**) Electropherograms of the standard protein mixture using CZE-MS with a 1.5-m capillary (top panel) and a 1-m capillary (bottom panel). (**b**) Electropherograms of the standard protein mixture using the 1.5-m separation capillary CZE-MS analyses with 1.5-m LPA-coated separation capillary in sextuplicate.

Next, we tested the reproducibility of the separation for the CZE-MS platform using the 1.5-m separation capillary using the standard protein mixture (**figure 3.1b**). The separation profiles

showed good reproducibility for the CZE-MS platform. The relative standard deviations (RSD) for the protein intensity were <25% and <2% for migration time.

Three sample different sample loading capacities were then tested on the 1.5-m separation capillary using the standard protein mixture. The CZE-MS platform used the sample stacking method dynamic pH junction as described in our previous paper. The protein mixture was dissolved in 50 mM ammonium bicarbonate (ABC, pH ~8.0), while the BGE for the CZE separation was 10% acetic acid (pH ~2.2). The sample loading volumes that were tested with the standard protein mix 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 3.2**). The separation window of the platform for the standard protein mixture was boasted by over 100% when the sample loading volume increased from 0.5-uL to 2-uL (**figure 3.2a**). As the sample loading volume increased, the migration time of the protein also increased (**figure 3.2b**). An explanation for the increased separation window and migration time is as follows. During the beginning of the CZE separation, the protein are being concentrated directly on the capillary based on the dynamic pH junction method[35,36]. During the dynamic pH junction method, the sample zone titrates due to the differences between the basic sample zone and acidic BGE causing analyte stacking between the two zones; once the titration of the sample zone completes, the normal CZE separation continues. As the sample loading volume increase, the sample zone titration time also increased causing a longer sample stacking time before the normal CZE separation continued. This phenomenon is what led to the longer separation window and protein migration time. **Figure 3.2C** illustrates the protein intensity change vs sample loading volume. As the sample loading volume increased from 0.5 to 1.0-uL, the average protein intensity was boasted by 3-folds. When the sample loading volume increased

form 1-uL to 2-uL, the average protein intensity increased by ~30%. The data here demonstrates that even at 2-uL sample loading volume that the dynamic pH junction based-CZE-MS/MS using the 1.5-m separation capillary could still effectively concentration the protein sample.

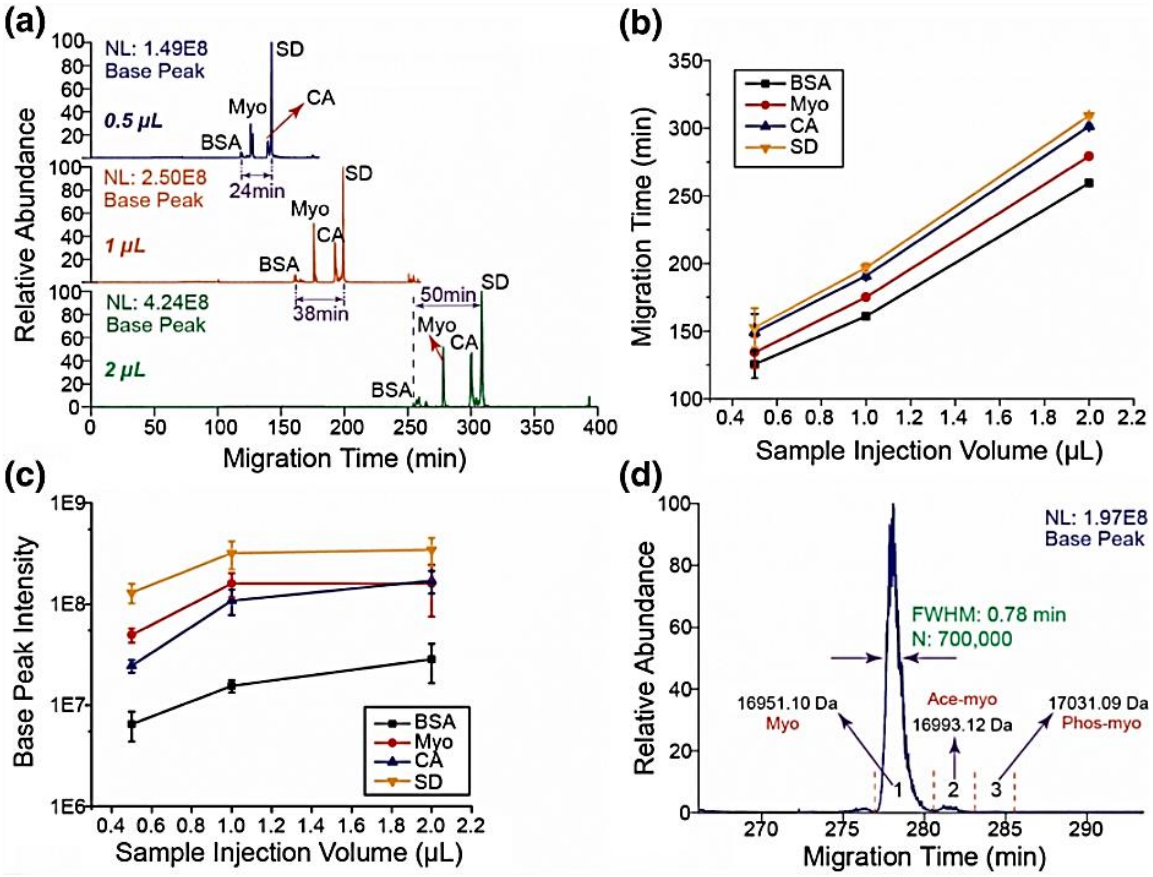


Figure 3.2. CZE-MS analyses of a standard protein mixture using different sample loading volumes (0.5-µL, 1-µL, and 2-µL). A 1.5-m LPA-coated separation capillary was used for all the CZE-MS runs in duplicate. (**a**) Electropherograms of the standard protein mixture with the three different sample loading 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.

The CZE-MS platform using the 1.5-m separation capillary produced extraordinarily high separation efficiency. For example, myoglobin produced theoretical plate values of 200,000,

700,000, and 1,000,000 for the 0.5, 1, and 2-uL sample loading volumes (**figure 3.2d**). The CZE-MS platform also detected, and baseline resolved three different proteoforms of myoglobin that had over 100-fold concentration dynamic range using the 2-uL sample injection volume. **Figure 3.2d** illustrates the three different forms of myoglobin that were detected, the three forms are labeled 1, 2, and 3. The three proteoforms of myoglobin that were detected are as follows: (1) myoglobin without PTMs (15951.10 Da), (2) acetylated myoglobin (16993.12 Da), and (3) phosphorylated myoglobin (17031.09 Da). The data here was also confirmed using bottom-up proteomics using a digested myoglobin sample; the bottom-up data also detected the acetylated and phosphorylation forms of myoglobin peptides with high confidence (**appendix figure 3.8**). There is also no experimental evidence of the acetylated and phosphorylated myoglobin (equine) reported within the UniProt database.

3.3.2 Top-down proteomics of *E. coli* cells using single-shot CZE-MS/MS with a 1.5-m long separation capillary

Due to the longer separation window and migration time of the standard protein mixture (**figure 3.2a and 3.2c**) discussed in the earlier section, the time that required for one single-shot CZE-MS using the 1.5-m separation capillary was significantly increased. The instrument time needed to be controlled therefore for the analysis of the *E. coli* and zebrafish brains, the sample injection volume was kept at 500-nL for all CZE-MS/MS runs using the 1.5-m LPA-coated separation capillary.

Figure 3a shows the separation profile of the *E. coli* proteome; the CZE-MS/MS platform produced reproducible detection and separation profiles. The CZE-MS/MS platform using the 1.5-m separation capillary produced ~180 min separation window (**figure 3.3b**) This separation

window is ~100% wider than the one that was produced in our previous work using the 1-m LPA-coated separation capillary[25]. Using 1-ug of *E. coli* proteome that was injected onto the capillary, the single-shot CZE-MS/MS platform identified 804 ± 10 proteoforms and 266 ± 6 proteins (n = 3) (**figure 3.3c**). Compared to our previous work with the 1-m LPA coated capillary, the proteoform and protein identification were boasted by ~30% on average[25]. Multiple mass shifts were detected in this work such as N-terminal methionine excision, signal peptide cleavage, truncations, and various PTMs including acetylation, methylation, oxidation, and phosphorylation. Two proteins that were identified with high confidence based on the number of fragment ions were YqjC and chaperone protein DnaK. The fragmentation and sequence patterns are shown in **figure 3.3d**; the protein YqjC had a signal peptide cleavage detected, while there was detection of a N-terminal methionine excision, acetylation, and C-terminal truncation for the chaperone protein DnaK.

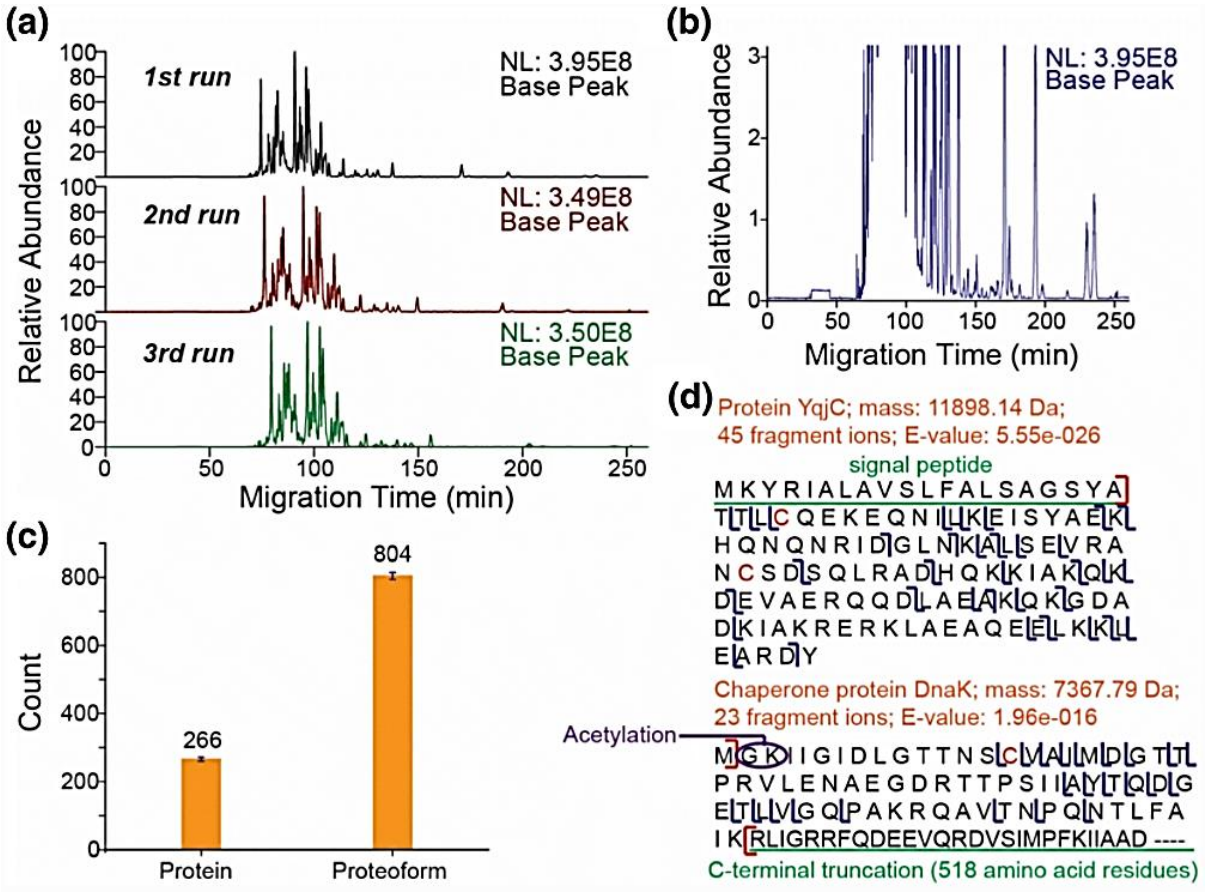


Figure 3.3. CZE-MS analyses of the *E. coli* proteome using a 1.5-m-long LPA-coated separation capillary. One microgram of proteins was injected per triplicate CZE-MS/MS run. (**a**) Base peak electropherograms of the triplicate CZE-MS/MS runs. (**b**) A zoomed-in electropherogram of one CZE-MS/MS run showing the separation window. (**c**) Protein and proteoform identifications of the triplicate CZE-MS/MS runs. 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. Carbamidomethylation modification are marked in red on the cysteine (C) residues. DnaK has one acetylation modification on either the G or K residue.

Next, we investigated the performance of the CZE-MS/MS platform for top-down proteomics using a 1.5-m LPA-coated separation capillary for mass limited *E. coli* proteome samples. Using 1-ug of *E. coli* proteins were dissolved in 2-uL of 50 mM ABC (pH ~8.0) was put into a CZE sample vial for duplicate CZE-MS/MS analyses. 500-nL of the sample was injected onto the 1.5-m separation capillary, which corresponds to 25% of the total sample volume within in the CZE sample vial. This injection volume corresponds to 250-ng of proteins onto the separation

capillary per the duplicate runs of the CZE-MS/MS platform. **Figure 3.4a** shows the base peak electropherogram of one CZE-MS/MS run with 250-ng injected. Comparing this data to the data obtained in **figure 3.3a**, both figures show comparably separation profiles and base peak intensities, however the sample loading amount in **figure 3.3a** was 4-fold higher than in **figure 3.4a.** Roughly 800 proteoforms and over 250 proteins were identified using the single-shot CZE-MS/MS platform with the 1.5-m long separation capillary when only 250-ng of *E. coli* proteins were injected (**figure 3.4b**). Substantially, the proteoform and protein identification numbers that were obtained with the CZE-MS/MS platform using 250-ng of proteins were comparable with the CZE-MS /MS platform using 1-ug of proteins. Based on this data, we further tested the capabilities of the CZE-MS/MS platform using the 1.5-m LPA-coated separation capillary for mass limited proteome samples. Next, we used 100-ng of the *E. coli* proteome as the starting material using the CZE-MS/MS platform using the 1.5-m LPA-coated separation capillary. Once again, 100-ng of the *E. coli* proteins were dissolved in 2-uL of 50 mM ABC (pH ~8.0) and placed within a CZE sample vial. 500-nL of the sample was injected on the 1.5-m separation capillary for duplicate CZE-MS/MS analysis; only 25-ng of the *E. coli* proteome was injected onto the separation capillary per CZE-MS/MS run. **Figure 3.4a** (bottom panel) shows the base peak electropherogram of one single-shot CZE-MS/MS run using 25-ng as the starting material. Comparing the 25-ng and 250-ng sample runs, the base peak intensity of the 25-ng sample was ~4x lower than that of the 250-ng sample run (1.39E8 vs 5.91E8). The CZE-MS/MS platform using 100-ng as the sample starting material identified 449 proteoforms and 173 proteins (**figure 3.4b**). The data that is presented in this study demonstrates the power of CZE-MS/S for top-down proteomics of mass -limited proteome samples.

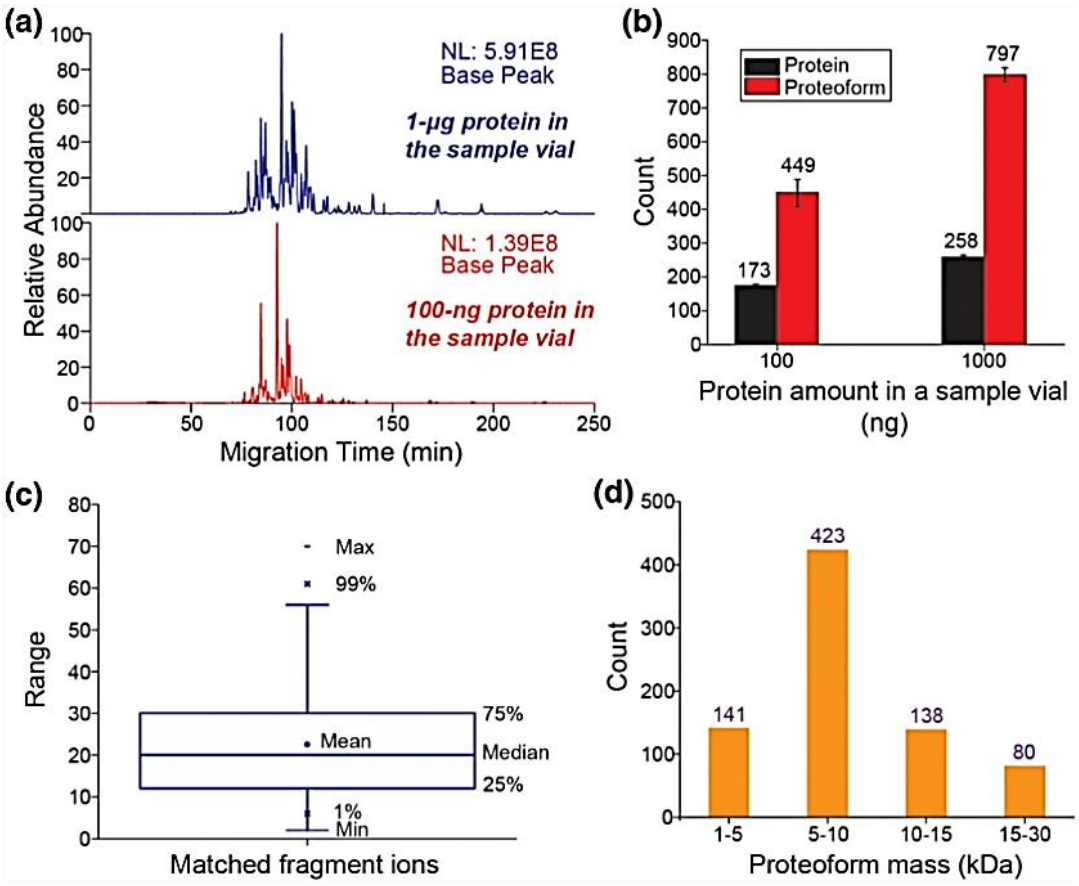


Figure 3.4. CZE-MS/MS analyses using the 1.5-m LPA-coated separation capillary of mass-limited *E. coli* proteome samples. (**a**) Base peak electropherograms of the *E. coli* proteome sample with 1-µg (top panel) and 100-ng proteins (bottom panel) as the starting materials. (**b**) Proteoform and protein identifications from the analyses of 1-µg and 100-ng *E. coli* samples. The error bars are the standard deviations 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 using the 1-µg *E. coli* sample. (**d**) Mass distribution of the identified proteoforms from one CZE-MS/MS analysis using the 1-µg *E. coli* sample.

Using one run of the CZE-MS/MS with the 250-ng of *E. coli* proteins injected, the number of matched fragment ions of identified and the mass distribution of the identified proteoforms were found; **figure 3.4c** shows the distribution of the match fragment ions. Approximately, 12 or fewer matched fragments were used for identification of ~25% of the proteoforms with the mean and median of the match fragment ions being 23 and 20. **Figure 3.4d** shows the proteoform mass distribution; roughly 90% of the proteoforms that were identified had masses below 15 kDa, while only 80 proteoforms were identified with masses higher than 15 kDa. Top-down

proteomics still has issues with the detection and identification of large proteoforms from complex mixture to due several reasons such as: the S/N ratio of proteoforms decreases as the proteoform molecular weight increases[37], the comigration/coelution of small and large proteoforms during liquid-phase separations causing issues with the MS detection of the larger proteoforms, and mass analyzers still have limited resolution for larger proteoforms due to the difficulty of determining the accurate mass for identification.

3.3.3 Quantitative top-down proteomics of zebrafish brain Cb and Teo regions using CZE-MS/MS with a 1.5-m long separation capillary

We dissected and collected the cerebellum (Cb) and optic tectum (Teo) from three separate mature zebrafish brains (**figure 3.5a**). The three separate sections of Cb were combined and the three separate sections of Teo were combined for protein extraction. The Cb and Teo extraction proteins were dissolved in 50 mM ABC (pH ~ 8.0) to reach a final concentration of 1 mg/mL for the CZE-MS/MS analysis using the 1.5-m LPA-coated separation capillary. Each CZE-MS/MS run was done in triplicate with a 500-nL sample injection volume (500-ng injected onto the separation capillary).

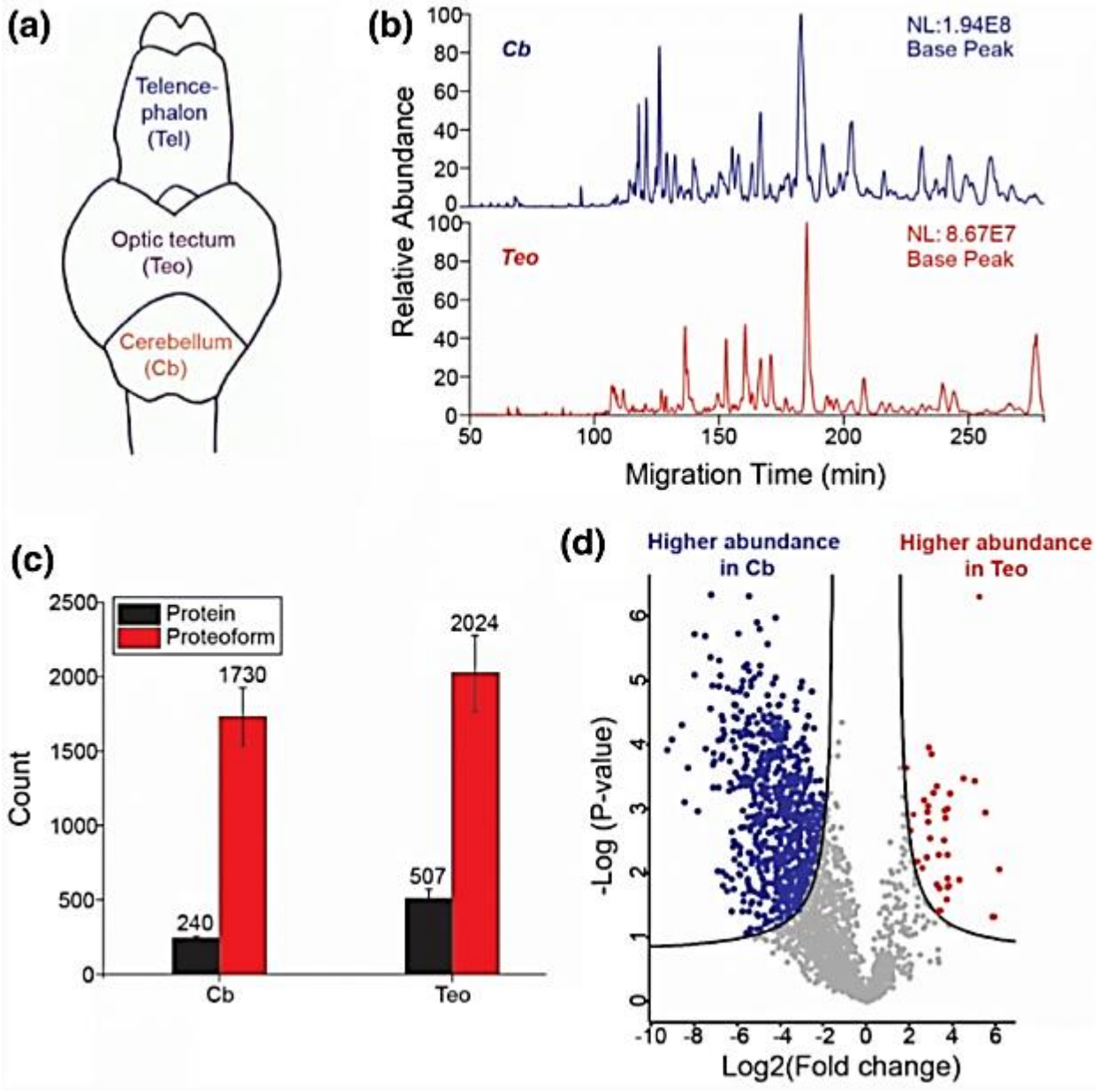


Figure 3.5. CZE-MS/MS analyses using the 1.5-m LPA-based separation capillary of zebrafish brain Cb and Teo regions. (**a**) An illustration of a mature zebrafish brain. (**b**) Base peak electropherograms of zebrafish brain Cb (top panel) and Teo (bottom panel) after CZE-MS/MS analyses. (**c**) Protein and proteoform identifications from the Cb and Teo samples. The error bars represent the standard deviations from triplicate CZE-MS/MS runs. (**d**) Volcano plot of the quantified proteoforms. The proteoforms with higher abundance in Cb are marked in blue and the proteoforms with higher abundance in Teo are marked in red.

Figure 3.5b shows the base peak electropherogram for Cb (top panel) and Teo (bottom panel).

Separation window for the Cb and Teo samples reached ~180-mins for the CZE-MS/MS

platform. The separation profiles for the Cb and Teo samples are strictly different from each and produced drastically different base peak intensities (1.9E8 vs 8.7E7) even though the same amount of total proteins for the Cb and Teo samples were injected onto the separation capillary. The CZE-MS/MS platform produced thousands of proteoform identifications both brain samples; the Cb sample identified 1,730 proteoforms corresponding to 240 proteins and the Teo sample identified 2,024 proteoforms corresponding to 507 proteins (**figure 3.5c)**. Based on the data presented, this illustrates that there are significant differences between the Cb and Teo proteome samples. Once again, the significant majority of the proteoforms identified show proteoform masses <5 kDa. This can be due to a combination of the top-down proteomic issues that was discussed in the earlier section, but mostly due to the face that many brain proteins are small so they can more readily pass the blood-brain barrier. This also demonstrates the potential ability of the CZE-MS/MS platform using 1.5-m LPA-coated separation capacity for the identification of thousands of proteoforms from complex proteome samples using only nanograms of starting material.

Next, we applied a label-free approach based on proteoform feature intensity to quantitatively compare the Cb and Teo samples. About 4,000 proteoforms were identified when all the proteoform identifications from the six CZE-MS/MS runs were combined for proteoform quantification. However, when considering the proteoforms for quantification only the feature intensities from proteoforms that were identified in all six of the CZE-MS/MS were examined for comparison across the two samples. When taking this information into account, we quantified ~2,000 proteoforms across all the samples. For each quantified proteoform, the feature intensity was normalized to the quantified proteoform intensity from the first CZE-MS/MS run of the Cb sample. Then a log2 transformation was applied to the quantified proteoform. We used the

Perseus software to perform a t-test verifying the proteoform intensity difference between the Cb and Teo samples with 1% FDR and S0 = 1. S0 illustrates any artificial variance within groups (Cb and Teo samples) and controls how important three statistical tests are (t-test, p-value, and the difference between means). For instance, when s0 = 0, then only the p-value matters, however a nonzero s0 allows the difference of means to play a role within the statistical values[38].

The volcano plot of the quantified proteoforms is shown in **figure 3.5d**. Of the 2,000 proteoforms that were quantified, 786 proteoforms illustrated drastic abundance differences between the Cb and Teo samples. For the Cb samples, 749 proteoforms corresponding to 131 proteins demonstrated higher abundance; for the Teo samples, 37 proteoforms corresponding to 26 proteins demonstrated higher abundance. Using the David Bioinformatics Resource 6.8 for analysis, we performed biological enrichment analysis on the 131 upregulated proteoforms in the Cb brain region. These upregulated proteins within the Cb were highly enriched in these biological processes: including muscle contraction (p value: 1E-4), glycolytic process (p value: 5E-16), and mesenchyme migration (p value: 0.01). However, more experiments using many biological replicates need to be completed and validated before any solid conclusions about the biological mechanism presented within in this study. This experiment is the first attempt at quantitative top-down proteomics of complex proteomes using CZE-MS/MS.

Next, we further evaluated the proteoform abundance from the single-shot CZE-MS/MS using the 1.5-m LPA-coated separation capillary of the Cb and Teo brain samples by estimating the dynamic range using the proteoform feature intensity (**figure 3.6**). There were roughly six orders of magnitude dynamic range produced from the proteoform abundance of the Cb and Teo samples using the single-shot CZE-MS/MS platform (**figure 3.6a**). The accuracy of the determine dynamic range can be influenced by the proteoforms mass difference, there we

manually checked two proteoforms of parvalbumin with similar masses (11,558 Da vs 11,542 Da). Both proteoforms were identified with high confidence (e-values of 6.59E-03 vs 1.55E-27 with FDR's of 0% and 0.2%). The two proteoforms had nearly a 5 fold difference in magnitude from each other; Proteoform 1 (11,558 Da) had a proteoform feature intensity of 2.7E5 (**figure 3.6b**), while Proteoform 2 (11,542 Da) had a proteoform feature intensity of 1.9E10 (**figure 2.6c**). Two conclusions can be drawn from this data: (1) Single-shot CZE-MS/MS using the 1.5-m LPA-coated separation capillary cab reach 5 orders of magnitude difference in proteoform abundance, and (2) significantly different abundance can be produced from different proteoforms that arise from the same gene.

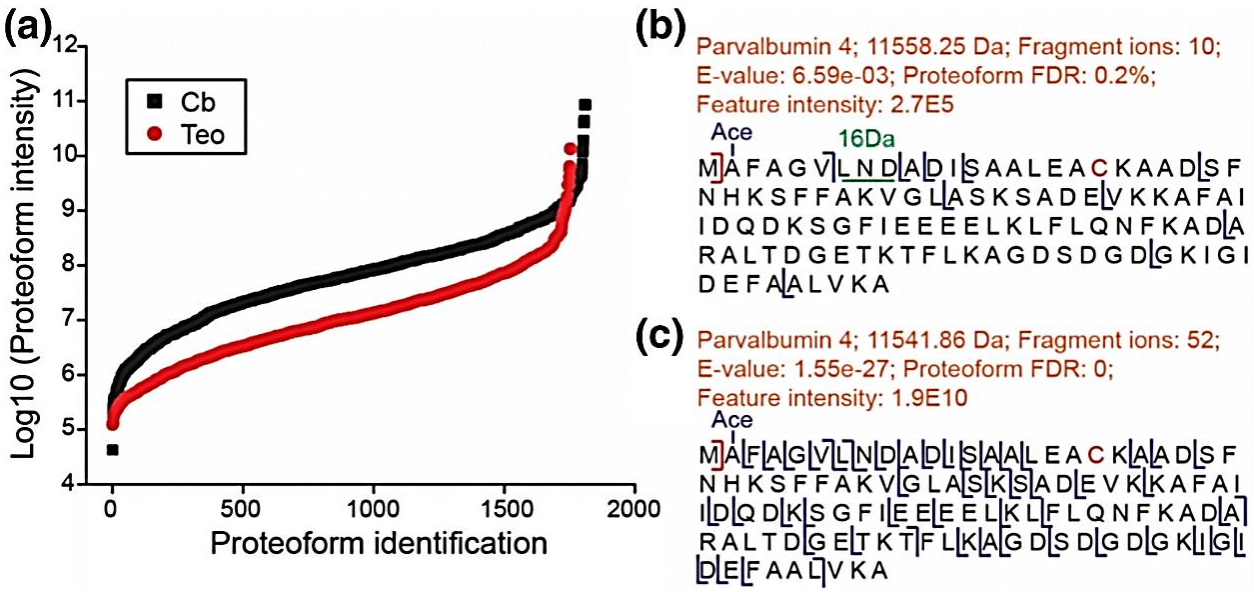


Figure 3.6. (**a**) Proteoform abundance dynamic range from single-shot CZE-MS/MS analysis using the 1.5-m LPA-coated separation capillary of Cb and Teo samples. The proteoform feature intensity was used to approximation the dynamic range. (**b, c**) The sequence and fragmentation pattern of two proteoforms of parvalbumin 4.

3.4 Conclusions

In this study, we introduce a CZE-MS/MS platform using a 1.5-m LPA-coated separation capillary for large-scale top-down proteomics of an *E. coli* proteome and two regions of a

zebrafish brain proteome. The single-shot CZE CZE-MS/MS platform using a 1.5-m LPA-coated separation capillary produced a 2-uL leading capacity for the analysis of a standard protein mixture and a 180-min separation window for complex proteome samples. The *E. coli* proteome identified ~800 proteoforms corresponding to 258 proteins with only 250-ng inject and ~450 proteoforms corresponding to 178 proteins with only 25-ng injecting using the single-shot CZE-MS/MS platform. Single-shot CZE-MS/MS using the 1.5-m LPA-coated separation capillary identified and quantified thousands of proteoforms from two different regions (Cb and Teo) of the Zebrafish brain proteome using only 500-ng of protein material. This is the first time that quantitative top-down proteomics using CZE-MS/MS was attempted and this is also the largest proteoform identification data set using top-down based CZE-MS/MS reported yet. The data that is produced within this study demonstrates the ability of our CZE-MS/MS platform for large-scale top-down proteomics of mass limited samples. The CZE-MS/MS platform can also be used for the top-down characterization of proteoforms from other mass-limited samples such as single cells, circulating tumor cells, and laser capture microdissection.

3.4 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).

APPENDIX

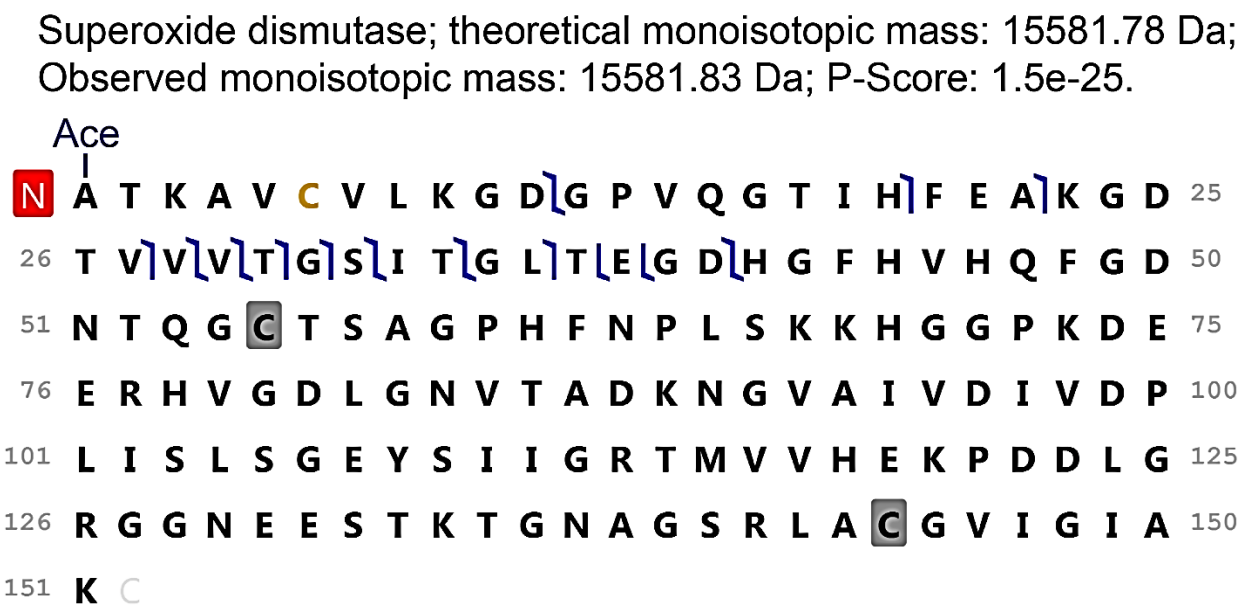


Figure 3.7. Fragmentation and sequence pattern of superoxide dismutase using (UniProt ID: P00442). A N-terminal acetylation was identified and is labeled on the alanine (A) residue. Highlighted in grey is the carbamidomethylation on two cysteine residues.

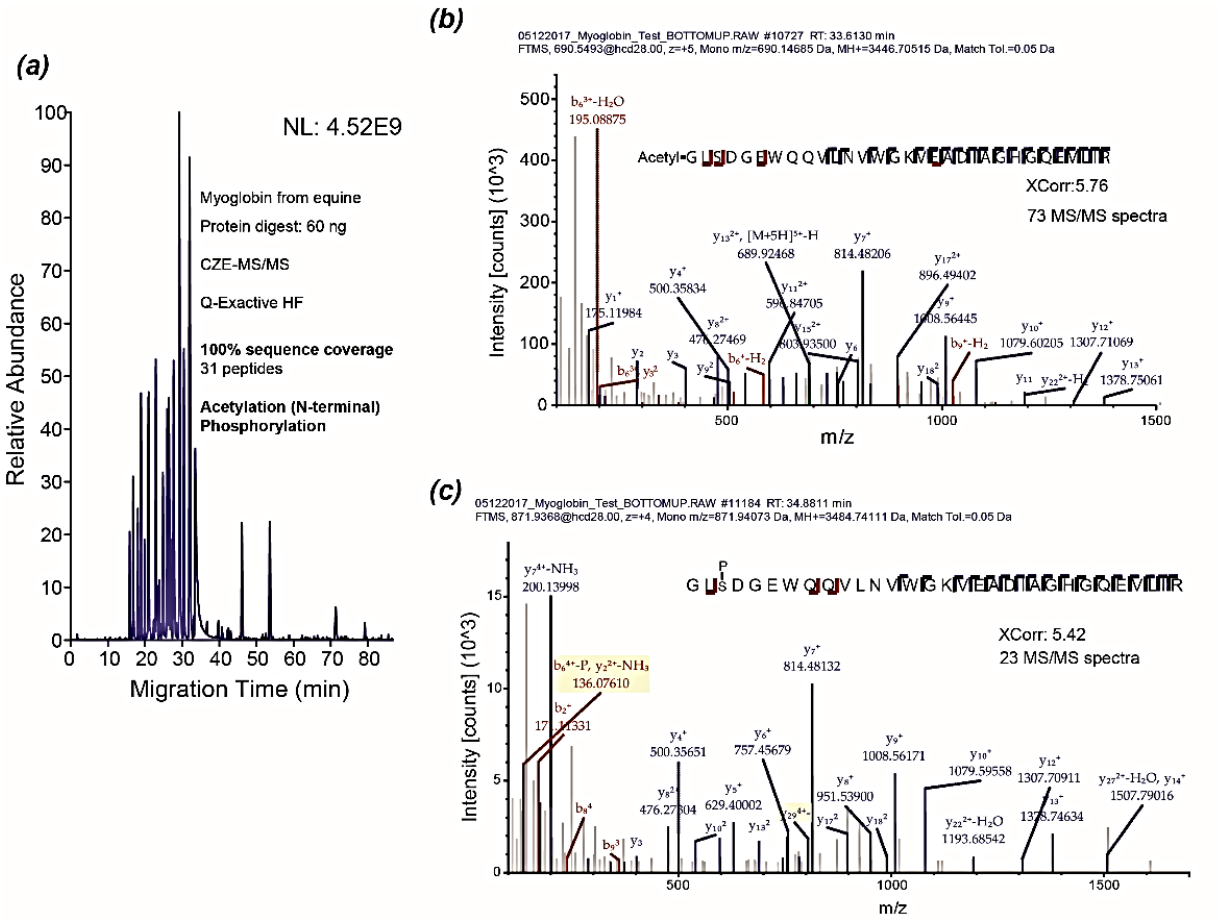


Figure 3.8. (**a**) Base peak electropherogram of the tryptic digest of myoglobin analyzed by bottom-up based CZE-MS/MS. (**b**) Myoglobin peptide with a N-terminal acetylation with the annotated MS/MS spectrum using bottom-up proteomics. (**c**) A phosphopeptide myoglobin with the annotated MS/MS spectrum using bottom up proteomics. Confirmation of the proteoforms that were detected using top-down based CZE-MS with a 1.5-m LPA-coated separation capillary.

REFERENCES

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. 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), 254–258.

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.* **2016**, *15* (3), 976–982.

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.* **2017**, *89* (10), 5467–5475.

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.* **2013**, *110* (25), 10153–10158.

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* **2017**, *1498*, 99–110.

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.* **2017**, *16* (2), 609–618.

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.* **2017**, *16* (2), 1087–1096.

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.* **2018**, *17* (10), 3526–3536.

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.* **2018**, *90* (14), 8553–8560.

10. Jorgenson, J. W.; Lukacs, K. D. Capillary Zone Electrophoresis. *Science* **1983**, *222* (4621), 266–272.

11. Valaskovic, G. A.; Kelleher, N. L.; McLafferty, F. W. Attomole Protein Characterization by Capillary Electrophoresis-Mass Spectrometry. *Science* **1996**, *273* (5279), 1199–1202.

12. Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavallée-Adam, M.; Yates, J. R., 3rd. Sheathless Capillary Electrophoresis-Tandem Mass Spectrometry for Top-down Characterization of Pyrococcus Furiosus Proteins on a Proteome Scale. *Anal. Chem.* **2014**, *86* (22), 11006–11012.

13. Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates, J. R., 3rd. In-Line Separation by Capillary Electrophoresis Prior to Analysis by Top-down Mass Spectrometry Enables Sensitive Characterization of Protein Complexes. *J. Proteome Res.* **2014**, *13* (12), 6078–6086.

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.* **2016**, *15* (10), 3679–3685.

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* **2014**, *14* (10), 1158–1164.

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.* **2013**, *85* (12), 5989–5995.

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.* **2013**, *85* (4), 2289–2296.

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-B1. *Anal. Chem.* **2016**, *88* (2), 1138–1146.

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* **2013**, *12* (9), 2640–2656.

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.* **2014**, *86* (10), 4873–4878.

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: Capillary Electrophoresis Electrospray Interface. *Rapid Commun. Mass Spectrom.* **2010**, *24* (17), 2554–2560.

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.* **2015**, *14* (5), 2312–2321.

23. Moini, M. Simplifying CE-MS Operation. 2. Interfacing Low-Flow Separation Techniques to Mass Spectrometry Using a Porous Tip. *Anal. Chem.* **2007**, *79* (11), 4241–4246.

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* **2017**, *142* (12), 2118–2127.

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.* **2017**, *89* (22), 12059–12067.

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.* **2018**, No. 140. https://doi.org/10.3791/58644.

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.* **2018**, *90* (9), 5529–5533.

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* **2016**, *146*, 839–843.

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. Weinheim Bergstr. Ger.* **2013**, *125* (51), 13906–13909.

30. 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.

31. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, *24* (21), 2534–2536.

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.* **2002**, *74* (20), 5383–5392.

33. Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214.

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* **2016**, *13* (9), 731–740.

35. Aebersold, R.; Morrison, H. D. Analysis of Dilute Peptide Samples by Capillary Zone Electrophoresis. *J. Chromatogr.* **1990**, *516* (1), 79–88.

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.* **2000**, *72* (6), 1242–1252.

37. Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L. On the Scalability and Requirements of Whole Protein Mass Spectrometry. *Anal. Chem.* **2011**, *83* (17), 6868–6874.

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.* **2001**, *98* (9), 5116–5121.

39. Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* **2009**, *4* (1), 44–57.

40. 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. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (8), 1435–144

CHAPTER 4. Development of a Highly Sensitive Top-Down Proteomic Workflow Using Capillary-Zone Electrophoresis-Electrospray Ionization Tandem Mass Spectrometry for Spatially Resolved Proteomics of Zebrafish Brain

4.1 Introduction

Cellular systems, such as the brain, are heterogeneous creating specific microenvironments with distinct molecular and functional characteristics to perform biological functions[1-5]. This feature results in varied cellular responses and therefore unique pathology. To better understand the molecular and cellular architecture within brains and its biological function, there is a need for high resolution spatially resolved molecular imaging of tissue section[1-5]. Highly resolved spatial proteomics is a promising method for understanding health and disease within the human brain. The brain proteome is highly complex and compartmentalized organ; many genes within the brain are expressed in a temporospatial fashion, meaning that most likely different proteoforms are expressed in different brain regions[9]. Presently, we have limited knowledge on how protein function affects neurological disorders; for example, we know that neurological disorders do exhibit a phenomenon called "selective vulnerability" and while the cause of this phenomenon is unknown, one of the defining characteristics is protein homeostasis dysfunction[6-9]. Therefore, understanding the brain proteome may unlock key characteristics for disease pathology that may result in therapeutic answers for these diseases.

On-tissue spatially resolved proteomics can provide valuable information about the tissue microenvironment and how biological functions can fluctuate from small cellular changes. Most proteomic studies utilize bottom-up proteomics (BUP) due its high sensitivity, throughput, and

robustness[10-12]. However, this technique introduces a protein inference problem, often inferring proteins based on a small number of peptides that cover limited regions of the full-length protein. BUP has issues with distinguishing protein isoforms and proteoforms that have homologous sequences and can lose information on the combinations of PTMs on the protein sequence after digestion[10-12]. Top-down proteomics (TDP) overcomes the protein inference issue by analyzing intact proteins and identifying proteoforms and their PTM's with higher confidence than BUP[13-17]. Presently, TDP provides opportunities to gain valuable insight into biological mechanisms by analyzing proteoform abundances.

Currently, targeted methods such as immunohistochemistry (IHC) staining, fluorescence in situ hybridization (FISH) or imaging mass cytometry, are used for characterization of the molecular landscape, however these methods rely on previous knowledge of the working system and availability of suitable antibodies. Matrix-assisted laser desorption/ionization (MALDI) can be used for discovery experiments and has been shown to measure hundreds of different species for molecular imaging, though there are challenges due to lack of a coupled separation dimension and ionization suppression issues[2,20,23].

Reversed phase liquid chromatography tandem mass spectrometry (LC-MS/MS) based top-down proteomics has been shown to identify more than >1,000 proteoforms, however this technique does require higher concentration levels than what is typically sampled when attempting to map protein expression levels with high spatial resolutions. For instance, Shen et al. used RPLC-MS/MS for the top-down proteomics of a simple bacterial lysate; this approach produced over 900 proteoforms using a few μg of proteome sample[18]. There are issues with sample loss in the RPLC-MS system (e.g., on the stationary phase and in the injection valve). Alternative platforms with better sensitivity for proteoform detection could enable TDP of nanograms of complex

proteome samples. CZE-ESI-MS/MS has been shown to identify and quantify hundreds to thousands of proteoforms with nanograms of starting material. Our group recently demonstrated that CZE-ESI-MS/MS can identify ~800 proteoforms from 250-ng of *E. coli* cell proteins and can quantify thousands of proteoforms via consuming 500-ng of zebrafish brain proteome material[24], making this a viable candidate for mass-limited proteome samples.

There have been advances for optimizing the workflow for mass-limited proteome samples using top-down proteomics. However, intact proteoforms exhibit poor recovery due to their wide range of chemical properties preventing uniform solubilization. One key point is the use of MS-compatible detergents, such as degradable or non-ionic detergents, to reduce sample loss because there is no need for the extra steps to remove the detergent before MS analysis. Zhou et al. coupled the nanoPOTS (nanodroplet processing in one pot for trace samples) method [21,22], which employed degradable Rapigest detergent and only 200-nL sample processing volume, with RPLC-MS/MS for TDP of ~100 human cancer cells where over 150 proteoforms were identified[19]. Delcourt et al. used ProteaseMax, an acid-cleavable detergent, for the identification ~70 proteoforms from laser capture microdissection (LCM) of rat brain tissue (1 mm^2 with a thickness of 30-μm)[2].

LCM is an interesting technique that can sample tissue with high spatial resolution. LCM consists of an inverted light microscope coupled with a laser beam that can be used to isolate cell from the surrounding tissue[25,26]. Cells of interest, predefined by the user, is isolated and then cut away from the adjacent tissue[25,26]. After which, a small laser pulse under the cut tissue cells will lift the tissue on to an adhesive cap for extraction. LCM is an interesting technique that it can extract specific tissue regions, even single cells, can be extracted with high accuracy[25,26].

Here, we describe a modified top-down proteomic workflow that can applied to mass-limited proteome samples by using a non-ionic detergent Octyl glucopyranoside (OG) for cell lysis and CZE-ESI-MS/MS. The extracted proteins were directly analyzed by the CZE-MS/MS without any sample cleanup, resulting in a higher throughput and reduced sample loss. This workflow was applied to LCM sample of zebrafish brain. Zebrafish are a useful model organism due to three reasons: (1) high genetic similarity to humans, (2) high fecundity rate and fast development, and (3) cheap to maintain, easy to fertilize and develops outside of the mother for easy examination[27]. Any information we can learn about proteoform differences within the Zebrafish brain, we can also possibly apply to human brains to learn more about how protein function affects neurological disorders. We used a 20-μm-thick section of brain with a 500- μm^3 area to demonstrate the modified workflow along with our CZE-ESI-MS/MS platform for the identification and quantification of proteoforms of mass-limited samples.

4.2 Experimental

4.2.1 Materials and Reagents

Acrylamide was purchased from Acros Organics (NJ, USA). Standard proteins, ammonium bicarbonate (NH4HCO3), 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). Aqueous mixtures were filtered with Nalgene Rapid-Flow Filter units (Thermo Scientific) with 0.2 μm CN membrane and 50 mm diameter. Fused silica capillaries (50 μm i.d./360 μm o.d.) were obtained from Polymicro Technologies (Phoenix, AZ).

4.2.2 Tissue Preparation

Zebrafish brains were collected from two mature female zebrafish (AB/Tuebingen line). The zebrafish were provided by Professor Jose Cibelli's group at the department of Animal Science at Michigan State University. All the animal related experiments were performed using a protocol approved by the Institutional Animal Care and Use Committee of Michigan State University. After extraction, the brains were rinsed with phosphate buffered saline, placed into a cyrostat block and covered with Optimal Cutting Temperature Compound (OCT compound) and frozen using dry ice. The brains in the OCT compound were stored at -80 ^{o}C until use. A cryostat was used to cut the zebrafish brain tissue to a thickness of 20-μm and placed onto PEN membrane slides. The brain tissues were then stained with Cresyl Violet Staining.

4.2.3 Laser Capture Microdissection

Zeiss Palm MicroBeam IV laser capture microdissection system was employed to cut areas of the brain tissue. The PEN membrane slides were placed onto the microscope slide adaptor and a computer mouse, using a closed-shape manual drawing tool, was cut to cut square slices (500 μm^2) at specific regions of interest on the brain tissue. To collect the tissue, we used an adhesive cap to stock the tissue sections on the adhesive caps. Ten different samples were collected from two Zebrafish brain sections. All tissue samples were stored at -80 ^{o}C until ready for CZE-MS analysis.

4.2.4 Sample Preparation for CZE-MS/MS

Octyl glucopyranoside, OG, detergent was dissolved in 100 mM ABC (4.5% (w/v)) and the tissue samples from the laser microdissection were re-suspended in 5-μL by placing the detergent directly onto the adhesive cap. The detergent was then pipetted up and down several times and

the sample were spun down in a centrifuge. The samples were then prepared by three different approaches (**figure 4.1**): the typical top-down workflow and a modified workflow using two different types of extraction methods. For the typical workflow 10 samples from the same Zebrafish brain slice were used; this workflow consisted of in-cap solubilization using 4.5% (w/v) OG followed by reduction and alkylation, then acetone precipitation. The modified workflow consisted of in-cap solubilization using 4.5% (w/v) OG, but removed the reduction, alkylation and acetone precipitation. The modified workflow consisted of two different extraction methods: ultrasonication and a repeated freeze/thaw method. For the modified workflow, ten samples from the same Zebrafish brain slice were used in the modified workflow (**figure 4.2**). Five samples originating from the same Zebrafish brain slice were placed on ice and ultrasonicated for 20 minutes. The other 5 samples originating from the Zebrafish same brain slice went through a repeated freeze/thaw protocol; the samples were placed into liquid nitrogen and thawed at 37 ^{o}C for 2 minutes for a total of 6 rounds of the freeze/thaw. All 10 samples were then vortexed and spun down. The sample was diluted 2x with water to reach a total ABC concentration of 50 mM with a total volume of 10-μL.

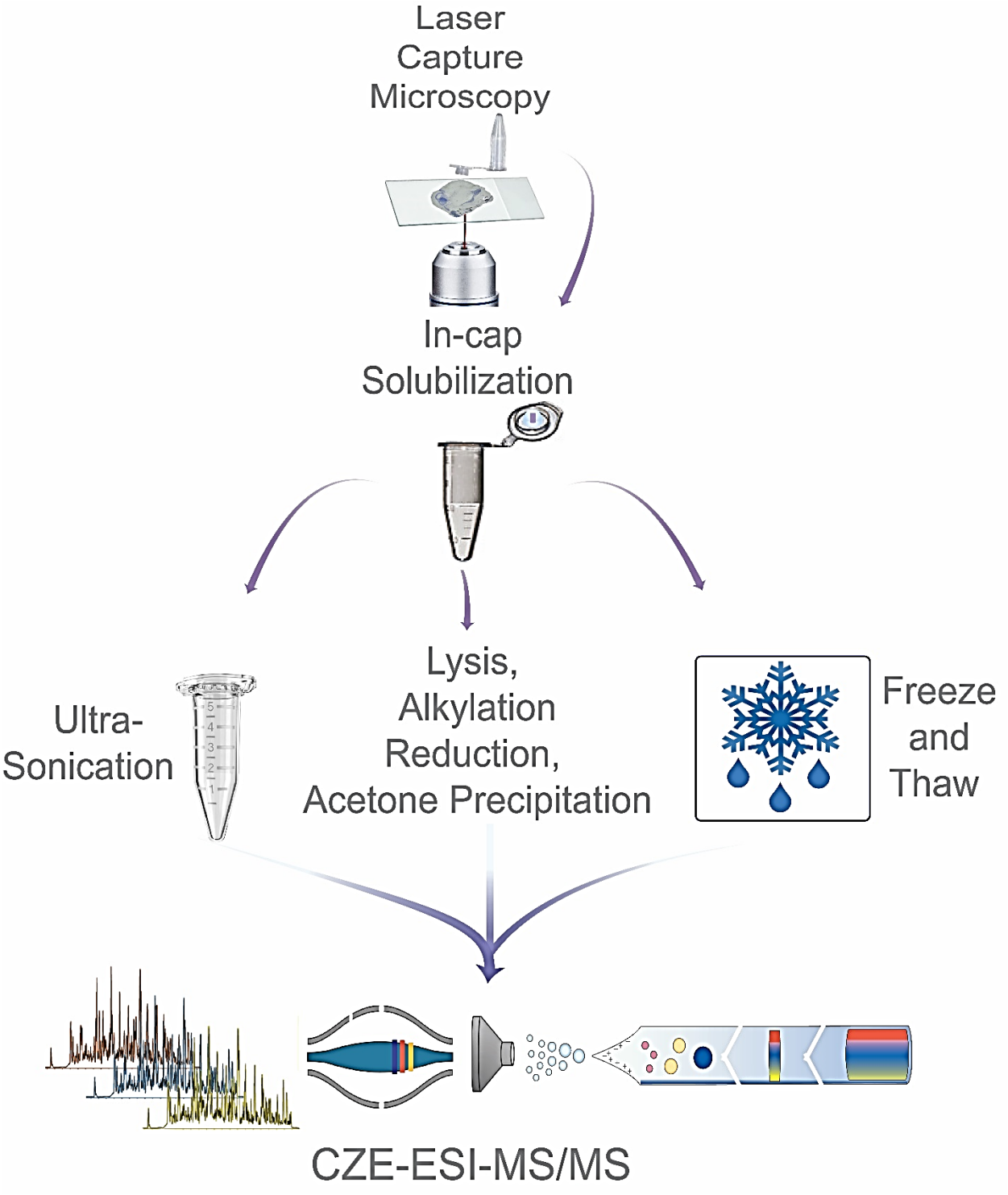


Figure 4.1. Schematic overview of the LCM workflow. Two workflows were used: the standard top-down proteomic workflow and a modified top-down proteomic workflow. The modified workflow used two different extraction methods: ultrasonication and freeze/thaw.

4.2.5 CZE-ESI-MS/MS

An CESI 8000 CE autosampler (Bruker) was used for automated CE operation. The CE autosampler was coupled to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) through a commercialized electro-kinetically pumped sheath flow CE-MS interface (CMP Scientific, Brooklyn, NY). A fused silica capillary (50-μm i.d., 360-μm o.d., 75 cm in length) with a linear polyacrylamide (LPA) coating on the inner wall was used for the CZE separation. The outlet end of the capillary was etched with hydrofluoric acid to reduce the outer diameter. *(Caution: use appropriate safety procedures while handling hydrofluoric acid solutions.)* The CZE background electrolyte (BGE) was 10% acetic acid (pH ~2.2) and the sheath buffer 0.2% (v/v) formic acid containing 10% (v/v) methanol. ESI emitters were pulled to an opening of 30-40-μm using 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. Each sample vial was coated with 1 mg/mL Bovine Serum Albumin (BSA) to reduce sample adsorption to the inner walls during analysis. Sample injection was carried out by applying 5-psi for 70-s corresponding to ~370-nL of sample material injected onto the capillary. 30 kV was used for separation and was applied at the injection end of the capillary, while 2.0-2.2 kV was applied to the sheath buffer for ESI. After each CZE run, the capillary was flushed with the BGE by applying a 20-psi pressure for 10 mins.

A Q-Exactive HF mass spectrometer was operated in data dependent mode; the top 5 most intense ions in one MS spectrum were sequentially isolated in the quadrupole and fragments using higher energy collision dissociation (HCD). Only charge states higher than 3 were selected for HCD fragmentation. MS was set to a mass resolution of 120,000 (at *m/z* 200), 3 microscans, 1E6 AGC target value, 50-ms maximum injection time and 600-2000 *m/z* scan range. For the MS/MS, the mass resolution was 60,000, 3 microscans, AGC value of 1E5, 200-ms maximum

injection time, 4 *m/z* isolation window, and a normalized collision energy (NCE) of 20%. The dynamic exclusion was set to 30s and the exclude isotopes function was on.

4.2.6 Data Analysis

All samples were analyzed using Xcalibur software (Thermo Fisher Scientific) to get intensity and migration time of proteins. The electropherograms were exported from Xcalibur and were formatted using Adobe Illustrator to report the final figures.

The LCM sample RAW files were analyzed using TOP-down mass spectrometry based proteoform identification and characterization (TopPIC) platform for the proteoform identification and quantification. The RAW files were converted to mzML files using the MsConvert tool and spectral deconvolution was done using the TOP-down mass spectrometry feature detection (TopFD) to generate msalign files. The msalign files were used as the input for TopPIC (version 1.4.0) for database searching. UniProt databases for zebrafish (AUP000000437) were used for the database search. The maximum number of unexpected modifications were set to 2 and the maximum mass shift of unknown modification was set to 500 Da, while the precursor and fragment mass error tolerances were 15 ppm. The target decoy approach was used for the false discovery rate (FDR) and a 5% proteoform level FDR was used to filter proteoform identifications.

Label-free quantification for top-down proteomics was performed using the TopDiff tool to group spectral peaks into isotopomer envelopes and then combine these envelopes from the same proteoform with different retention times and charge states. The combined envelopes were reported a CZE-MS feature intensity. A feature intensity is calculated as a sum using peaks from

all scans and charge states as described in my previous work[24]. Only proteoforms that that had feature intensities in all CZE-MS/MS runs were considered for quantification.

4.3 Results and Discussion

4.3.1 Effectiveness of OG for Qualitative Top-Down Proteomics using Single-Shot CZE-ESI-MS/MS of Laser Capture Microdissection Tissue Samples

Intact proteoforms exhibit a lower sensitivity due their physiochemical properties causing issues such as 'charge state dilution' and 'isotope dilution' effects[14]. During the ESI process, a single proteoform will acquire multiple charges causing the protein's total signal to be divided over all the different charge states (i.e., charge state dilution). The larger the protein, the more signal distribution will occur across multiple channels. Additionally, proteoform's large molecular weight causes a broad isotopic distribution that lowers the S/N ratio (i.e., isotope dilution). Therefore, minimizing sample loss during sample preparation is crucial if we are to apply top-down proteomics to mass-limited proteome samples, such as LCM samples. One way that this can be accomplished is by streamlining the workflow by utilizing MS-compatible detergents, where there is no need for detergent removal before sample analysis, resulting in lower sample loss.

OG detergent is a nonionic surfactant that is MS-compatible; it is composed of an octanol with a glucose attached via a glycosidic bond. It is a neutral molecule and shouldn't affect the CZE separation as it will migrate out of the capillary during the flushing step, therefore there is no need to remove the detergent before CZE-ESI-MS/MS analysis. Here, we evaluate how different workflows and extraction methods affect proteoform identification (**figure 4.1**). We compared a normal top-down workflow to a modified top-down workflow, where the modified workflow

used either ultra-sonication or a freeze/thaw extraction method to extract the protein material from the LCM tissue samples. To evaluate the different work, we dissected and analyzed three different Zebrafish brain regions (cerebellum (Cb), optic tectum (Teo), and telencephalon (Tel)) from a 20-μm thick brain slice. The brain tissue samples were microdissected as square regions with an area of 500-μm^2 corresponding to roughly 5,000 cells (~500-ng total protein content) (**Figure 4.2B**).

For the typical workflow 10 samples from the same Zebrafish brain slice were used; this workflow consisted of in-cap solubilization using 4.5% (w/v) OG followed by reduction and alkylation, then acetone precipitation. The modified workflow consisted of in-cap solubilization using 4.5% (w/v) OG, but removed the reduction, alkylation and acetone precipitation. The modified workflow consisted of two different extraction methods: ultrasonication and a repeated freeze/thaw method to compare proteoform identifications. For the modified workflow, ten microdissected samples from the same Zebrafish brain slice were used in the modified workflow (**figure 4.2**). The ultra-sonification method utilized five microdissected samples originating from the same Zebrafish brain slice (**figure 4.2A**). The microdissected regions consisted of two tissue sections from the cerebellum (Cb2 and Cb3) and three tissue sections from the optic tectum (Teo3, Teo4, and Teo5). The five tissue sections were placed on ice and ultrasonicated for 20 minutes. The other five microdissected samples originating from the same Zebrafish brain slice as the modified workflow were used in the freeze/thaw protocol (**figure 4.2A**). The dissected regions consisted of one tissue section from the cerebellum (Cb1), two tissue sections from the optic tectum (Teo1 anTeo2), and two tissue sections from the telencephalon (Tel1 and Tel2).

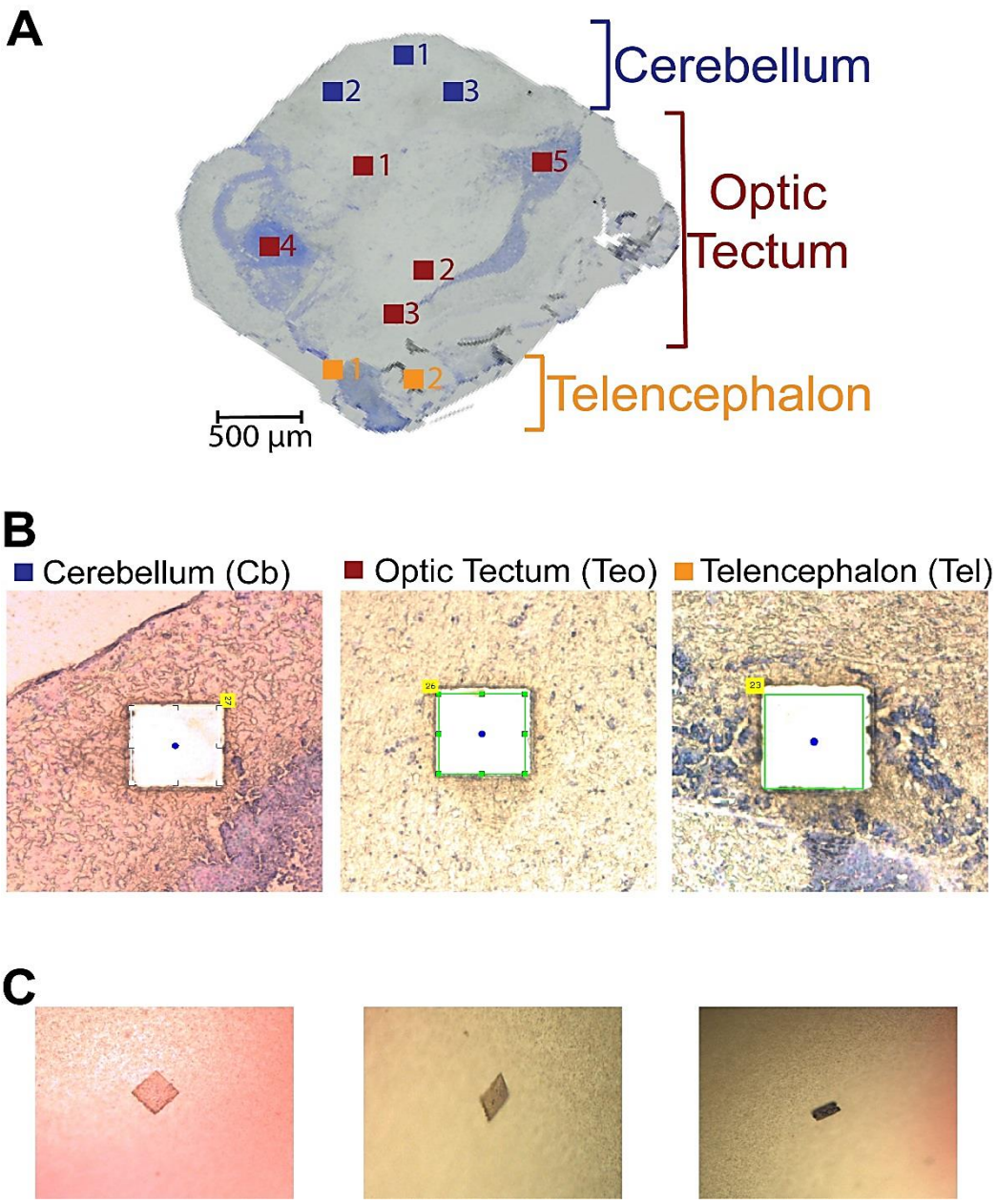


Figure 4.2. (**A**) A 20-µm-thick Zebrafish brain slice used in the study. Three separate regions of the Zebrafish brain, cerebellum (Cb), Optic Tectum (Teo), and Telencephalon (Tel), were microdissected with a spatial resolution of 500 μm^2. (**B**) The microscopic images of the three square regions of brain tissue regions after the microdissection. (**C**) Corresponding microdissected tissue section on the LCM cap.

First, the LCM samples underwent a normal TDP workflow (i.e., in-cap lysis with OG, reduction, alkylation, and acetone precipitation), however this method produced <10 proteoform identifications, meaning that there was significant sample loss on the walls of the tube and during the acetone precipitation step. Therefore, this workflow wasn't used.

Next, we evaluated a modified top-down workflow where OG in 100 mM ABC would be used as the lysis buffer, and after lysis the OG would be diluted 2x to reached a final concentration of 50 mM ABC with a total sample volume of 10-µL and 370-nL (~25% of the total capillary length) would be directly injected onto our CZE-ESI-MS/MS platform. We assessed the extraction efficiency of two method: ultra-sonication and freeze/thaw on the five different microdissected tissue samples isolated from different brain regions but originating from the same brain slice using LCM. The different extraction methods perform differently when using proteoform identifications as a comparison approach.

For the ultrasonication method, we detected between 12-40 proteoforms identifications corresponding to 5-15 proteins (**figure 4.3A**). This extraction method performed poorly in terms of identification numbers, indicating that either the extraction method didn't yield good protein amount or that there was significant sample loss. The ultrasonication extraction could have issues with the low sample volume, meaning that the high frequency vibrations applied to the sample couldn't disrupt the cell walls efficiently to fully extract the material. One more reason is during

this step the adsorption onto the walls of the tube due splashing caused by the vibrations could also cause significant sample loss.

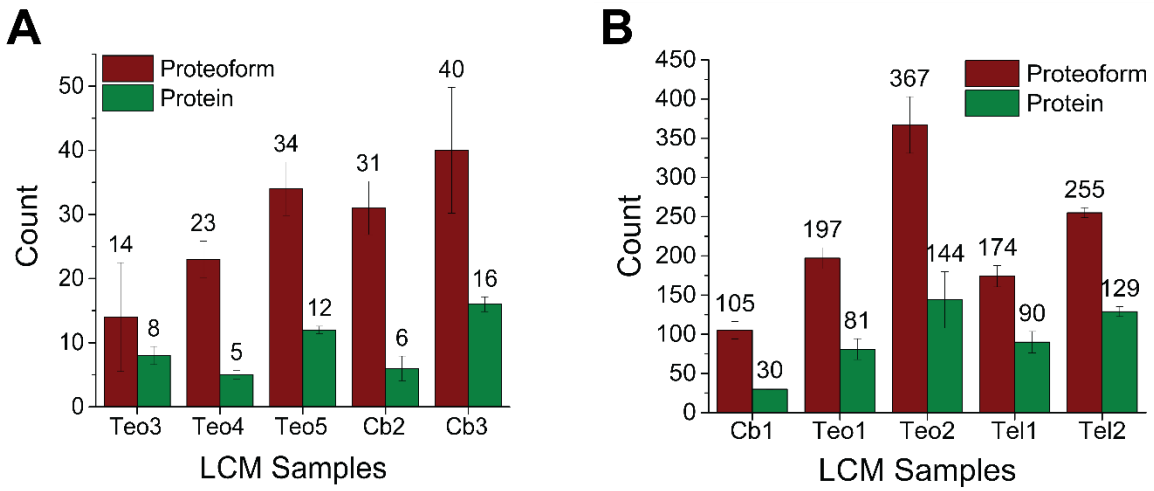


Figure 4.3. Comparison between proteoforms and proteins identified from the LCM cap using the ten tissue samples originating from the same Zebrafish brain slice utilizing OG detergent for protein extraction with either (**A**) ultra-sonication and (**B**) freeze/thaw extraction method. The error bars are the standard deviations from duplicate CZE-ESI-MS/MS runs.

The freeze/thaw method produced a higher proteoform identification rate. We detected between 105-367 proteoforms corresponding to 30-144 proteins. (**figure 4.3B**). The freeze/thaw method performed the best compared to the typical top-down workflow and the modified workflow using ultrasonication. The tissue sections contained roughly 5,000 cells equating to ~500-ng of protein content within the tissues. We lysed the tissue samples with 5-µL and then diluted the sample 2x, resulting in a total volume of 10-µL. During CZE analysis, we injected <5% of the sample volume onto the capillary (~250 cells). Our CZE-ESI-MS/MS platform produced highly sensitive separation of the zebrafish brain cells producing an average proteoform identification of ~220 proteoforms. Recently, Zhou et al. applied their NanoPOTS platform for the top-down proteomics of ~100 HeLa cell using nLC-MS/MS identifying 174 proteoforms. This study utilized pure HeLa cell cultures which have homogenous genotypic and phenotypic

characteristics and are easier to handle for protein analysis. Our study utilized primary tissues obtained in vitro for analysis; primary tissues require more external extraction mechanisms to both extract the cells from the tissue matrix and to extract the protein from the cells. Therefore, our results are consistent with the sensitivity that was acquired in the NanoPOTS study using nLC-MS/MS.

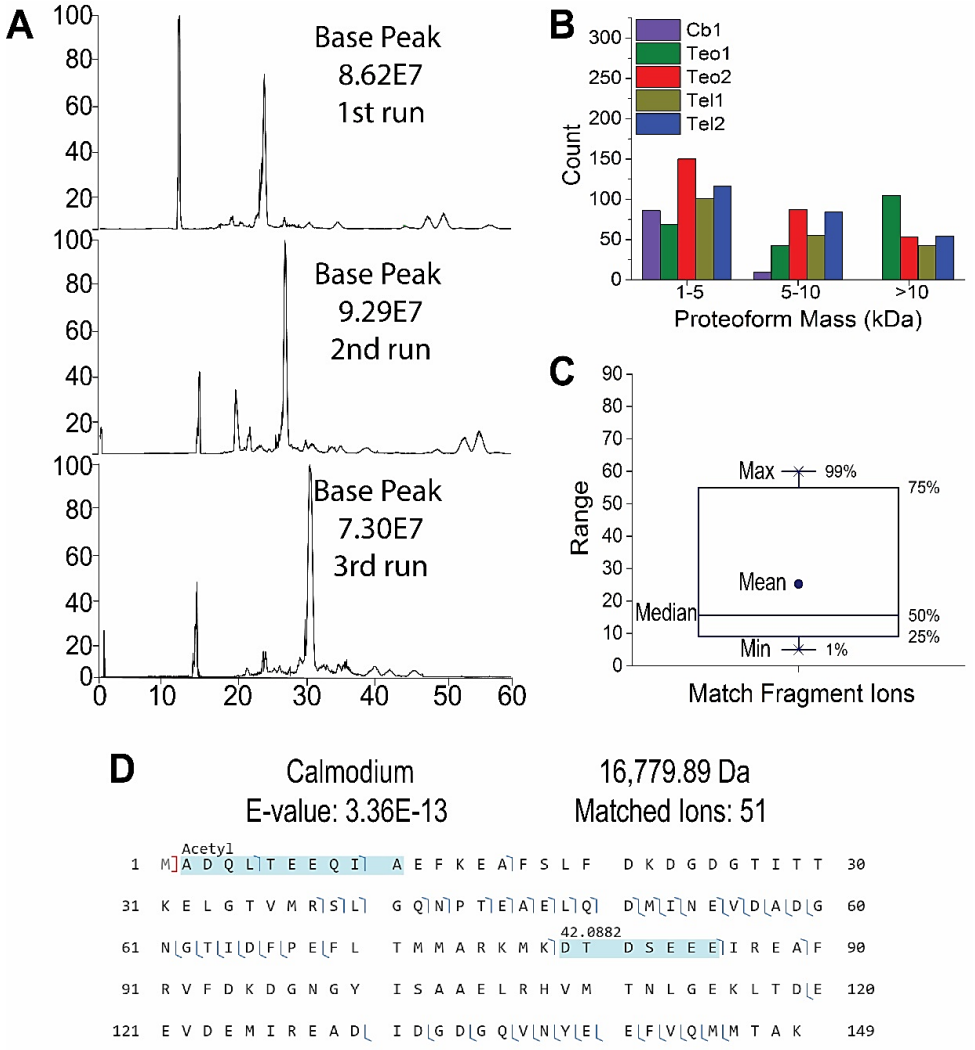


Figure 4.4. CZE-MS analyses of the LCM samples using the freeze/thaw method using a 75-cm-long and LPA-coated separation capillary. The CZE-MS/MS analyses were performed in triplicate. (**A**) Base peak electropherograms of the triplicate CZE-MS/MS runs of the Tel2 microdissected tissue sample. (**B**) Mass distribution of the identified proteoforms from one CZE-MS/MS analysis from each of the LCM samples using the freeze/thaw method. (**C**) Distribution of matched fragment ions of one CZE-MS/MS run. (**D**) Sequence and

[Figure 4.4. Continued] fragmentation pattern of protein Calmodium. There is one acetylation and a mass shift of +42 Da.

The identification rate between the microdissected tissue samples result in vastly different identification rates even when the tissue sample was dissected from the same brain region and followed the same sample handling procedure and CZE-ESI-MS/MS analysis. For the Teo brain section we identified ~197 and ~367 proteoforms from the tissue sections 1 and 2, respectively and for the Tel brain region we identified ~174 and ~255 proteoforms from tissue sections, respectively (**figure 4.3B**). The results indicate that protein population isn't homogenous throughout a specific brain region and could account for the vast difference in the proteoform identification even though the samples were obtained from the same brain region.

The TDP platform using the freeze/thaw extraction method achieved an ~35-min average separation window between the microdissected tissue samples with theoretical plate numbers of ~50,000 (**Figure 4.4A**). **Figure 4.5B** shows the reproducibility of the LFQ intensity between technical runs of the different microdissected tissue sample in the Teo region and **Figure 4.5D** shows the reproducibility of the LFQ intensity between technical runs of the different microdissected tissue sample in the Tel region. The LFQ intensity reproducibility between technical runs illustrated high reproducibility (ρ ~0.88-99), indicting a high system precision. The CZE-ESI-MS/MS system identified hundreds of proteoform identifications from the LCM samples; the system identified proteoforms with N-terminal truncations and N-terminal methionine excisions. While several common PTMs, including acetylation (+42 Da), oxidation (+16 Da), methylation (+14) was also identified.

To further investigate the freeze/thaw method, the proteoform properties were investigated. Looking at the charge state of the proteoforms for one CZE-MS/MS run from each LCM sample, ~20-40% of the proteoforms had a charge state >10, while 50-80% of the samples had charge

states <10. The majority of the proteoforms that were identified from the LCM samples did have mass lower than 10 kDa (**Figure 4.4B**). On average from all LCM samples, 26% of the proteoform identified had masses over 10 kDa. The matched fragment ions distribution of identified proteoforms from one CZE-MS/MS run is shown in **Figure 4.4C**. The mean was 25, and the median was 13. One-fourth of the proteoforms identified were matched with 10 or fewer fragment ions. The sequence and fragmentation pattern of the protein Calmodium is shown in **figure 4.4D**. This protein was identified with high confidence: E-value of 3.36E-13 and over 50 matched fragment ions. Acetylation and a mass shift of +42 were also identified on this protein.

4.3.2 Spatially Resolved Quantitative Top-Down Proteomic Analysis of Microdissected Zebrafish Brain Tissue

Next, we asked whether there were molecular differences between the two separate microdissected tissues isolated Teo and the two separate microdissected tissues isolated from Tel using our label-free quantitative top-down data. The data used was obtained from our modified workflow using the freeze/thaw extraction method (**figure 1**). Currently, there is no study that has performed using LCM coupled to CZE-ESI-MS/MS for quantitative top-down proteomics of tissues isolated with the same Zebrafish brain region. This study examined two microdissected tissue samples isolated from the same brain region to illustrate the proteoform distribution difference between the samples within the same region. We then quantitatively compared each microdissected tissue samples within the same brain region based on the proteoform feature intensity. The proteoforms from the technical duplicates that were found in all runs were combined for proteoform quantifications. A total of 29 and 30 proteoforms were quantified for Teo1/Teo2 and Tel1/Tel2, respectively.

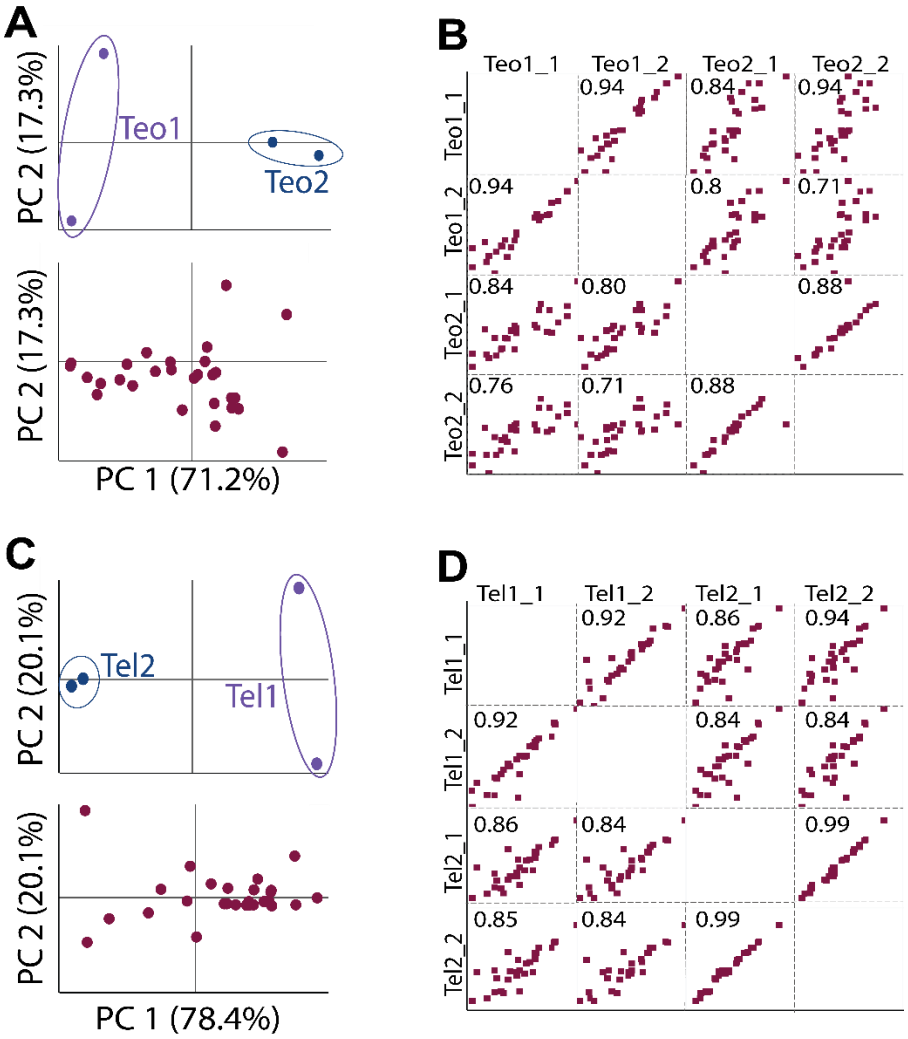


Figure 4.5. Principle component analysis of proteoform expression of the (**A**) the two different microdissected tissue samples from Teo brain region and (**C**) the two different microdissected tissue samples from Tel brain region. Pairwise correlation matrix using the log2-transformed LFQ intensities for (**B**) the two different microdissected tissue samples from Teo brain region and (**D**) the two different microdissected tissue samples from Tel. The numbers in the corner correspond to Pearson's correlation brain region.

The LFQ intensity from the brain regions can be used for comparing any kind of proteomic difference by normalizing to total proteoform amount (**Figure 4.5B and 4.5C**). However, the total protein amount that was obtained after extraction was difficult to measure due to the small volume size that was used for the proteoform extraction, therefore a correlation between the pairs was also used to validate our quantification data. As stated above, the LFQ intensity

reproducibility between technical runs illustrated high reproducibility (ρ ~0.88-99), indicting a high system precision. Pearson correlation coefficients between the two microdissected tissue samples within the same brain region show marked proteomic differences by displaying lower correlation coefficients. Between microdissected tissue samples within the same brain regions, Tel1 and Tel2 displayed correlation coefficients of 0.84 - 0.86, whereas Teo1 and Teo2 had lower correlation coefficients of 0.8 – 0.71. The correlation coefficient differences between the microdissected tissue samples within the same region of the brain illustrate a marked difference for the proteoforms quantified and display the proteoform distribution difference within the same brain region.

The quantitative data was used to compare the proteomic profiles between the microdissected tissue samples within the same region of the brain. **Figure 4.5A and 4.5C** shows the principal component analysis (PCA) of the calculated LFQ proteoform intensities for Teo1/Teo2 and Tel1/Tel2. Only the two most important principal components (PC) were chosen and it accounted for a combined total variance of 88.5% and 98.5% between the two microdissected tissue samples for Teo and Tel regions. The data points corresponding to the microdissected tissue samples originating from the same brain region formed groups that were clearly distinct from each other (**figure 4.5A and 4.5C, top panel**). The microdissected tissue samples from the same brain region clustered together within their own component without overlap with each other, suggesting proteomic differences in tissue and cell type within the same brain region based on their proteoform expression levels. **Figure 4A bottom panel** illustrates proteins that were comparably expressed between the Teo1/Teo2 at PC1 = 0, while **Figure 4C bottom panel** illustrates proteins that were comparably expressed between the Tel1/Tel2 at PC1 = 0. Proteoforms that were enriched within Teo2 (**Figure 4A bottom panel**) and Tel1 (**Figure 4C**

bottom panel) are shown at PC1 > 0, while the proteoforms enriched within Teo1 (**Figure 4A bottom panel**) and Tel2 (**Figure 4C bottom pane)** are shown at PC < 0. The proteoforms that had greater expression differences between the microdissected tissue samples can be found along PC2.

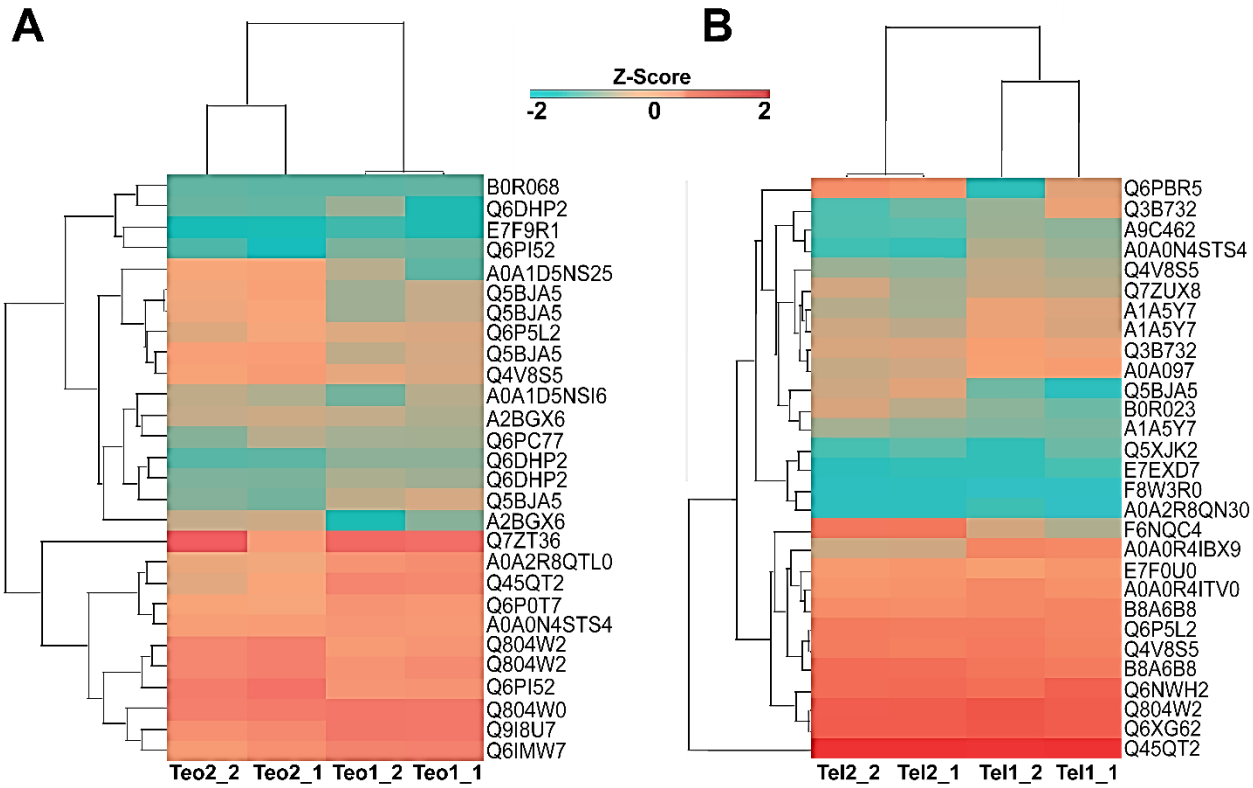


Figure 4.6. Z-score hierarchical clustering based on squared Euclidean distance measure. Each row represents one proteoform and each column represents one sample. The color scale means the proteoform expression standard deviations from the mean, with blue for low expression and red for the high expression levels. (**A**) Hierarchical clustering analysis (HCA) of the two tissue samples and their technical replicates microdissected from the Teo brain region, and (**B**) HCA of the two tissue samples and their technical replicates microdissected from the Tel brain region.

We applied a hierarchical clustering analysis (HCA) on the significantly upregulated proteoforms using Z-score log2 transformed abundances to visualize the proteoform expression differences (**Figure 4.6)**. Each of the technical duplicates of the microdissected tissue samples from the same brain regions clustered together, following the same trend as the PCA plot. When comparing the two separate microdissected tissue from the Tel region (**figure 4.6A**) and the Teo

(**figure 4.6B**), each microdissected tissue sample had a clear grouping of proteoforms with higher abundances relative to each other. This indicates the apparent differences in proteoform distribution abundances and biological function related to each specific microdissected tissue sample within the same brain regions.

Gene ontology using David Bioinformatics was performed on the upregulated proteins from each brain section; the proteins from each region of the brain showed marked differences in the two microdissected tissue samples for Teo and Tel for the molecular functions, respectively. Teo1 has proteins that were enriched in nucleotide binding and motor activity. Proteins within Teo2 had actin monomer binding and metal/calcium ion binding. Proteins in Tel1 were enriched in molecular functions such as DNA binding and nucleosome assembly, while Tel 2's upregulated proteins showed functions dealing with actin filament organization, sequestering of actin monomers, and actin cytoskeleton organization.

Table 4.1. Upregulated proteins of the microdissected tissue samples from the Teo brain region. Fold change values are log2 scale.

Protein ID	Brain Region	log_2 Fold Change	p-value	Gene ID	Protein Name	Molecular Weight (kDa)
H2B1	Teo2	1.5	2.64E-02	zgc:112234	Histone H2B	3.7
H2B1	Teo2	1.6	2.43E-02	zgc:112234	Histone H2B	4.1
H2B1	Teo2	1.8	1.49E-02	zgc:112234	Histone H2B	12.1
CALM	Teo2	2.07	2.36E-02	calm1a	Calmodulin	16.7
A0A0N4STS	Teo1	1.62	4.53E-02	ubb	Ubiquitin B	8.5
A0A1D5NS	Teo2	1.34	4.73E-04	tnnt3b	Troponin T	5.5
A0A2R8QTL	Teo1	3.32	3.81E-04	mylpfa	Myosin light chain, phosphorylatable	18
A2BGX6	Teo2	3.41	2.30E-04	myhc4	Myosin heavy chain	2.8

Table 4.1 Table 1 shows the differentially expressed proteoforms for the two microdissected tissue samples of the Teo brain region. The optic tectum is the largest part of the brain and acts as the primary visual centers. Myosin heavy chain (A2BGX6) Troponin T (A0A1D5NS), were

both highly expressed within the Teo2 section. Myosin heavy chain can associate with Troponin T which is a type of myosin cytoskeleton motor proteins, to help with neuronal migration and growth cone motility for the maturation of the CNS[31]. Calmodulin (Q6PI52) is a protein that contain EF hands, which is a specific motif that is found in a family of calcium-binding proteins. While the role of calcium-binding proteins is not fully understood, there have been studies that have shown that this family of proteins do influence retinal development and regulation[28-30].

Table 4.2. Upregulated proteins of the microdissected tissue samples from the Tel brain region. Fold change values are log2 scale.

Protein ID	Brain Region	log_2 Fold Change	p-value	Gene ID	Protein Name	Molecular Weight (kDa)
Q6XG62	Tel1	1.9	1.25E-02	icn	Protein S100	10.3
Q4V8S5	Tel1	1.37	4.19E-02	acbd7	Acbd7	9.8
Q6P5L2	Tel1	1.08	8.31E-02	dbi	Diazepam-binding	9.5
B8A6B8	Tel1	1.6	2.49E-02	tmsb2	Thymosin beta	5.1
A0A0R4IBX9	Tel1	3.31	4.85E-04	si:ch73-1a9.3	Si:ch73-1a9.3	5.3
F6NQC4	Tel2	2.4	3.98E-03	nrgna	Neurogranin, protein kinase C substrateRC3	6.9
A0A097	Tel1	2.06	8.57E-03	tmsb1	Thymosin beta	5.4
Q3B732	Tel1	1.44	3.63E-02	hmgn2	High mobility group nucleosomal-binding domain 2	3.1
A1A5Y7	Tel1	1.46	3.40E-02	hmgn7	High mobility group nucleosomal-binding domain 2	8
A1A5Y7	Tel1	2	9.94E-03	hmgn7	High mobility group nucleosomal-binding domain 2	8.1
Q3B732	Tel1	2.78	1.66E-03	hmgn2	High mobility group nucleosomal-binding domain 2	2.7
H2B1	Tel2	3.06	8.69E-04	zgc:112234	Histone H2B 1/2	4.1
Q4V8S5	Tel1	1.37	4.20E-02	acbd7	Acbd7 protein	9.9
B0R023	Tel2	1.3	4.99E-02	nedd8l	NEDD8	8.5
A9C462	Tel1	1.85	1.43E-02	cox6a1	Cytochrome c oxidase subunit	3.6
MRPB	Tel1	1.72	1.89E-02	marcksl1b	MARCKS-related protein 1-B	10.5
A0A0R4ITV0	Tel1	1.56	2.74E-02	si:dkey-46i9.1	Si:dkey-46i9.1	7.1

Table 4.2 shows the differentially expressed proteoforms for the two microdissected tissue samples of the Tel brain region. The telencephalon contains the largest portion of the central nervous system and consists of the cerebral cortex, subcortical white matter and basal nucleus. The protein Q3B732, which is a high mobility group nucleosomal-binding domain 2, and Thymosin (Q45QT2) were found to be highly expressed in Tel1 compared to Tel2. In addition, four proteoforms of this protein family were upregulated within Tel1. This family of proteins protect against microcephaly by maintaining accessibility to chromatin[26], while Thymosin has a protective effect for the CNS by regulating neurogenesis and tissue growth by upregulating miR-200a expression[27]. A0A0R4IBX9 is a non-characterized protein that is predicted to have nucleosomal DNA binding activity, which correlates with being upregulated in Tel1 considering that multiple isoforms of high mobility group nucleosomal-binding domain 2 were also found to be upregulated. NEDD8 (B0R023) was found to be upregulated in Tel2 is highly expressed in the hippocampal pyramidal neurons; and a study found that when NEDD8 migrates from the nucleus to aggregate within the cytoplasm of the cell, it may lead to excessive levels of interleukin-1β causing hyper-ubiquitination which may act as a driver in early Alzheimer-related neuropathogenesis of Down's syndrome pathogenesis.

4.5 Conclusion

In this study, we present for the first time a highly sensitive modified sample preparation workflow for top-down proteomics using laser capture microdissected tissue samples of Zebrafish brain tissue. This workflow utilized OG detergent, a MS-compatible detergent, for cell lysis using a freeze/thaw method for protein extraction. This workflow eliminated the necessity of detergent removal before CZE-ESI-MS/MS analysis resulting in a lower sample loss for mass limited samples using top-down proteomics. This modified workflow identified an average of

~220 proteoforms of laser captured microdissected tissue sections (500-µm tissue section) when <250 cells were injected onto the capillary demonstrating the sensitivity of this platform for mass limited samples. The LFQ intensity reproducibility between technical runs illustrated high reproducibility (ρ ~0.88-99). The CZE-ESI-MS/MS system identified proteoforms with N-terminal truncations and N-terminal methionine excisions and several common PTMs, including acetylation (+42 Da), oxidation (+16 Da), methylation (+14) was also identified. This procedure facilitated quantitative top-down proteomics that produced protein expression profiles that can efficiently distinguish between different microdissected tissue sections even when the sample were isolated from the same brain region. The work here described a process that can used to speculate about the tissue microenvironment and how biological functions can fluctuate from small cellular changes in the spatial context of primary tissue samples such as the human brain.

4.6 Acknowledgements

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 would also like to thank Ms. Amy Porter from the Histology lab at Michigan State University for her help with the cryosection of the Zebrafish brain and Dr. Melinda Frame from the Center for Advanced Microscopy from Michigan State University for her help with the microdissecting of the brain tissue using LCM.

REFERENCES

REFERENCES

1. Davis, S.; Scott, C.; Ansorge, O.; Fischer, R. Development of a Sensitive, Scalable Method for Spatial, Cell-Type-Resolved Proteomics of the Human Brain. *J. Proteome Res.* **2019**, *18* (4), 1787–1795.

2. Delcourt, V.; Franck, J.; Quanico, J.; Gimeno, J.-P.; Wisztorski, M.; Raffo-Romero, A.; Kobeissy, F.; Roucou, X.; Salzet, M.; Fournier, I. Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions. *Mol. Cell. Proteomics* **2018**, *17* (2), 357–372.

3. Clair, G.; Piehowski, P. D.; Nicola, T.; Kitzmiller, J. A.; Huang, E. L.; Zink, E. M.; Sontag, R. L.; Orton, D. J.; Moore, R. J.; Carson, J. P.; Smith, R. D.; Whitsett, J. A.; Corley, R. A.; Ambalavanan, N.; Ansong, C. Spatially-Resolved Proteomics: Rapid Quantitative Analysis of Laser Capture Microdissected Alveolar Tissue Samples. *Sci. Rep.* **2016**, *6* (1).

4. Zhu, Y.; Dou, M.; Piehowski, P. D.; Liang, Y.; Wang, F.; Chu, R. K.; Chrisler, W. B.; Smith, J. N.; Schwarz, K. C.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Smith, R. D.; Qian, W.-J.; Kelly, R. T. Spatially Resolved Proteome Mapping of Laser Capture Microdissected Tissue with Automated Sample Transfer to Nanodroplets. *Mol. Cell. Proteomics* **2018**, *17* (9), 1864–1874.

5. Satija, R.; Farrell, J. A.; Gennert, D.; Schier, A. F.; Regev, A. Spatial Reconstruction of Single-Cell Gene Expression Data. *Nat. Biotechnol.* **2015**, *33* (5), 495–502.

6. Brockington A.; et al. Unravelling the enigma of selective vulnerability in neurodegeneration: motor neurons resistant to degeneration in ALS show distinct gene expression characteristics and decreased susceptibility to excitotoxicity. *Acta Neuropathol.* **2013**, 125, 95–109.

7. Surmeier D. J.; Obeso J. A.; Halliday G. M. Selective neuronal vulnerability in Parkinson disease. *Nat. Rev. Neurosci.* **2017**, 18, 101–113.
8. Akila Parvathy Dharshini S.; Taguchi Y.; Michael Gromiha M. Exploring the selective vulnerability in Alzheimer disease using tissue specific variant analysis. *Genomics* **2018**

9. Hosp F.; et al. Spatiotemporal Proteomic Profiling of Huntington's Disease Inclusions Reveals Widespread Loss of Protein Function. *Cell Rep*. **2017**, 21, 2291–2303

10. Dupree, E. J.; Jayathirtha, M.; Yorkey, H.; Mihasan, M.; Petre, B. A.; Darie, C. C. A Critical Review of Bottom-up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes* **2020**, *8* (3), 14.

11. Gillet, L. C.; Leitner, A.; Aebersold, R. Mass Spectrometry Applied to Bottom-up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **2016**, *9* (1), 449–472.

12. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R., 3rd. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394.

13. Armirotti, A.; Damonte, G. Achievements and Perspectives of Top-down Proteomics. *Proteomics* **2010**, *10* (20), 3566–3576.

14. Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **2016**, *9* (1), 499–519.

15. Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445* (4), 683–693.

16. 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.; Paša-Tolić, 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. *Nat. Methods* **2019**, *16* (7), 587–594.

17. Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; Sun, L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z.; Yu, D.; Shortreed, M. R.; Smith, L. M. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19* (10), e1800361.

18. 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* **2017**, *1498*, 99–110.

19. Zhou, M.; Uwugiaren, N.; Williams, S. M.; Moore, R. J.; Zhao, R.; Goodlett, D.; Dapic, I.; Paša-Tolić, L.; Zhu, Y. Sensitive Top-down Proteomics Analysis of a Low Number of Mammalian Cells Using a Nanodroplet Sample Processing Platform. *Anal. Chem.* **2020**, *92* (10), 7087–7095.

20. Molecular Imaging by Mass Spectrometry — Looking beyond Classical Histology. *Nat. Rev. Cancer*. **2010**, *10*, 639–646.

21. Zhu, Y.; Piehowski, P. D.; Zhao, R.; Chen, J.; Shen, Y.; Moore, R. J.; Shukla, A. K.; Petyuk, V. A.; Campbell-Thompson, M.; Mathews, C. E.; Smith, R. D.; Qian, W.-J.; Kelly, R. T. Nanodroplet Processing Platform for Deep and Quantitative Proteome Profiling of 10–100 Mammalian Cells. *Nat. Commun.* **2018**, *9* (1).

22. Zhu, Y.; Clair, G.; Chrisler, W. B.; Shen, Y.; Zhao, R.; Shukla, A. K.; Moore, R. J.; Misra, R. S.; Pryhuber, G. S.; Smith, R. D.; Ansong, C.; Kelly, R. T. Proteomic Analysis of Single

Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS. *Angew. Chem. Int. Ed Engl.* **2018**, *57* (38), 12370–12374.

23. Ye, H.; Mandal, R.; Catherman, A.; Thomas, P. M.; Kelleher, N. L.; Ikonomidou, C.; Li, L. Top-down Proteomics with Mass Spectrometry Imaging: A Pilot Study towards Discovery of Biomarkers for Neurodevelopmental Disorders. *PLoS One* **2014**, *9* (4), e92831.

24. 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. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (8), 1435–1445.

25. Xu, B. J. Combining Laser Capture Microdissection and Proteomics: Methodologies and Clinical Applications. *Proteomics Clin. Appl.* **2010**, *4* (2), 116–123.

26. De Marchi, T.; Braakman, R. B. H.; Stingl, C.; van Duijn, M. M.; Smid, M.; Foekens, J. A.; Luider, T. M.; Martens, J. W. M.; Umar, A. The Advantage of Laser-Capture Microdissection over Whole Tissue Analysis in Proteomic Profiling Studies. *Proteomics* **2016**, *16* (10), 1474–1485.

27. Meyers, J. R. Zebrafish: Development of a Vertebrate Model Organism: Zebrafish: Development of a Vertebrate Model Organism. *Curr. Protoc. Essent. Lab. Tech.* **2018**, *16* (1), e19.

28. Gao, X.-L.; Tian, W.-J.; Liu, B.; Wu, J.; Xie, W.; Shen, Q. High-Mobility Group Nucleosomal Binding Domain 2 Protects against Microcephaly by Maintaining Global Chromatin Accessibility during Corticogenesis. *J. Biol. Chem.* **2020**, *295* (2), 468–480.

29. Gui-hong Zhang, Krishna Dilip Murthy, Rahmawati Binti Pare, Yi-hua Qian. Protective Effect of Tβ4 on Central Nervous System Tissues and Its Developmental Prospects. *European Journal of Inflammation* **2020**

30. Bhoyar, R. C.; Jadhao, A. G.; Sivasubbu, S.; Singh, A. R.; Sabharwal, A.; Palande, N. V.; Biswas, S. Neuroanatomical Demonstration of Calbindin 2a- and Calbindin 2b- Calcium Binding Proteins in the Early Embryonic Development of Zebrafish: MRNA Study. *Int. J. Dev. Neurosci.* **2017**, *60*, 26–33

31. Di Donato, V.; Auer, T. O.; Duroure, K.; Del Bene, F. Characterization of the Calcium Binding Protein Family in Zebrafish. *PLoS One* **2013**, *8* (1), e53299.

32. Britto, L. R.; Gobersztejn, F.; Karten, H. J.; Cox, K. Depletion and Recovery of the Calcium-Binding Proteins Calbindin and Parvalbumin in the Pigeon Optic Tectum Following Retinal Lesions. *Brain Res.* **1994**, *661* (1–2), 289–292.

33. Sittaramane, V.; Chandrasekhar, A. Expression of Unconventional Myosin Genes during Neuronal Development in Zebrafish. *Gene Expr. Patterns* **2008**, *8* (3), 161–170.

34. Balasubramaniam, M.; Parcon, P. A.; Bose, C.; Liu, L.; Jones, R. A.; Farlow, M. R.; Mrak, R. E.; Barger, S. W.; Griffin, W. S. T. Interleukin-1β Drives NEDD8 Nuclear-to Cytoplasmic Translocation, Fostering Parkin Activation via NEDD8 Binding to the P-Ubiquitin Activating Site. *J. Neuroinflammation* **2019**, *16* (1), 275

CHAPTER 5: The Future of Capillary-Zone Electrophoresis-Electrospray Ionization Tandem Mass Spectrometry for Top-Down Proteomics

The use of CZE-MS for proteomic research has been renewed due to the advances in MS technologies that has provided a need for separation techniques that provide high resolving power and speed. The advancements in CE-MS interfaces have been proven to be robust, reproducible and sensitivity, enabling the detection of complex and mass limited sample material resulting in highly confident identification and detection of peptides, proteoforms and protein complexes with ease. The sensitivity issue of top-down proteomics necessitates the need for multiple dimensions of separations to reduce the sample complexity, therefore combinations of multidimensional LC separations combined with CZE will provide a basis for reducing sample complexity and improving peak capacity. This improved peak capacity will provide better proteoform separation for the identification and characterization of low abundance and larger proteoforms for deep proteome coverage.

Top-down proteomics suffers from a lack of comprehensive sequence coverage for proteoforms identified from complex mixtures. Mass spectrometers with a combination of different gas phase fragmentations techniques is one way to increase proteoform sequence coverage for highly confident characterization. Integrating CZE separation before mass spectrometry analysis for multiple fragmentation will drastically improve the identification and characterization of proteoforms and significantly improve top-down proteomics.

Given that native separations are compatible with CZE separations, applying CE to native top-down proteomics will help to provide information about protein complex conformation and

dynamics. While reduced top-down proteomics provides structural information about proteins and in addition about proteins covalently attached to each other, native top-down proteomics can provide information about non-covalent interactions between protein subunits. Integration of native and reduced top-down proteomics will provide complementary information to provide more comprehensive structural information for interdisciplinary studies, such as proteomics and structural biology.

Due to the proteomics field, the technological future of CZE-MS is promising. Further developments for CZE-MS for top-down proteomics is necessary, however this separation technique is quite promising for the delineation of proteoforms and peptides. Soon, this technique will move into biological application and away from technical development. CZE-MS for top-down proteomics will aid in understanding the various roles that proteoforms and protein complexes play in clinical developments, such as disease pathology.

ProQuest Number: 28492038

INFORMATION TO ALL USERS
The quality and completeness of this reproduction is dependent on the quality
and completeness of the copy made available to ProQuest.



Distributed by ProQuest LLC (2021).
Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license
or other rights statement, as indicated in the copyright statement or in the metadata
associated with this work. Unless otherwise specified in the copyright statement
or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization
of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA