ADVANCING CAPILLARY ELECTROPHORESIS-MASS SPECTROMETRY FOR TOP-DOWN PROTEOMICS: TECHNICAL DEVELOPMENT AND BIOLOGICAL APPLICATIONS

By

Tian Xu

A DISSERTATION

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

Chemistry - Doctor of Philosophy

2023

ABSTRACT

Top-down proteomics (TDP) enables the proteome profiling of biological subjects at the proteoform level and understanding of differential functions associated with proteoform heterogeneity, such as sequence variation, post-translational modifications (PTMs), etc. Drastic advances on TDP technologies (e.g. sample preparation, separation/fractionation, fragmentation, bioinformatics, etc.) have been achieved in the past decades. Further improvements in separation remain desired for better analysis throughput and deeper proteome coverage. Capillary electrophoresis (CE), including capillary zone electrophoresis (CZE) and capillary isoelectric focusing (cIEF), provide superior separation performance for proteoforms. This dissertation focuses on the advancement of CE-MS-based tools on throughput, separation resolution, and capacity for TDP and utility of these tools for biological applications.

In Chapter 2, we developed high-throughput and high-capacity cIEF-MS/MS platforms. The high-throughput platform enables efficient identification and quantification of proteoforms (less than one hour per run), whereas the high-capacity cIEF-MS/MS provides large number of proteoform identifications (IDs, more than 700 proteoforms in a single shot analysis) which is valuable for deep TDP. In Chapter 3, we further improved the stability and robustness of cIEF-MS platform using optimized linear polyacrylamide (LPA) capillary coating and catholyte with lower pH (pH~10). The work achieved high-resolution characterization and accurate isoelectric point (pI) determination of charge variants (~0.1 pI difference) of monoclonal antibodies (mAbs). In Chapter 4, we developed a nondenaturing cIEF-MS platform for ultrahigh resolution characterization of microheterogeneity of a variety of protein complexes. Typically, pI determinations of variants in protein complexes allow us to decipher how sequence or PTM variations modulate the pIs of the protein complexes. In Chapter 5, while CZE-MS/MS is a well-developed approach, for the first time, we coupled FAIMS to CZE-MS/MS to facilitate online gas-phase fractionation of proteoforms. The FAIMS greatly enhanced the sensitivity of the system and expanded the number of proteoform IDs, especially large proteoform IDs. The work renders CZE-FAIMS-MS/MS as a new powerful multidimensional platform for deep TDP.

In Chapters 6 and 7, we applied cIEF-MS/MS and CZE-MS/MS for studying the sexual dimorphism of zebrafish brains and proteoform-level differences between metastatic and nonmetastatic colorectal cancer (CRC) cells, respectively. In Chapter 6, quantitative TDP of thousands of proteoforms from male and female zebrafish brains by cIEF-MS/MS based approach discovered various overexpressed proteoforms in male or female brains that are closely associated with hormone activity. In Chapter 7, We performed deep TDP study of non-metastatic and metastatic CRC cells (SW480 and SW620) using CZE-MS/MS based multidimensional

platform and identified more than 20,000 proteoforms of over 2,000 proteins from the two cell lines, which presents around 5-folds higher number of proteoform IDs in comparison with previous TDP studies of human cancer cells. The work revealed significant discrepancies between the two isogenic cell lines regarding proteoform and single amino acid variant (SAAV) profiles. Quantitative data disclosed differentially expressed proteoforms between the two cell lines and their corresponding genes were connected to cancer pathways and networks.

This thesis is dedicated to my parents and my friends who gave me unconditional support throughout my life.

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest appreciation to my advisor Prof. Liangliang Sun for his support and guidance in my research and graduate education. I feel so lucky to join his lab where I found my interest in amazing CE-MS technologies. In the past years, he taught me extensive instrumental knowledge and troubleshooting skills. With his support, I got precious opportunities to present my work at conferences and interact with people in the field. I have become more confident under his encouragement and learned to enjoy both my research and life. His optimism and passion for science inspired me to overcome every obstacle and continue to dig into my potential in science.

I would like to thank my collaborators Dr. Linjie Han, Dr. Alayna Geoge Thompson, and Dr. Qunying Zhang from AbbVie for their help and funding support on the projects of developing high-resolution cIEF-MS for mAbs and ADC characterization. Particularly, many thanks to Dr. Linjie Han for valuable discussions and suggestions on improving our data quality and manuscripts.

I would like to thank Prof. Xiaowen Liu from Tulane University, who helped us with data analysis on zebrafish brain and colorectal cancer projects. I really appreciate Prof. Amanda Hummon from Ohio State University for providing colorectal cancer cells and Prof. Jose Cibelli from Michigan State University for the zebrafish samples. I want to thank Dr. James Xia from CMP Scientific for the suggestions on CE-MS, and Dr. Joseph Beckman, Dr. Valery Voinov, Blake Hakkila, and Mike Hare from e-MSion for helping with the development of ECciD on Q-TOF.

I am also thankful to my committee members Prof. Heedeok Hong, Prof. Xiangshu Jin, and Prof. Greg Swain. Their instructions in my first-year committee meeting and comprehensive exam are very important to me for being a qualified Ph.D. candidate and making the progress on my research projects.

Thanks should also go to all of my group members. Xiaojing, Daoyang, Zhichang, Eli, and Rachele taught me plenty of experimental skills when I joined the lab. In particular, I gained precious experience on native CE-MS from Xiaojing. Daoyang and Zhichang shared with me information and advice on career development after their graduation. I would like to thank my friend Qianjie, who is my source of happiness and provided massive help in my life. I also want to thank Qianyi Wang, Jorge A. Colón-Rosado, Dr. Fei Fang, Olivia Gordon, Seyed Sadeghi. I really enjoyed our time working, helping each other, and exchanging ideas in the lab.

I want to thank Dr. Fan Zhang and Dr. Daniela Tomazela from Merck for mentorship during my internship. I really benefited from this program with deeper insight into the potential of CE-MS

for studying a variety of interesting biologics. Dr. Fan Zhang particularly helped me a lot in both experiments and building networks with talented scientists in biopharma.

Lastly, I am so grateful to my parents for giving me endless love, always being supportive to my decisions, and cheering me up in the journey of life. I am so sorry that I could not keep them company during this period and really wish I have more chances to stay with them and bring them happiness in the future.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AA	Acetic acid
ABC	Ammonium bicarbonate
ACN	Acetonitrile
ADC	Antibody-drug conjugate
AI-ETD	Activated ion electron transfer dissociation
BGE	Background electrolyte
BPE	Base peak electropherograms
BUP	Bottom-up proteomics
CA	Carbonic anhydrase II
CCS	Collision cross-sections
CE	Capillary electrophoresis
CID	Collision-induced dissociation
cIEF	Capillary isoelectric focusing
CRC	Colorectal cancer
CV	Compensation voltage
Cyt c	Cytochrome c
CZE	Capillary zone electrophoresis
DAR	Drug-to-antibody ratio
DDA	Data-dependent acquisition
DTT	Dithiothreitol
DV	Dispersion voltage
E. coli	*Escherichia coli*
ECD	Electron capture dissociation
EIE	Extracted ion electropherograms
EMR	Orbitrap extended mass range
EOF	Electroosmotic flow
ESI	Electrospray ionization
ETD	Electron transfer dissociation
ETnoD	Electron transfer without dissociation
FA	Formic acid
FAIMS	High-field asymmetric waveform ion mobility spectrometry
FASS	Field-amplified sample stacking
FDR	False discovery rate

FT	Fourier transform
FTICR	Fourier transform ion cyclotron resonance
GELFrEE	Gel-eluted liquid fraction entrapment electrophoresis
GO	Gene ontology
HCD	Higher energy collisional dissociation
HIC	Hydrophobic interaction chromatography
HPC	Hydroxypropyl cellulose
IAA	Iodoacetamide
ID	Identification
IEX	Ion exchange chromatography
LC	Liquid chromatography
LCM	Laser capture microdissection
LOD	Limit of detections
LPA	Linear polyacrylamide
LTQ	Linear ion trap
mAb	Monoclonal antibody
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
Myo	Myoglobin
PEPPI	Polyacrylamide gels as Intact species
pI	Isoelectric point
PTM	Post-translational modification
PVA	Poly (vinyl alcohol)
RNA	Ribonucleic acid
RPLC	Reversed-phase liquid chromatography
RSD	Relative standard deviation
SA	Streptavidin
SAAV	Single amino acid variant
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TDP	Top-down proteomics
TFA	Trifluoroacetic acid

tITP	Transient isotachophoresis
TMT	Tandem Mass Tag
TOF	Time-of-flight
UV	Ultraviolet
UVPD	Ultraviolet photodissociation

CHAPTER 1. Introduction

1.1 Top-down proteomics (TDP) and technological challenges

Proteins direct participants in biological events and are central intermediaries connecting genotype and phenotype. A single protein-coding gene can derive various forms of proteins, also termed proteoforms, due to genetic variations/single nucleotide polymorphisms (SNPs), alternatively spliced ribonucleic acid (RNA) transcripts, post-translational modifications (PTMs), and truncations (Figure 1.1).[1,2] The proteoform diversity can result in distinct protein functions and activities.[3-5] For instance, dynamic phosphorylated histones in humans, such as phosphorylated H3Y41, regulate the transcription processes by promoting the unwrapping of nucleosomes.[6] Certain PTMs (e.g. phosphorylation, ubiquitination) on alpha-synuclein are associated with a high incidence of neurodegeneration in Parkinson's disease.[7] Mapping the proteoforms is crucial for understanding their functions in the biological systems, deciphering the disease pathologies, and developing targeted therapeutics.[4-9] Recently, Human Proteoform Project has been proposed to construct human proteoform atlases and discover the proteoforms related to the disease.[3] As the human proteome is extremely complex and diverse, which comprises millions of proteoforms across a billion-fold dynamic range,[5] the development and advancement of technologies are highly desired to enlarge proteoform characterization for the project.

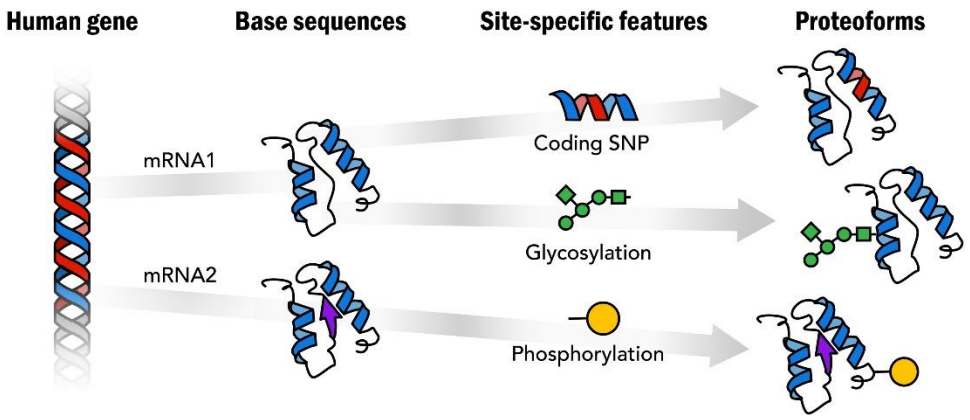

Figure 1.1 Diverse proteoforms derived from a single gene. Reproduced with permission from reference (3).

Bottom-up proteomics (BUP) is well-established for precisely interrogating the PTMs on peptides from digested proteins but fails to capture the combinatorial PTMs of individual proteoforms due to the "peptide-to-protein inference problem".[4,10,11] In contrast, top-down proteomics (TDP), which delineates intact proteoforms, reflects comprehensive information on

primary sequence and PTM variations. TDP relies on mass spectrometry (MS)-based approaches for proteoform identification and quantification.[12-16] A typical MS-based TDP workflow is shown in Figure 1.2. Briefly, proteins are extracted from cells or tissues and then separated using either liquid chromatography (LC) or capillary electrophoresis (CE). After protein molecules are ionized through electrospray ionization (ESI) and introduced into a mass spectrometer (e.g. Orbitrap), the masses of intact proteoforms are measured through full MS acquisition, and proteoforms of interest are further isolated and fragmented by tandem mass spectrometry (MS/MS). The proteoforms can be identified based on accurate masses of precursor and fragment ions by database searching.

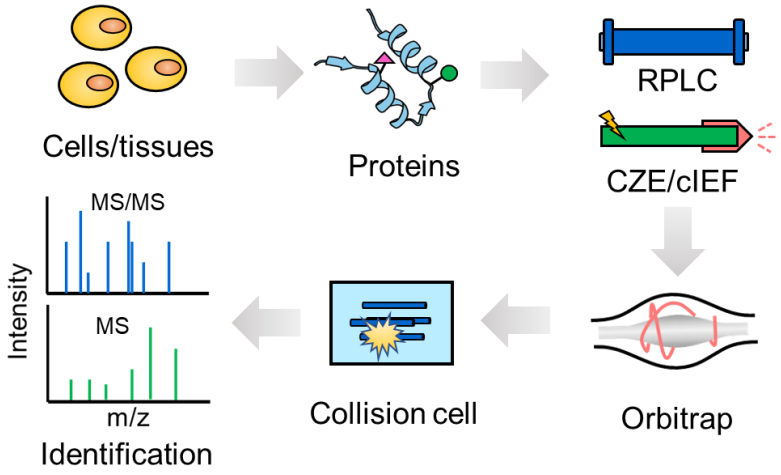

Figure 1.2 Workflow of TDP.

At present, TDP remains restricted in the throughput and depth of proteoform analysis.[3,8,17] TDP studies generally require a long sample preparation (several days) and instrumentation time (several days to months). Developments of high-throughput TDP workflows with simple sample preparation procedures and rapid sample fractionation/ separation are crucial for efficient profiling of proteome and facilitating clinic applications. On the other hand, considering the high complexity and dynamic range of proteome, the approaches for deep TDP are required to expand proteoform identification. Currently, interrogation of the low-abundance or large proteoforms (>30 kDa) is extremely difficult.[16,18,19] The advancement of separation resolution and capacity of front-end fractionation/separation is the key to improving proteome coverage by reducing sample coelution.[4,14,16] Different LC (e.g. reversed-phase liquid chromatography, RPLC) and CE (e.g. capillary zone electrophoresis, CZE) methods have been extensively coupled to MS for TDP. CE is potentially better suitable for proteoform separation than LC. The separation of CE in capillary causes less sample diffusion and sample loss than an LC column with the stationary phase,

thereby providing higher separation resolution and sensitivity.[20-24] However, technical issues of CE-MS for TDP remain. For example, the sample loading capacity of CZE is low and needs to be improved for complex samples.[21-23] For another example, the automation, sensitivity, and stability issues of capillary isoelectric focusing (cIEF)-MS needs to be addressed.[25-27] Besides liquid-phase separation, high-field asymmetric waveform ion mobility spectrometry (FAIMS), a gas-phase fractionation technique, has recently emerged as a promising option for constructing multidimensional platforms with LC/CE to implement high throughput or deep TDP.[28-31]

More details about the principles, technological challenges, and method improvements in separation, MS and MS/MS, CE-MS interface for ESI, as well as applications of CE-MS will be introduced in the following sections.

1.2 Separation approaches

1.2.1 Liquid chromatography (LC)

Liquid chromatography (LC) is a prevalent separation technique with high capacity and reproducibility for TDP. LC utilize packed column for separation and the principle of separation can be explained by van Deemter equation (Equation 1.1):

$$H = A + B/v + Cv \qquad \textbf{Equation 1.1}$$

where H is plate height, A is the eddy diffusion parameter, which is related to different paths that molecules flow through the column, B is the longitudinal diffusion parameter, which is associated with the band broadening in the mobile phase from the central region to neighboring zones, C is the resistance to mass transfer between mobile and stationary phase, v is the linear velocity.[32]

The plate number is used to describe the separation efficiency of LC (Equation 1.2):

$$N = L/H \qquad \textbf{Equation 1.2}$$

where N is the number of theoretical plates, L is column length, and H is the plate height. Lowering the plate height benefits higher separation efficiency.[32] LC typically achieves plate numbers around 10^3 to 10^4 plates/m.[33]

LC provides various separation options based on proteins' hydrophobicity (e.g. reversed-phase liquid chromatography, RPLC[17,34,35]; hydrophobic interaction chromatography, HIC), size (e.g. size exclusion chromatography, SEC[36-37]), and ionic strength (e.g. ion exchange chromatography, IEX[38-41]). Among them, RPLC and SEC are the most popular separation approaches for TDP.

RPLC employs the non-polar stationary phase and polar mobile phase to perform the separation. Proteoforms with different hydrophobicity have different retention on the stationary

phase. By altering the compositions of organic solvents in a gradient, they are separated at different times. RPLC for TDP remains limited by a low peak capacity (<100) and irreversible sample adsorption.[14,42] Utilizing packing beads with shorter alky chains (C1-C4) can improve protein recovery.[43] Elevating the temperature to 50-70 °C can benefit protein solubility and minimize protein adsorption.[44] The smaller size of beads provides better separation efficiency by reducing eddy diffusion and resistance to mass transfer.[34] For larger proteins, larger particle pore size generates better separation resolution by providing larger binding areas.[43] Shen et al. previously systematically investigated the factors (column length, particle size, pore size, functional group of beads) that impact the separation of *S. oneidensis* lysate.[43] They found increasing the length of the column (> 1 meter) and using a long gradient (> 10 hours) achieved the most significant improvement in separation capacities (>400) and the number of proteoform identifications (IDs) below 50 kDa (~900). However, the method requires ultrahigh-pressure (14k psi) and is time-consuming.

Alternatively, SEC enables fast separation of analytes according to their sizes. SEC columns are packed with porous silica particles coated with a neutral and hydrophilic layer. In SEC separation, molecules larger than the pores are excluded from the packed bed and elute first in the void volume. In contrast, the smaller molecules can penetrate the pores to various degrees depending on their size, with the smallest molecules diffusing furthest into the pore structure and eluting last. Multiple offline SEC or online SEC-MS/MS studies have been carried out for resolving standard protein mixtures and lysates from heart tissues, greatly benefiting the identification of various large proteoforms up to 223 kDa.[39,40] SEC has a relatively low resolution compared to many other LC methods. In many cases, it was adopted as a protein fractionation method in TDP or coupled with several other SEC columns with different pore sizes to enhance the separation resolution.[40] In addition, the undesired secondary interaction/ionic interaction between protein and stationary phase is also a great concern for SEC.[45] High concentrations of acid or additives (e.g. trifluoroacetic acid, TFA) are previously employed in studies to mitigate the ion paring effect. [40, 46]

1.2.2 Capillary electrophoresis (CE)

1.2.2.1 Capillary zone electrophoresis (CZE)

Capillary zone electrophoresis (CZE) separates proteins based on their electrophoretic mobilities in an electric field, which relate to their charge-to-size ratios (Equation 1.3). CZE separation in a fused silica capillary is determined by two factors: electrophoretic mobility (Equation 1.3), and electroosmotic flow (EOF, Equation 1.4).

$$\mu_{ep} = \frac{q}{6\pi\eta r}$$ **Equation 1.3**

where μ_{ep} is electrophoretic mobility, q is the charge of the proteoform, η is the viscosity of the background electrolyte (BGE), and r is the proteoform's radius.[47]

$$\mu_{eo} = \frac{\varepsilon\zeta}{4\pi\eta}$$ **Equation 1.4**

Where μ_{ep} is EOF, ε is dielectric constant, ζ is zeta potential, and η is the viscosity of the BGE.[48]

The EOF is induced by negatively charged silanol groups on the capillary inner wall. The negatively charged surface attracts the cations from the BGE to form a double layer, which drives the BGE along with proteoforms to migrate toward the cathode under the electric field.[49] EOF causes a narrower separation window and lower separation resolution in CZE.[20] To suppress the influence of EOF, many CZE studies modified the capillary inner wall with a layer of neutral coating (e.g. linear polyacrylamide, LPA).[20] Therefore, the separation of CZE is mainly determined by the electrophoretic mobilities of proteoforms, which are associated with their charge-to-size ratios. A single-shot CZE-MS/MS analysis of *Escherichia coli* (*E. coli*) using LPA coating previously achieved peak capacity of around 280 in a 90 min separation window and 600 proteoform identifications, presenting much better performance than the CE-MS/MS works using bare fused silica capillaries. [20,21]

CZE has better separation performance than LC (theoretical plate number: 10^5-10^6 plates/m [33] vs. 10^3-10^4 plates/min) since separation in the open tubular capillary avoids eddy diffusion and resistance to mass transfer. In particular, CZE presents high-resolution separation for some proteoforms with heterogeneity on PTMs (e.g. phosphorylation, acetylation). For example, Drown et al. compared both LC-MS/MS and CZE-MS/MS analysis of human heart tissue and indicated three phosphoproteoforms of cardiac troponin I (cTnI) which are baseline separated in CZE-MS/MS but coeluted in RPLC-MS/MS.[50] In addition, CZE-MS/MS using a 1.5-meter capillary for CZE-MS/MS previously achieved the separation of three proteoforms of myoglobin with heterogeneity on acetylation and phosphorylation.[51]

However, the sample loading capacity of CZE is much lower than LC. A typical CZE separation allows only 1% of the total capillary volume for sample loading to guarantee separation resolution, which is equivalent to 20 nL for a 1-meter capillary (50 μm i.d.).[20] In comparison, RPLC can facilitate the microliter level of sample loading, due to the capability of trapping the sample in the stationary phase. The development of high-capacity CZE methods are highly desired for better identification of low-abundance proteoforms.

Different sample preconcentration methods have recently been introduced for CZE, such as field-amplified sample stacking (FASS), transient isotachophoresis (tITP), dynamic pH junction, etc.[20] The dynamic pH junction attracts the most attention for CE-based TDP studies due to its high stacking and separation performance.[51,52] In dynamic pH junction (Figure 1.3), the sample is dissolved in a basic sample buffer (e.g. ammonium bicarbonate, pH 8), whereas the acidic solution (e.g. 5% acetic acid, pH 2.4) is employed as BGE.[53-55] The majority of proteoforms in the sample buffer are negatively charged and migrate towards pH boundary I when a positive potential is applied. Meanwhile, the protons in the BGE start to titrate the sample zone. The pH boundary I gradually shrink towards the pH boundary II and the sample is concentrated in a narrower zone. Eventually, when the titration is complete, the two boundaries are combined together and the proteoforms are positively charged to facilitate regular CZE separation. Dynamic pH junction improves sample injection to a maximum of 50% of total capillary volume.[51] Our recent work applied dynamic pH to CZE in a 1.5-meter capillary and achieved 2000 proteoform identifications from microliter level sample loading of zebrafish brains.[51] A comparison of dynamic

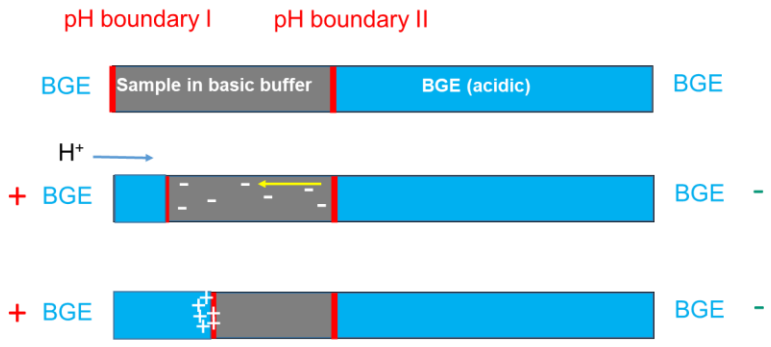

Figure 1.3 Process of dynamic pH junction.

pH junction and FESS in TDP of a standard protein mixture showed a 4-fold higher theoretical plate number in dynamic pH junction than FESS.[51] However, for CZE, the room for further improving sample loading capacity is very limited. Therefore, digging into the potential of other CE separation methods with high loading capacity such as cIEF for TDP can be an alternative solution.

1.2.2.2 Capillary isoelectric focusing (cIEF)

cIEF offers the separation of proteins on basis of their isoelectric points (pIs). cIEF provides several attractive advantages. First, cIEF can achieve ultrahigh-resolution separation for proteins with minor pI differences (as low as 0.01 pH unit)[56] and differentiate proteoforms that are structurally similar but heterogenous on charges, such as proteoforms with different PTMs.

Second, cIEF provides a higher sample loading capacity (up to 100% of capillary volume) than regular CZE (typically ~1% of capillary volume).[57] After focusing, a concentration factor of 50 -100 times can be easily obtained for analytes, which makes cIEF well suitable for analyzing low-abundance proteins.

A typical cIEF separation is performed with a two-step process: focusing and mobilization. Briefly, in the electric field, a linear pH gradient is built up with the assistance of amphoteric compounds with high buffering capacity; meanwhile, the charged proteoforms migrate until they focus on narrow zones where pHs equal to their pIs, i.e., proteoforms become neutral. Subsequentially, either hydrodynamic flow (pressure /gravity) or chemical mobilization is applied to drive proteins into a detector.[26]

Conventional cIEF systems are equipped with one-point imaging (ultraviolet (UV), fluorescence) or whole-column imaging for monitoring signals of analytes. The methods were widely applied for pI measurement and stoichiometric characterization. Integrating ultrahigh resolution cIEF with MS/MS is very appealing for protein characterization and TDP study. cIEF-MS has been pioneered by Lee and Smith group in 1990s.[56-62] However, for a long time, the cIEF-MS was performed in a semi-online manner, where the capillary outlet was first inserted in a basic catholyte for focusing, and then manually transferred to an interface filled with acidic sheath liquid for chemical mobilization and ionization. The appearance of the “sandwich” injection configuration in 2009 makes it possible to implement the fully automated cIEF-MS analysis.[63] The method was carried out by filling the capillary with a plug of MS-compatible catholyte buffer such as ammonia hydroxide, followed by a plug of the sample-ampholyte mixture (Figure 1.4). Thus, the cIEF focusing could be facilitated after applying voltage even though its outlet was installed in an interface with an acidic sheath liquid. After focusing, a low pressure (~50 mbar) or chemical mobilization was employed to drive focused proteins toward MS for detection (Figure 1.4). The chemical mobilization can automatically be initiated when cations from anolyte and anions from sheath liquid enter the capillary and gradually disrupt the pH gradient.[64]

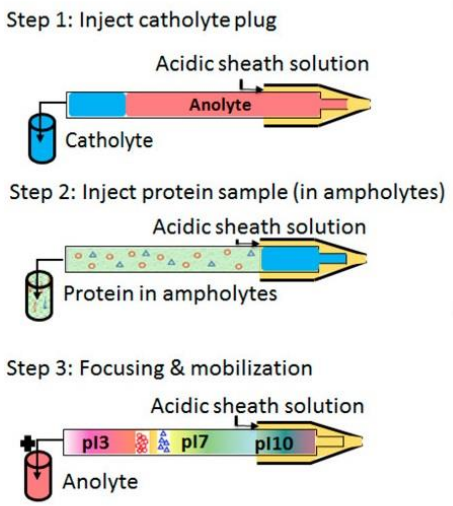

Figure 1.4 cIEF separation with "sandwich" injection configuration. Reproduced with permission from reference (116).

The widespread adoption of cIEF-MS requires drastic improvement of the system's sensitivity. First, a highly sensitive and robust ESI interface is desired for constructing cIEF-MS system. In addition, optimization of the ampholyte concentration is always necessary to mitigate signal suppression by ampholytes. Ampholytes are indispensable additives for maintaining pH gradients in cIEF. Reducing the concentration of ampholyte can significantly improve MS signal but adversely impacts separation resolution. Therefore. the ampholyte concentration has to be optimized by carefully balancing the MS signal and separation resolution. Furthermore, two-dimensional CE platforms have been recently introduced by coupling icIEF with CZE-MS, which effectively reduces the influence of ampholytes on MS analysis.[65-67]

1.2.3 High-field asymmetric waveform ion mobility spectrometry (FAIMS)

FAIMS, as a fractionation strategy in the gas phase, provides rapid and online filtering of ions based on their differential mobilities in oscillating high and low electric fields.[68-70] The FAIMS is composed of two parallel electrodes (planar or cylindrical shape) with dispersion voltage (DV) applied to one of the electrodes to deliver an asymmetric waveform (Vmax≠Vmin) (Figure 1.5A).[68] The ions carried by the carrier gas have different mobilities in high and low field segments of waveform and eventually end up colliding with the electrodes (Figure 1.5B).[68] The fractionation of ions can be facilitated by applying compensation voltage (CV) on the other electrode to offset the drift of ions and selectively allow the transmission of specific groups of ions.[68]

The differential mobility of an ion is determined by a variety of properties, such as the mass, shape, center of mass, and dipole.[29,68,71] Ideally, the dependence of ion mobility on the

electric field can be divided into three types in a specific range of electric field strengths (Figure 1.5C): motility accelerates with an increasing electric field (A-type); mobility initially increases and then gradually declines with an increasing electric field (B-type); mobility initially decreases with an increasing electric field (C-type).[68] The mobility behaviors of 10 proteins (8-66kDa) in FAIMS were evaluated previously.[71] The proteins lower than 30 kDa presented in C-type behavior, whereas proteins above 30 kDa showed A/B-type behavior. Large proteins have dipoles that can align in the strong electric field. Their mobility is determined by the collision cross sections (CCS) in the plane orthogonal to the dipole, rather than the averaged CCS. Therefore, FAIMS has the potential to separate proteins with different molecular weights based on their mobility difference.

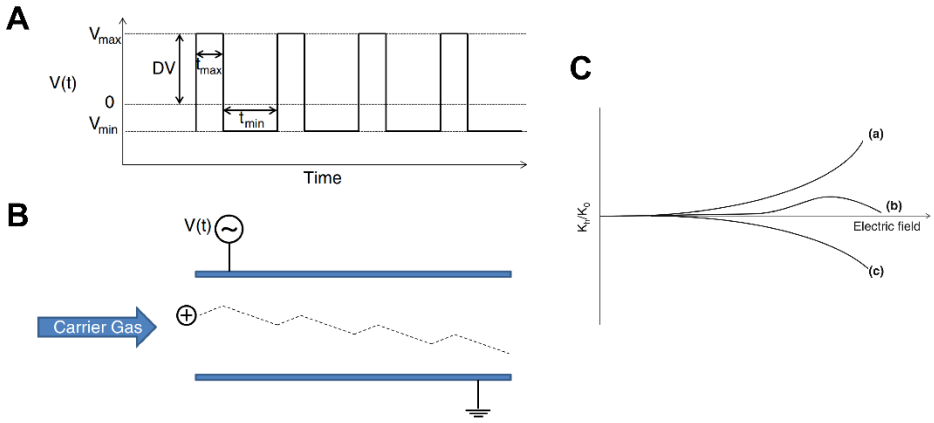

Figure 1.5 Separation mechanism in FAIMS. (A) asymmetric waveform applied to the FAIMS electrodes; (B) ion trajectory in FAIMS; (C) three types of mobility behavior in FAIMS. Reproduced with permission from reference (68).

1.3 Mass spectrometry (MS) for TDP

1.3.1 MS instrumentation

High-resolution and sensitive MS instruments are highly preferred for TDP studies. Various mass spectrometers, including Fourier transform ion cyclotron resonance (FTICR), Orbitrap, hybrid linear ion trap (LTQ) Orbitrap, and time-of-flight (TOF), have previously been applied for proteoform analysis.[13,72-74]

Early TDP was mostly carried out on FTICR mass spectrometers because of ultrahigh mass resolving power (10^5-10^6 at m/z 400) and high mass accuracy (mass error lower than 1ppm).[75,76] The principle of FTICR can be concluded in the following processes. The ions are trapped in a cell that is composed of a magnetic field and orthogonal electric trapping plates, and then are excited to a larger cyclotron radius by oscillating the electric field with radio frequency (RF) pulse. When the RF is turned off, the ions continue rotating at cyclotron frequencies which

are proportional to their z/m. The image current from ions is recorded and converted to the frequency domain using Fourier Transform (FT) and further transformed to a mass spectrum.

Orbitrap is another high-resolution FTMS analyzer (10^5 at m/z 400) and contains a barrel-like outer electrode and a spindle-like inner electrode.[77,78] Orbitrap MS utilizes a curved linear trap (C-trap) to collect and pulse the ions into the orbitrap. After injection from C-trap to the Orbitrap, the positive ions oscillate around the inner electrode with applied negative static potentials in various frequencies that relate to m/z of ions. Different ions are separated according to their differential frequencies and their image currents are determined. Compared to FTICR, the orbitrap is more affordable and has received wide applications in TDP of cells, tissues, clinic samples, and in both denaturing and native conditions. [14,17,50]

Both FTICR and Orbitrap have slow scan speeds (up to 20 Hz). In contrast, TOF can serve as a complementary instrument with fast data acquisition (higher than 1000 Hz). In TOF, ions obtain kinetic energy from acceleration by an electric field, followed by traveling through a drift tube. Their time of flight is:

$$t = \frac{d}{\sqrt{2U}}\sqrt{\frac{m}{q}} \qquad \textbf{Equation 1.5}$$

where d is the path of flight, U is the voltage for accelerating ions, m is the mass of the ion, and q is the charge of the ion.

The ions with different m/z have distinct velocities and reach the detector at different times. The length of the drift tube dictates separation resolution. The reflectron TOF typically has an ion mirror at the opposite end of the tube which can reverse the direction of ions to increase the length of the drift path.[79] In addition, the same ions may carry a distribution of kinetic energy due to different starting points during acceleration. The ions with higher energy can travel deeper into the reflectron whereas the ions with less energy have a shallow path.[79] Thus, reflectron corrects their flight time and better focuses them on the detector. Most TOF instruments equip with microchannel plate (MCP) detectors, where ions hit the plate and electrons are generated to produce amplified signal.[79]

TOF has a lower resolution (10^3-10^4)[80], compared to FTICR and Orbitrap. However, unlike FTICR and Orbitrap which have resolution decrease dramatically with increasing m/z, TOF experiences a very small decline of resolution at high m/z.[81,82] In addition, the TOF allows a much wider mass range, which makes it popular for native MS analysis of protein complexes (up to 500 MDa).[82]

1.3.2 Tandem mass spectrometry (MS/MS)

Extensive fragmentation of proteoforms by MS/MS is crucial for interrogating their sequence and PTM information during database search. In MS/MS, the ions of interest (precursor ions) are isolated in a mass analyzer and sent to the collision cell for fragmentation. The produced fragments were further introduced to the second mass analyzer for detection. A variety of techniques have been developed for fragmenting protein backbones, including collision-induced dissociation (CID), higher energy collisional dissociation (HCD), electron capture dissociation (ECD), electron transfer dissociation (ETD), ultraviolet photodissociation (UVPD), etc.[14,74]

CID and HCD are similar approaches that collide protein precursor ions with neutral gas (nitrogen, helium, or argon) for fragmentation. Concisely, the isolated precursor ions are accelerated by potential and undergo multiple collisions with neutral gas. The process deposits the kinetic energy to the backbone as vibrational internal energy. The backbone cleavage occurs when the energy is sufficient to overcome the barrier for dissociation. CID and HCD typically cause cleavage of C−N amide bonds and generate b and y type of fragments.[14] As CID and HCD favor cleavage of the most labile bonds of a protein, they generally present limited sequence coverage for large proteins.[14]

ECD and ETD are electron-driven fragmentation techniques that provide extensive fragmentation of protein backbone and better preservation of the labile PTMs, such as phosphorylation and glycosylation.[83-86] In ECD , electrons are captured at the protein protonated sites. The energetic hydrogen atoms (H•) are ejected from the proteins and are further captured at high-affinity sites of the proteins (e.g. backbone amide), leading to cleavages of the N-Cα bond and generation of c and z• ions.[87] The fragmentation mechanism of ETD is similar to ECD. However, unlike ECD, where protein directly captures the electrons ejected from the filament, ETD uses radical anions to deliver the electrons to protein ions. ECD and ETD are frequently coupled with activation/dissociation methods to further improve fragmentation efficiency. For example, in ETD, noncovalent interactions across the precursor ion can prevent the separation of fragment ions, resulting in electron transfer without dissociation (ETnoD). Using higher energy infrared photons to vibrationally activate the precursor, also called activated ion (AI)-ETD, can effectively disrupt the noncovalent interactions during ETD. [88-90] Alternatively, all ETD products are further fragmented using HCD (EThcD) to increase sequence coverage.[91]

UVPD utilizes UV laser source (wavelength of 193nm or 213 nm) to excite precursor ions with high-energy photons.[92] UVPD can generate all types of ions (a, x, b, y, c, z•) from direct backbone cleavage and other ions (d, v, w ions) from secondary fragmentation, thereby providing high fragmentation efficiency. In addition, UVPD has a low dependency on the charge of the

protein and can well preserve the labile PTMs. Therefore, UVPD has been considered a powerful fragmentation tool for proteoform characterization. UVPD and HCD were compared previously in TDP of HeLa cell lysate.[93] Particularly, UVPD presented better sequence coverage and confidence in proteoform identification than HCD.[93]

1.4 CE-MS interfaces for electrospray ionization (ESI)

ESI is a soft ionization process that produces intact multiply charged gas-phase ions from analytes in solution.[94,95] The process of ESI is illustrated in Figure 1.6. For ESI, an electric potential is applied between the emitter and mass spectrometer. The solution at the tip of the emitter is occupied with positively charged ions and forms a Taylor cone where a jet of liquid droplets is emitted (micrometer radius size). With the assistance of heat and drying gas, the solvent evaporates from the droplets quickly and the droplets shrink to smaller sizes. When the charge density in a droplet increases to a limit at which the surface tension of the droplet cannot hold charge repulsion on the surface (Rayleigh limit), the droplet will experience an explosion and form even smaller droplets (nanometer radius size). After certain rounds of the explosion, the analytes in the droplets are eventually released into the gas phase.

Online hyphenation of CE and MS or MS/MS requires a well-designed interface that establishes electrical continuity in CE separation and facilitates ESI. Two types of ESI interfaces are available for CE-MS: sheath-flow and sheathless.96,97 Sheath-flow interfaces use sheath liquid to maintain electric contact and assist ionization. They are most widely adopted in CE-MS for their high robustness and good stability. However, sheath-flow interfaces are challenged by decreased sensitivity due to the dilution of effluent by sheath liquid. In contrast, sheathless interfaces can achieve superior sensitivity as it does not use sheath liquid, but they encounter several issues such as fragile tips, instability of electrospray owing to low flow rate, and bubbles formation.

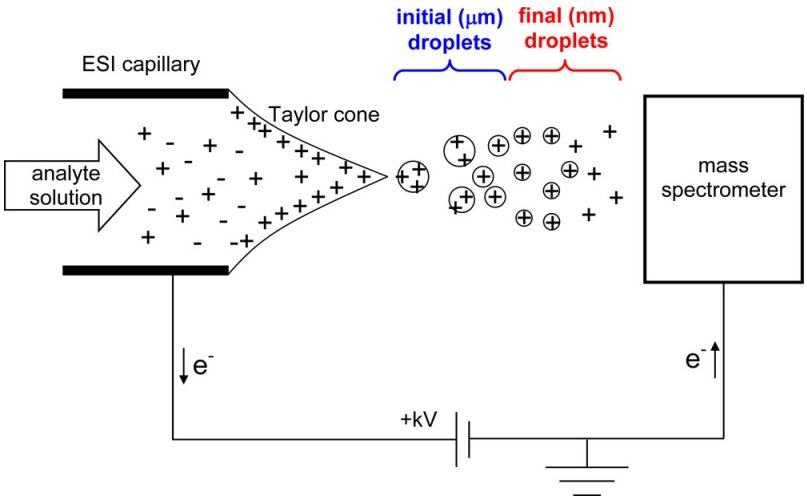

Figure 1.6 ESI process in the positive ion mode. Reproduced with permission from reference (95).

In decades, tremendous efforts have been invested to improve the sensitivity of the sheath-flow interface. The coaxial sheath-flow interface, designed by the Smith group in 1988, is the earliest version of the interface used for CE-MS.[98] The coaxial configuration allows the introduction of sheath liquid around the terminal end of the capillary to promote desolvation and ionization (Figure 1.7A). However, significant dilution of analytes can occur using this interface, as sheath liquids are typically pumped at a high flow rate (~1~10 μL/min), which is much higher than the flow rate in CE (20~100 nL/min). Improvement in the sensitivity has been achieved by reducing the flow rate of sheath liquid to nL/min level in later CE-MS interface development. In 2010, the Chen group constructed a flow-through microvial interface by placing the separation capillary in a stainless-steel emitter (Figure 1.7B).[99,100] The sheath buffer is delivered through the gap between the capillary and the emitter via a syringe pump at a very low flow rate (100~300 nL/min). By applying this interface with a flow rate of 200 nL/min, they achieved a five times higher limit of detections (LODs) for amino acids compared to using a commercialized sheath flow interface (flow rate~1 μL/min).[100] In the same year, the Dovichi group introduced an electrokinetically pumped sheath flow interface to the field.[101,102]The interface is designed with the capillary inserted through a cross-unit into a glass emitter (Figure 1.7C). The side arm of the cross-unit is connected to a vial containing acidic sheath liquid. When the voltage is applied to the sheath buffer, it drives electro-osmotic flow in the emitter to produce an electrospray. The interface was later upgraded by adjusting of emitter orifice size and distance between the terminus of the capillary and emitter orifice. The third generation of the interface reported in 2015 achieved as low as ~50 nL/min flow rate and ~300 zmole LODs for peptides.[102] Compared to the flow-

through microvial interface, the electrokinetically pumped sheath flow interface does not require hydrodynamic forces (such as pressure) to aid flow, thus avoiding post-column band broadening. Both the two sheath-flow interfaces have been successfully applied to CZE-MS and cIEF-MS analysis. In particular, the electrokinetically pumped sheath flow interface, has presented outstanding robustness and sensitivity in CZE-MS-based TDP and BUP. In addition, sheath-flow interfaces are well-suited for cIEF-MS studies. Apart from assisting ionization, the sheath liquids can serve as chemical mobilizers to facilitate protein mobilization after cIEF focusing and reduce the impact of ampholyte on sensitivity by mixing cIEF effluent with sheath liquids.

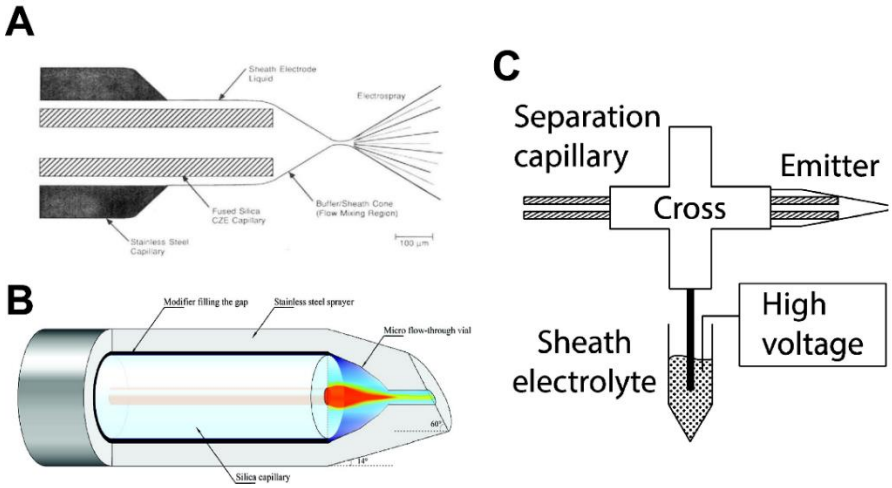

Figure 1.7 CE-MS interface. (A) Coaxial sheath flow interface. (B) Flow-through microvial interface. (C) Electrokinetically pumped sheath flow interface. Reproduced with permission from reference (98), (100), and (102).

1.5 Advancement of CE-MS and their applications in TDP

CZE-MS and cIEF-MS have been developed for protein analysis since the 1990s. In recent years, CZE-MS have been improved in sensitivity and capacity through the advancement of the CE-MS interface and the development of sample preconcentration methods. Using an electrokinetically pumped sheath flow interface and dynamic pH junction, a single-shot CZE-MS/MS of *E.coli* proteome was able to achieve around 600 proteoform IDs and 200 protein IDs.[21] The performance of CZE-MS/MS was enhanced by the application of advanced fragmentation techniques. CZE-MS/MS with UVPD (213 nm) for TDP of zebrafish brain identified more than 227 proteoforms from 139 proteins with high sequence coverage.[103] CZE-MS/MS system equipped with AI-ETD presented around 1000 proteoform IDs of 300 proteins from one SEC fraction of *E. coli*.[104] The capacity of CZE-MS/MS was also expanded using a long separation capillary. A 1.5-meter capillary for CZE-MS/MS significantly increased the sample loading volume from hundreds

of nL (1-meter capillary) to 2 μL in large-scale TDP of zebrafish brains.[51] In addition, constructing multidimensional platforms by integrating various liquid-phase phase separation/sample prefractionation methods with CZE-MS/MS further boosts peak capacity, which is crucial for deep TDP. Coupling SEC and RPLC sample prefractionation with CZE-MS/MS previously generated high peak capacity (~4000) and attributed around 6000 proteoform IDs from *E.coli*, presenting a 10-fold improvement compared with many one-dimensional CZE-MS/MS studies.[52]

CZE-MS/MS and CZE-MS/MS-based multidimensional platforms have been applied for a variety of biological applications. CZE-MS/MS for quantitative analysis of zebrafish brain cerebellum (Cb) and optic tectum (Teo) and discovered 700 differentially expressed proteoforms between the two regions, which provides insight into different functions of different brain regions.[51] Furthermore, the platform was used for the analysis of brain sections isolated from optic tectum (Teo) and telencephalon (Tel) regions using laser capture microdissection (LCM), enabling spatial resolving of proteoform distribution of brains.[105] Another study was carried out for comprehensive mapping of the proteoform landscape in five human tissues (heart, lungs, kidneys, small intestines, and spleen) by combining CZE-MS/MS and nano-flow RPLC-MS/MS analysis.[50] The work identified in total, 11,466 proteoforms and found various important proteoforms associated with tissue-specific functions (e.g. muscle contractility, host-pathogen interaction, etc.). Chen et al. applied two-dimensional SEC-CZE-MS/MS to globally investigate histone proteoforms from a calf histone sample. CZE-MS/MS showed 30-fold higher sensitivity than RPLC-MS/MS and achieved 400 histone proteoform IDs with diverse PTMs, such as acetylation, methylation, phosphorylation, and succinylation.[106] Johnson et al. initialized a pilot study using an on-capillary cell lysis workflow for CZE-MS/MS analysis of single cells, resulting in 23-50 proteoform IDs from replicate runs of single HeLa cells.[107] The low protein loss sample processing combined with the high sensitivity of CZE-MS/MS will no doubt expedite a better understanding of the single-cell heterogeneity at the proteoform level in the future. CZE-MS/MS was also performed at native conditions and has been used for the characterization of the standard protein mixture, protein complexes in *E. coli* proteome, and endogenous nucleosomes (200 kDa complex of DNA and histone proteins).[108-110]

In addition, the CZE-MS/MS has emerged as a powerful tool in pharmaceutical analysis. Native CZE-MS/MS was employed for characterizing the proteoforms in SigmaMAb and monomer and homodimer in NISTmAb.[111] CZE-MS on the microfluidic device (ZipChip system of 908 devices) facilitates fast screening (less than 15 minutes) of fragments and PTM variants in monoclonal antibodies (mAbs) and bispecific antibodies (bsAbs), and drug-to-antibody ratio (DAR) species in a lysine-linked antibody-drug conjugate (ADC).[112-115]

Compared to CZE-MS/MS, cIEF-MS/MS remains less popular although cIEF is the CE mode with the highest resolving power. cIEF-MS/MS has long been challenging because of manual operations, lack of sensitive CE-MS interface, and ampholyte impact on MS detection.[27] In recent years, the technique has received breakthrough improvements in those aspects. The introduction of the "sandwich" injection configuration to cIEF enabled fully automated separation and mobilization. The upgraded sheath flow CE-MS interface also enhanced sensitivity for MS analysis and minimized the influence of ampholytes. cIEF-MS has been applied for the characterization of various mAbs and presented superior separation resolution on their charge variants (pI variation less than 0.1) with heterogeneity on deamidation, incomplete lysine clipping, and cyclization of N-terminal glutamic acid, etc.[116,117] In addition, Lecoeur et al. performed qualitative and quantitative analysis of whey proteins in bovine milk using cIEF-MS in glycerol-water media.[118] Typically, using glycerol as an additive in cIEF greatly benefited the solubility of hydrophobic whey proteins during separation. Furthermore, despite that early cIEF-MS studies attempted to utilize cIEF-MS for the mass analysis of proteins in complex samples (e.g. *E. coli*), they generally lack identification of separated proteins, which was limited by efficient fragmentation and bioinformatics tools.[56-62] Therefore, the combination of state-of-the-art top-down (TD)-MS technologies with ultra-high resolution cIEF will be appealing for large-scale TDP of complex proteome and targeted TD characterization of proteoform heterogeneity.

1.6 Summary

TDP is a crucial strategy for interrogating complex proteomes at the proteoform level. MS-based TDP needs to be improved in throughput and depth for efficient analysis and better proteome coverage. The advancement of separation in TD-MS analysis is critical for achieving the expectation above. CZE and cIEF are considered highly attractive for TDP for their high separation efficiency and low sample loss compared to LC. CZE-MS/MS (or MS)-based platforms have been frequently employed for TDP and pharmaceutical applications in the past five years. Typically, coupling LC prefractionation with CZE-MS/MS largely boosted the peak capacity of the platform, which is highly valuable for the deep TDP of biological subjects. In addition, the current CZE-based multidimensional approach requires large sample materials and is time and labor-consuming. More efficient sample fractionation to couple with CZE-MS/MS is highly desired for further enhancing the performance and making it feasible to use. Besides, cIEF-MS/MS remains a less developed but highly promising technique for TDP. cIEF has a higher sample capacity and separation resolution than CZE. However, cIEF-MS/MS still needs to be further improved in the configuration for automated separation and the sensitivity and stability of the platform.

The research works in this dissertation are dedicated to advancing cIEF-MS/MS and CZE-MS/MS (or MS) based platforms for TDP. Two biological applications using cIEF-MS/MS and CZE-MS/MS were carried out for studying the sexual dimorphism of zebrafish brains and metastatic and non-metastatic colorectal cancer cell lines, respectively.

REFERENCES

(1) Smith, L. M.; Kelleher, N. L.Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10*, 186-187.

(2) Smith, L. M.; Kelleher, N. L.Proteoforms as the next proteomics currency. *Science* **2018**, *359*, 1106-1107.

(3) Smith, L. M.; Agar, J. N.; Chamot-Rooke, J.; Danis, P. O.; Ge, Y.; Loo, J. A.; Paša-Tolić, L.; Tsybin, Y. O.; Kelleher, N. L.; Proteomics, C. f. T.-D.The human proteoform project: defining the human proteome. *Sci. Adv.* **2021**, *7*, eabk0734.

(4) 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.Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011**, *480*, 254-258.

(5) 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.How many human proteoforms are there? *Nat. Chem. Biol* **2018**, *14*, 206-214.

(6) Millán-Zambrano, G.; Burton, A.; Bannister, A. J.; Schneider, R.Histone post-translational modifications—cause and consequence of genome function. *Nat. Rev. Genet.* **2022**, 1-18.

(7) Schmid, A. W.; Fauvet, B.; Moniatte, M.; Lashuel, H. A.Alpha-synuclein post-translational modifications as potential biomarkers for Parkinson disease and other synucleinopathies. *Mol. Cell Proteomics* **2013**, *12*, 3543-3558.

(8) Melby, J. A.; Roberts, D. S.; Larson, E. J.; Brown, K. A.; Bayne, E. F.; Jin, S.; Ge, Y.Novel strategies to address the challenges in top-down proteomics. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 1278-1294.

(9) He, L.; Rockwood, A. L.; Agarwal, A. M.; Anderson, L. C.; Weisbrod, C. R.; Hendrickson, C. L.; Marshall, A. G.Top-down proteomics—a near-future technique for clinical diagnosis? *Ann. Transl. Med.* **2020**, *8*.

(10) Kelleher, N. L.; Thomas, P. M.; Ntai, I.; Compton, P. D.; LeDuc, R. D.Deep and quantitative top-down proteomics in clinical and translational research. *Expert Rev Proteomics* **2014**, *11*, 649-651.

(11) Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D.Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Mol. Cell Proteomics* **2016**, *15*, 45-56.

(12) Chait, B. T.Mass spectrometry: bottom-up or top-down? *Science* **2006**, *314*, 65-66.

(13) Zhou, H.; Ning, Z.; E. Starr, A.; Abu-Farha, M.; Figeys, D.Advancements in top-down proteomics. *Anal. Chem.* **2012**, *84*, 720-734.

(14) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y.Top-down proteomics: ready for prime time? *Anal. Chem.* **2017**, *90*, 110-127.

(15) 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.Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat. Methods* **2019**, *16*, 587-594.

(16) 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.Identification and quantification of proteoforms by mass spectrometry. *Proteomics* **2019**, *19*, 1800361.

(17) Melani, R. D.; Gerbasi, V. R.; Anderson, L. C.; Sikora, J. W.; Toby, T. K.; Hutton, J. E.; Butcher, D. S.; Negrão, F.; Seckler, H. S.; Srzentić, K.The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* **2022**, *375*, 411-418.

(18) 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*, 3465-3473.

(19) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L.Top down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445*, 683-693.

(20) Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L.Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *Trends Anal Chem* **2019**, *120*, 115644.

(21) 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*, 12059-12067.

(22) Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L.Recent advances (2019–2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. *Mass Spectrom. Rev.* **2021**.

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

(24) Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates III, 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.

(25) Simpson, D. C.; Smith, R. D.Combining capillary electrophoresis with mass spectrometry for applications in proteomics. *Electrophoresis* **2005**, *26*, 1291-1305.

(26) Hühner, J.; Lämmerhofer, M.; Neusüß, C.Capillary isoelectric focusing-mass spectrometry: Coupling strategies and applications. *Electrophoresis* **2015**, *36*, 2670-2686.

(27) Xu, T.; Sun, L.A mini review on capillary isoelectric focusing-mass spectrometry for top-down proteomics. *Front. Chem.* **2021**, *9*, 651757.

(28) Gerbasi, V. R.; Melani, R. D.; Abbatiello, S. E.; Belford, M. W.; Huguet, R.; McGee, J. P.; Dayhoff, D.; Thomas, P. M.; Kelleher, N. L.Deeper Protein Identification Using Field Asymmetric Ion Mobility Spectrometry in Top-Down Proteomics. *Anal. Chem.* **2021**, *93*, 6323-6328.

(29) Fulcher, J. M.; Makaju, A.; Moore, R. J.; Zhou, M.; Bennett, D. A.; De Jager, P. L.; Qian, W.-J.; Paša-Tolić, L.; Petyuk, V. A.Enhancing top-down proteomics of brain tissue with FAIMS. *J. Proteome Res.* **2021**, *20*, 2780-2795.

(30) Kaulich, P. T.; Cassidy, L.; Winkels, K.; Tholey, A.Improved Identification of Proteoforms in Top-Down Proteomics Using FAIMS with Internal CV Stepping. *Anal. Chem.* **2022**, *94*, 3600-3607.

(31) Takemori, A.; Kaulich, P. T.; Cassidy, L.; Takemori, N.; Tholey, A.Size-Based Proteome Fractionation through Polyacrylamide Gel Electrophoresis Combined with LC–FAIMS–MS for In-Depth Top-Down Proteomics. *Anal. Chem.* **2022**, *94*, 12815-12821.

(32) Moody, H. W.The evaluation of the parameters in the van Deemter equation. *J. Chem. Educ.* **1982**, *59*, 290.

(33) Chen, Z.; Wu, J.; Baker, G. B.; Parent, M.; Dovichi, N. J.Application of capillary electrophoresis with laser-induced fluorescence detection to the determination of biogenic amines and amino acids in brain microdialysate and homogenate samples. *J. Chromatogr. A* **2001**, *914*, 293-298.

(34) Wang, Z.; Ma, H.; Smith, K.; Wu, S.Two-dimensional separation using high-pH and low-pH reversed phase liquid chromatography for top-down proteomics. *Int. J. Mass spectrom.* **2018**, *427*, 43-51.

(35) Wang, Z.; Yu, D.; Cupp-Sutton, K. A.; Liu, X.; Smith, K.; Wu, S.Development of an online 2D ultrahigh-pressure nano-LC system for high-pH and low-pH reversed phase separation in top-down proteomics. *Anal. Chem.* **2020**, *92*, 12774-12777.

(36) Xiu, L.; Valeja, S. G.; Alpert, A. J.; Jin, S.; Ge, Y.Effective protein separation by coupling hydrophobic interaction and reverse phase chromatography for top-down proteomics. *Anal. Chem.* **2014**, *86*, 7899-7906.

(37) Chen, B.; Lin, Z.; Alpert, A. J.; Fu, C.; Zhang, Q.; Pritts, W. A.; Ge, Y.Online hydrophobic interaction chromatography–mass spectrometry for the analysis of intact monoclonal antibodies. *Anal. Chem.* **2018**, *90*, 7135-7138.

(38) Chen, X.; Ge, Y.Ultrahigh pressure fast size exclusion chromatography for top-down proteomics. *Proteomics* **2013**, *13*, 2563-2566.

(39) Tucholski, T.; Knott, S. J.; Chen, B.; Pistono, P.; Lin, Z.; Ge, Y.A top-down proteomics platform coupling serial size exclusion chromatography and Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **2019**, *91*, 3835-3844.

(40) 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*, 5467-5475.

(41) Muneeruddin, K.; Thomas, J. J.; Salinas, P. A.; Kaltashov, I. A.Characterization of small protein aggregates and oligomers using size exclusion chromatography with online detection by native electrospray ionization mass spectrometry. *Anal. Chem.* **2014**, *86*, 10692-10699.

(42) Mohr, J.; Swart, R.; Samonig, M.; Böhm, G.; Huber, C. G.High-efficiency nano-and micro-HPLC–High-resolution Orbitrap-MS platform for top-down proteomics. *Proteomics* **2010**, *10*, 3598-3609.

(43) 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.

(44) Capriotti, A. L.; Cavaliere, C.; Foglia, P.; Samperi, R.; Laganà, A.Intact protein separation by chromatographic and/or electrophoretic techniques for top-down proteomics. *J. Chromatogr. A* **2011**, *1218*, 8760-8776.

(45) Uliyanchenko, E.Size-exclusion chromatography—from high-performance to ultra-performance. *Anal. Bioanal. Chem.* **2014**, *406*, 6087-6094.

(46) Liu, H.; Gaza-Bulseco, G.; Chumsae, C.Analysis of reduced monoclonal antibodies using size exclusion chromatography coupled with mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2258-2264.

(47) Jouyban-Gharamaleki, A.; Khaledi, M. G.; Clark, B. J.Calculation of electrophoretic mobilities in water–organic modifier mixtures in capillary electrophoresis. *J. Chromatogr. A* **2000**, *868*, 277-284.

(48) VanOrman, B. B.; Liversidge, G. G.; McIntire, G. L.; Olefirowicz, T. M.; Ewing, A. G.Effects of buffer composition on electroosmotic flow in capillary electrophoresis. *J. Microcolumn Sep.* **1990**, *2*, 176-180.

(49) Harvanová, J.; Bloom, L.Capillary electrophoresis technique for metal species determination: A review. *J. Liq. Chromatogr. Rel. Technol.* **2015**, *38*, 371-380.

(50) Drown, B. S.; Jooß, K.; Melani, R. D.; Lloyd-Jones, C.; Camarillo, J. M.; Kelleher, N. L.Mapping the Proteoform Landscape of Five Human Tissues. *J. Proteome Res.* **2022**, *21*, 1299-1310.

(51) 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*, 1435-1445.

(52) 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*, 5529-5533.

(53) 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*, 250-255.

(54) Wang, L.; MacDonald, D.; Huang, X.; Chen, D. D.Capture efficiency of dynamic pH junction focusing in capillary electrophoresis. *Electrophoresis* **2016**, *37*, 1143-1150.

(55) Zhu, G.; Sun, L.; Dovichi, N. J.Dynamic pH junction preconcentration in capillary electrophoresis-electrospray ionization-mass spectrometry for proteomics analysis. *Analyst* **2016**, *141*, 5216-5220.

(56) Jensen, P. K.; Paša-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*, 2076-2084.

(57) Shen, Y.; Xiang, F.; Veenstra, T. D.; Fung, E. N.; Smith, R. D.High-resolution capillary isoelectric focusing of complex protein mixtures from lysates of microorganisms. *Anal. Chem.* **1999**, *71*, 5348-5353.

(58) Pasa-Tolic, L.; Jensen, P. K.; Anderson, G. A.; Lipton, M. S.; Peden, K. K.; Martinovic, S.; Tolic, N.; Bruce, J. E.; Smith, R. D.High throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.* **1999**, *121*.

(59) Martinović, S.; Berger, S. J.; Paša-Tolić, L.; Smith, R. D.Separation and detection of intact noncovalent protein complexes from mixtures by on-line capillary isoelectric focusing-mass spectrometry. *Anal. Chem.* **2000**, *72*, 5356-5360.

(60) 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*, 3235-3241.

(61) Yang, L.; Lee, C. S.; Hofstadler, S. A.; Smith, R. D.Characterization of microdialysis acidification for capillary isoelectric focusing-microelectrospray ionization mass spectrometry. *Anal. Chem.* **1998**, *70*, 4945-4950.

(62) Tang, Q.; Harrata, A. K.; Lee, C. S.Two-dimensional analysis of recombinant E. coli proteins using capillary isoelectric focusing electrospray ionization mass spectrometry. *Anal. Chem.* **1997**, *69*, 3177-3182.

(63) Mokaddem, M.; Gareil, P.; Varenne, A.Online CIEF-ESI-MS in glycerol–water media with a view to hydrophobic protein applications. *Electrophoresis* **2009**, *30*, 4040-4048.

(64) Zhu, G.; Sun, L.; Dovichi, N. J.Simplified capillary isoelectric focusing with chemical mobilization for intact protein analysis. *J. Sep. Sci.* **2017**, *40*, 948-953.

(65) Montealegre, C.; Neusüß, C.Coupling imaged capillary isoelectric focusing with mass spectrometry using a nanoliter valve. *Electrophoresis* **2018**, *39*, 1151-1154.

(66) Hühner, J.; Jooß, K.; Neusüß, C.Interference-free mass spectrometric detection of capillary isoelectric focused proteins, including charge variants of a model monoclonal antibody. *Electrophoresis* **2017**, *38*, 914-921.

(67) Hühner, J.; Neusüß, C.CIEF-CZE-MS applying a mechanical valve. *Anal. Bioanal. Chem.* **2016**, *408*, 4055-4061.

(68) Cooper, H. J.To what extent is FAIMS beneficial in the analysis of proteins? *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 566-577.

(69) Shvartsburg, A. A.; Tang, K.; Smith, R. D.Modeling the resolution and sensitivity of FAIMS analyses. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 1487-1498.

(70) Swearingen, K. E.; Hoopmann, M. R.; Johnson, R. S.; Saleem, R. A.; Aitchison, J. D.; Moritz, R. L.Nanospray FAIMS fractionation provides significant increases in proteome coverage of unfractionated complex protein digests. *Mol. Cell Proteomics* **2012**, *11*.

(71) Shvartsburg, A. A.; Bryskiewicz, T.; Purves, R. W.; Tang, K.; Guevremont, R.; Smith, R. D.Field asymmetric waveform ion mobility spectrometry studies of proteins: Dipole alignment in ion mobility spectrometry? *J. Phys. Chem. B* **2006**, *110*, 21966-21980.

(72) Zhang, H.; Cui, W.; Wen, J.; Blankenship, R. E.; Gross, M. L.Native electrospray and electron-capture dissociation FTICR mass spectrometry for top-down studies of protein assemblies. *Anal. Chem.* **2011**, *83*, 5598-5606.

(73) Garcia, B. A.What does the future hold for top down mass spectrometry? *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 193-202.

(74) Siuti, N.; Kelleher, N. L.Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **2007**, *4*, 817-821.

(75) Mann, M.; Kelleher, N. L.Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18132-18138.

(76) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* **1998**, *17*, 1-35.

(77) Zubarev, R. A.; Makarov, A.; ACS Publications, 2013.

(78) Eliuk, S.; Makarov, A.Evolution of orbitrap mass spectrometry instrumentation. *Annu. Rev. Anal. Chem* **2015**, *8*, 61-80.

(79) Yin, R.; Kyle, J.; Burnum-Johnson, K.; Bloodsworth, K. J.; Sussel, L.; Ansong, C.; Laskin, J.High spatial resolution imaging of mouse pancreatic islets using nanospray desorption electrospray ionization mass spectrometry. *Anal. Chem.* **2018**, *90*, 6548-6555.

(80) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. An introduction to quadrupole–time-of-flight mass spectrometry. *J. Mass Spectrom.* **2001**, *36*, 849-865.

(81) Tamara, S.; den Boer, M. A.; Heck, A. J.High-resolution native mass spectrometry. *Chem. Rev.* **2021**, *122*, 7269-7326.

(82) Lössl, P.; Snijder, J.; Heck, A. J.Boundaries of mass resolution in native mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 906-917.

(83) Wiesner, J.; Premsler, T.; Sickmann, A.Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics* **2008**, *8*, 4466-4483.

(84) Medzihradszky, K.; Zhang, X.; Chalkley, R.; Guan, S.; McFarland, M.; Chalmers, M.; Marshall, A.; Diaz, R. L.; Allis, C. D.; Burlingame, A.Characterization of Tetrahymena histone H2B variants and posttranslational populations by electron capture dissociation (ECD) Fourier transform ion cyclotron mass spectrometry (FT-ICR MS). *Mol. Cell Proteomics* **2004**, *3*, 872-886.

(85) Brunner, A. M.; Lössl, P.; Liu, F.; Huguet, R.; Mullen, C.; Yamashita, M.; Zabrouskov, V.; Makarov, A.; Altelaar, A. M.; Heck, A. J.Benchmarking multiple fragmentation methods on an orbitrap fusion for top-down phospho-proteoform characterization. *Anal. Chem.* **2015**, *87*, 4152-4158.

(86) Mirgorodskaya, E.; Roepstorff, P.; Zubarev, R.Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal. Chem.* **1999**, *71*, 4431-4436.

(87) Shen, X.; Xu, T.; Hakkila, B.; Hare, M.; Wang, Q.; Wang, Q.; Beckman, J. S.; Sun, L.Capillary zone electrophoresis-electron-capture collision-induced dissociation on a quadrupole time-of-flight mass spectrometer for top-down characterization of intact proteins. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 1361-1369.

(88) Rush, M. J.; 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*, 8946-8953.

(89) Riley, N. M.; Westphall, M. S.; Coon, J. J.Sequencing larger intact proteins (30-70 kDa) with activated ion electron transfer dissociation. *J. Am. Soc. Mass Spectrom.* **2017**, *29*, 140-149.

(90) Lodge, J. M.; Schauer, K. L.; Brademan, D. R.; Riley, N. M.; Shishkova, E.; Westphall, M. S.; Coon, J. J.Top-down characterization of an intact monoclonal antibody using activated ion electron transfer dissociation. *Anal. Chem.* **2020**, *92*, 10246-10251.

(91) Fornelli, L.; Srzentić, K.; Huguet, R.; Mullen, C.; Sharma, S.; Zabrouskov, V.; Fellers, R. T.; Durbin, K. R.; Compton, P. D.; Kelleher, N. L.Accurate sequence analysis of a monoclonal antibody by top-down and middle-down orbitrap mass spectrometry applying multiple ion activation techniques. *Anal. Chem.* **2018**, *90*, 8421-8429.

(92) Lanzillotti, M.; Brodbelt, J. S.Comparison of Top-Down Protein Fragmentation Induced by 213 and 193 nm UVPD. *J. Am. Soc. Mass Spectrom.* **2023**.

(93) Cleland, T. P.; DeHart, C. J.; Fellers, R. T.; VanNispen, A. J.; Greer, J. B.; LeDuc, R. D.; Parker, W. R.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S.High-throughput analysis of intact human proteins using UVPD and HCD on an orbitrap mass spectrometer. *J. Proteome Res.* **2017**, *16*, 2072-2079.

(94) Bruins, A. P.Mechanistic aspects of electrospray ionization. *J. Chromatogr. A* **1998**, *794*, 345-357.

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

(96) Maxwell, E. J.; Chen, D. D.Twenty years of interface development for capillary electrophoresis–electrospray ionization–mass spectrometry. *Anal. Chim. Acta* **2008**, *627*, 25-33.

(97) Ramautar, R.; Heemskerk, A. A.; Hensbergen, P. J.; Deelder, A. M.; Busnel, J.-M.; Mayboroda, O. A.CE–MS for proteomics: Advances in interface development and application. *J Proteomics* **2012**, *75*, 3814-3828.

(98) Smith, R. D.; Barinaga, C. J.; Udseth, H. R.Improved electrospray ionization interface for capillary zone electrophoresis-mass spectrometry. *Anal. Chem.* **1988**, *60*, 1948-1952.

(99) 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.

(100) Zhong, X.; Maxwell, E.J.; Chen, D.D. Mass transport in a micro flow-through vial of a junction-at-the-tip capillary electrophoresis-mass spectrometry interface. *Anal. Chem* **2010**, 83(12), 4916-4923.

(101) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J.Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554-2560.

(102) 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*, 2312-2321.

(103) McCool, E. N.; Chen, D.; Li, W.; Liu, Y.; Sun, L.Capillary zone electrophoresis-tandem mass spectrometry with ultraviolet photodissociation (213 nm) for large-scale top–down proteomics. *Anal Methods* **2019**, *11*, 2855-2861.

(104) McCool, E. N.; Lodge, J. M.; Basharat, A. R.; Liu, X.; Coon, J. J.; Sun, L.Capillary zone electrophoresis-tandem mass spectrometry with activated ion electron transfer dissociation for large-scale top-down proteomics. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 2470-2479.

(105) Lubeckyj, R. A.; Sun, L.Laser capture microdissection-capillary zone electrophoresis-tandem mass spectrometry (LCM-CZE-MS/MS) for spatially resolved top-down proteomics: a pilot study of zebrafish brain. *Mol. Omics* **2022**, *18*, 112-122.

(106) Chen, D.; Yang, Z.; Shen, X.; Sun, L.Capillary Zone Electrophoresis-Tandem Mass Spectrometry As an Alternative to Liquid Chromatography-Tandem Mass Spectrometry for Top-down Proteomics of Histones. *Anal. Chem.* **2021**, *93*, 4417-4424.

(107) Johnson, K. R.; Gao, Y.; Gregus, M.; Ivanov, A. R.On-capillary Cell Lysis Enables Top-down Proteomic Analysis of Single Mammalian Cells by CE-MS/MS. *Anal. Chem.* **2022**, *94*, 14358-14367.

(108) Jooß, K.; McGee, J. P.; Melani, R. D.; Kelleher, N. L.Standard procedures for native CZE-MS of proteins and protein complexes up to 800 kDa. *Electrophoresis* **2021**, *42*, 1050-1059.

(109) Shen, X.; Kou, Q.; Guo, R.; Yang, Z.; Chen, D.; Liu, X.; Hong, H.; Sun, L.Native proteomics in discovery mode using size-exclusion chromatography–capillary zone electrophoresis–tandem mass spectrometry. *Anal. Chem.* **2018**, *90*, 10095-10099.

(110) Jooß, K.; Schachner, L. F.; Watson, R.; Gillespie, Z. B.; Howard, S. A.; Cheek, M. A.; Meiners, M. J.; Sobh, A.; Licht, J. D.; Keogh, M.-C.Separation and characterization of endogenous nucleosomes by native capillary zone electrophoresis–top-down mass spectrometry. *Anal. Chem.* **2021**, *93*, 5151-5160.

(111) Shen, X.; Liang, Z.; Xu, T.; Yang, Z.; Wang, Q.; Chen, D.; Pham, L.; Du, W.; Sun, L.Investigating native capillary zone electrophoresis-mass spectrometry on a high-end quadrupole-time-of-flight mass spectrometer for the characterization of monoclonal antibodies. *Int. J. Mass spectrom.* **2021**, *462*, 116541.

(112) Redman, E. A.; Batz, N. G.; Mellors, J. S.; Ramsey, J. M.Integrated microfluidic capillary electrophoresis-electrospray ionization devices with online MS detection for the separation and characterization of intact monoclonal antibody variants. *Anal. Chem.* **2015**, *87*, 2264-2272.

(113) Redman, E. A.; Mellors, J. S.; Starkey, J. A.; Ramsey, J. M.Characterization of intact antibody drug conjugate variants using microfluidic capillary electrophoresis–mass spectrometry. *Anal. Chem.* **2016**, *88*, 2220-2226.

(114) Carillo, S.; Jakes, C.; Bones, J.In-depth analysis of monoclonal antibodies using microfluidic capillary electrophoresis and native mass spectrometry. *J. Pharm. Biomed. Anal.* **2020**, *185*, 113218.

(115) Wu, Z.; Wang, H.; Wu, J.; Huang, Y.; Zhao, X.; Nguyen, J. B.; Rosconi, M. P.; Pyles, E. A.; Qiu, H.; Li, N.High-sensitivity and high-resolution therapeutic antibody charge variant and impurity characterization by microfluidic native capillary electrophoresis-mass spectrometry. *J. Pharm. Biomed. Anal.* **2023**, *223*, 115147.

(116) Dai, J.; Lamp, J.; Xia, Q.; Zhang, Y.Capillary isoelectric focusing-mass spectrometry method for the separation and online characterization of intact monoclonal antibody charge variants. *Anal. Chem.* **2018**, *90*, 2246-2254.

(117) Wang, L.; Bo, T.; Zhang, Z.; Wang, G.; Tong, W.; Da Yong Chen, D.High resolution capillary isoelectric focusing mass spectrometry analysis of peptides, proteins, and monoclonal antibodies with a flow-through microvial interface. *Anal. Chem.* **2018**, *90*, 9495-9503.

(118) Lecoeur, M.; Gareil, P.; Varenne, A.Separation and quantitation of milk whey proteins of close isoelectric points by on-line capillary isoelectric focusing—Electrospray ionization mass spectrometry in glycerol–water media. *J. Chromatogr. A* **2010**, *1217*, 7293-7301.

CHAPTER 2.* Development of automated cIEF-MS/MS and multidimensional SEC-cIEF-MS/MS approaches for TDP

2.1 Introduction

Mass spectrometry (MS)-based top-down proteomics (TDP) has emerged as a powerful tool for accurate identification and quantification of proteoforms, and provides comprehensive information about genetic variations, alternative splicing, post-translational modifications (PTMs).[1,2] Accurate characterization of proteoforms is critical for better understanding protein functions and discovering important protein signatures in the development of diseases.[3-5] Due to the extremely high complexity of proteomes, high-resolution proteoform separation is vital for large-scale TDP.

Besides the routinely used reversed-phase liquid chromatography (RPLC)-MS/MS, capillary zone electrophoresis (CZE)-MS/MS has been suggested as a valuable tool for TDP of complex proteomes with the identification of thousands of proteoforms [6-13]. Capillary isoelectric focusing (cIEF) is another classic electrophoresis technique, which separates proteoforms according to their isoelectric points (pIs).[14] Integrating cIEF with ESI-MS for protein study has been an important research area for two decades because cIEF has ultra-high resolution for proteoform separation.[15] The Lee and Smith groups performed the pioneering cIEF-MS works in the 1990s for the characterization of simple protein mixtures and complex proteome via the co-axial sheath flow CE-MS interface.[16-20] These pioneering works laid the foundation of using cIEF-MS for protein characterization. However, the technique has not been widely adopted for protein characterization in last two decades due to its manual operations, the ionization suppression of analytes from ampholytes, and lack of robust and highly sensitive CE-MS interface.

In recent years, cIEF-MS has attracted great attention again because of the drastic improvement of the CE-MS interface in sensitivity and the automated operations of cIEF-MS. The flow-through micro-vial CE-MS interface[21] and the electro-kinetically pumped sheath flow CE-MS interface [22,23] have been employed for cIEF-MS studies, in which “sandwich” injection methods were developed for automated cIEF-MS [24-26]. Several studies have successfully employed automated cIEF-MS for the high-resolution characterization of antibody charge variants[27-30].

---

* This chapter is partially adapted with permission from *Xu, T.; Shen, X.; Yang, Z.; Chen, D.; Lubeckyj, R. A.; McCool, E. N.; Sun, L. Anal. Chem. 2020, 92 (24), 15890-15898.*

While cIEF-MS presented great potential for delineating proteoforms, the previous cIEF-MS studies have been concentrated on measuring protein's mass without MS/MS analysis, impeding the confident proteoform identification in complex samples as well as the accurate localization of PTMs on proteoforms. In this study, we report the first work of applying automated cIEF-MS/MS in large-scale TDP of complex proteomes. The automated and on-line cIEF-MS/MS platform was developed using the electro-kinetically pumped sheath flow CE-MS interface, the "sandwich" injection configuration, and linear-polyacrylamide (LPA) coated separation capillaries. We developed high-throughput and high-capacity cIEF-MS/MS methods for large-scale TDP.

2.2 Experimental section

2.2.1 Sample preparation

A standard protein mixture (0.2 mg/mL) containing cytochrome c (11.7 kDa, pI 10.0, 0.1 mg/mL, Sigma-Aldrich) and myoglobin (16.9 kDa, pI 7, 0.1 mg/mL, Sigma-Aldrich) was prepared in 10 mM ammonium acetate solution (pH 6.9) for investigating cIEF separation under different conditions.

E. coli (strain K-12, substrain MG1655) was cultured in Lysogeny broth (LB) medium at 37 °C with 225 rpm shaking until the OD600 value reached to 0.7. The bacteria were collected by centrifugation (4,000 rpm, 10 min), then washed three times with phosphate-buffered saline (PBS). Afterward, the *E. coli* pellet was suspended in the lysis buffer containing 8 M urea, protease inhibitor (Roche), phosphatase inhibitor (Roche), and 100 mM ammonium bicarbonate (pH 8.0). The cells were lysed for 1 minute using a homogenizer 150 (Fisher Scientific) and then sonicated on ice for 10 minutes with a Branson Sonifier 250 (VWR Scientific). The *E. coli* lysate was centrifuged at 18,000 g for 10 minutes to collect the supernatant containing extracted proteins. The concentration of total proteins was measured by a bicinchoninic acid (BCA) kit (Fisher Scientific) according to manufacturer's instructions. 500 µg of *E. coli* proteins were denatured in lysis buffer at 37 °C for 30 minutes, reduced at 37 °C for 30 minutes after adding 1 µL of 1 M dithiothreitol (DTT, Sigma-Aldrich), and then alkylated at room temperature for 20 min after adding 2.5 µL of 1M iodoacetamide (IAA, Sigma-Aldrich). The reaction was quenched with an addition of 1 µL of 1 M DTT solution. The buffer exchange of protein sample was conducted by centrifugation with Microcon-30 kDa centrifugal filter (Merck Millipore) at 14,000 g for 10 minutes and then washing three times with 10 mM ammonium acetate (pH 6.9). Finally, the proteins retained on the centrifugal filter membrane were re-dissolved in 10 mM ammonium acetate (pH 6.9) for SDS-PAGE gel analysis and SEC fractionation.

2.2.2 SEC fractionation of *E. coli* proteome

SEC fractionation was performed on Agilent 1260 Infinity II HPLC system. An SEC column (Agilent, 4.6 × 300 mm, 3 μm particles, 150 Å porous) was used for protein separation. 200 μg (2 mg/mL, 50 μL × 2 injections) of *E. coli* proteins were injected into the SEC column and separated using 0.1% formic acid (FA) as mobile phase at a flow rate of 0.35 mL/min. The first fraction was collected from 2.0 minute to 5.0 minutes, the rest of the five fractions were collected from 5.0 to 15.0 minutes with 2 minutes per fraction. The samples were dried in the speed vacuum and re-dissolved in 10 μL of 10 mM ammonium acetate (pH 6.9).

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was conducted using 4~20% Mini-PROTEAN TGX precast gels (Bio-Rad) to evaluate the SEC fractionation efficiency. The fractions (10 μL, ~16 μg, protein per fraction) collected from of *E. coli* lysate were separately mixed with 10 μL loading buffer (10 mM Tris-HCl, pH 8.0,1% SDS, 40% glycerol, 0.1% DTT, and 0.05% bromophenol blue), and denatured at 95°C for 5 minutes. After these samples were loaded into an SDS-PAGE gel, the electrophoresis was performed with 1× SDS buffer at 140 V for 90 minutes. Finally, the gel was stained with Coomassie Blue for 2 hours and decolorized by deionized water for 12 hours. The size of proteins in each fraction was determined by comparing with standard proteins (Bio-Rad, Precision Plus Protein Dual Color Standards).

2.2.3 Preparation of linear polyacrylamide (LPA)-coated capillary

LPA-coated capillaries are commonly used in CZE since the neutral coating can effectively prevent proteins from adsorption onto the inner wall of the capillary and eliminate electroosmotic flow (EOF) in CE. In this study, LPA-coated capillaries were employed for cIEF separation. The bare fused silica capillaries (50 μm i.d., 360 μm o.d.) were coated with LPA according to the procedure previously described.[12-31] One end of the LPA-coated capillary was etched with hydrofluoric acid (HF) for about 90 minutes to reduce the outer diameter of the capillary end to around 70 μm according to our published procedure for cIEF-MS.[32]

2.2.4 Automated cIEF-MS/MS analysis

Automated cIEF-MS/MS analysis was performed via coupling a CESI 8000 Plus CE system (Beckman Coulter) to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) through a commercialized electrokinetically pumped sheath-flow CE-MS nano-spray interface (CMP Scientific Corp).[22-23] An ESI emitter (orifice size: 20~30 μm) was used to generate stable electrospray with assistance of sheath buffer containing 0.2%(v/v) formic acid and 10% (v/v) MeOH. One end of the LPA-coated capillary was inserted into the ESI emitter and the other end of the capillary was connected to the CE system. The “sandwich” injection configuration was employed to facilitate automated cIEF separation.[26,27] Generally, the capillary was sequentially

pumped with a plug of catholyte (0.3% v/v $NH_3 \cdot H_2O$, pH 11.8), protein-ampholyte (GE Healthcare Parmalyte 3~10 for IEF) mixture, and then inserted into a buffer vial containing 0.1% (v/v) formic acid (FA) or 5% (v/v) acetic acid (AA). Next, a voltage of 30 kV was applied across the capillary for protein focusing and mobilization.

A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for the cIEF-MS/MS analysis of complex proteomes using the data-dependent acquisition (DDA) mode. Full MS spectra were collected using the following parameters: m/z range of 800-3000, mass resolution of 240,000 (at m/z 200), a microscan number of 3, AGC target value of 1E6, and maximum injection time of 100 ms. The top 5 most intense precursor ions in full MS spectra were isolated with a window of 4 m/z and fragmented via higher-energy collisional dissociation (HCD) with normalized collision energy (NCE) of 20%. Only the precursor ions with an intensity higher than 5E4 and a charge state more than 3 were selected for fragmentation. Product ions were detected with a resolution of 60,000 (at m/z 200), a microscan number of 3, AGC target value of 1E6, and maximum injection time of 200 ms. The dynamic exclusion was enabled with a duration of 30 s and the isotopic peaks were excluded for DDA. A spray voltage of 2.0~2.3 kV was used, ion transfer tube temperature was set at 320°C, and the s-lens RF level was 55.

2.2.5 Database analysis

The MS raw files were converted to the mzXML files using Msconvert,[33] and further deconvoluted to the Msalign files using TopFD (TOP-down mass spectrometry feature detection), followed by database search using TopPIC (Top-down mass spectrometry-based proteoform identification and characterization).[34] UniProt databases of *E. coli* (UP000000625). Cysteine carbamidomethylation (+57) was set as a fixed modification and the maximum number of unexpected modifications was 2. The mass error of the precursor and product ions was within 15 ppm. The maximum and minimum mass shifts of unknown modifications were 500 Da and -500 Da, respectively. The false discovery rates (FDRs) were estimated using the target-decoy approach[35,36]. The data of six *E. coli* fractions were combined and filtered with a 5% proteoform-level FDR.

2.3 Results and discussions

2.3.1 Automated high-throughput and high-capacity cIEF-MS/MS

Figure 2.1A shows a diagram of the automated cIEF-MS system. In this platform, the outlet of an LPA-coated capillary is positioned into the electro-kinetically pumped sheath-flow CE-MS interface filled with an acidic sheath buffer containing 0.2% (v/v) formic acid (FA) and 10% (v/v) methanol, while its inlet is inserted into an acidic anolyte solution (0.1% (v/v) FA or 5% (v/v) acetic

acid (AA)). The focusing is carried out by applying a 30-kV voltage across the capillary after injecting a plug of basic catholyte (0.3% (v/v) $NH_3 \cdot H_2O$, pH 11.8) and a mixture of analytes and ampholyte into the capillary successively. After focusing, the separated proteoforms are mobilized out of the capillary for ESI-MS automatically when the pH gradient is gradually disrupted by the migration of hydrogen protons from the acidic anolyte and anions from the sheath buffer (chemical mobilization).

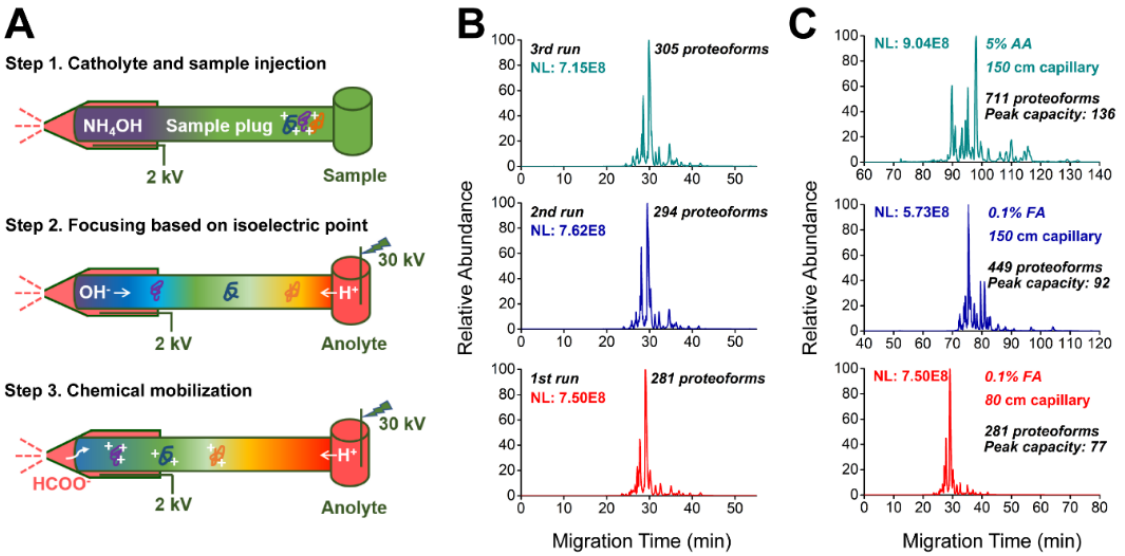

Figure 2.1 Development of cIEF-MS/MS methods with a single SEC fraction of an *E. coli lysate*. (A) Flowchart of automated cIEF-MS including basic catholyte and sample injection, focusing, and chemical mobilization. (B) Evaluation of reproducibility of cIEF-MS/MS system. The base peak electropherograms are from cIEF-MS/MS analysis of Fraction 3 of *E. coli* lysate in triplicate runs using an 80-cm capillary. (C) Base peak electropherograms of Fraction 3 using an 80-cm capillary plus 0.1% FA as anolyte (Red), a 150-cm capillary plus 0.1% FA as anolyte (Blue), and a 150-cm capillary plus 5% AA as anolyte (Dark cyan).

To improve proteoform separation and detection, critical experimental parameters of cIEF-MS were first investigated with a standard protein mixture, Figure 2.2. The results indicated that a 5-cm catholyte plug, a 40-cm sample plug (half of the total capillary volume), a 0.1% ampholyte concentration, and low protein concentration were the most appropriate conditions for cIEF separation balancing separation resolution and MS signal.

Using the optimized condition, one SEC fraction of an *E. coli* lysate (~0.4 mg/mL protein concentration) was analyzed by cIEF-MS/MS in triplicate. On average, nearly 300 proteoforms were identified in only 50 min with good reproducibility regarding the number of proteoform identifications (n=3 and RSD=4.1%), Figure 2.1B. We called the method high-throughput cIEF-

MS/MS. The high-throughput cIEF-MS/MS method also showed nice reproducibility regarding the label-free quantification (LFQ) intensity of proteoforms, Figure 2.3.

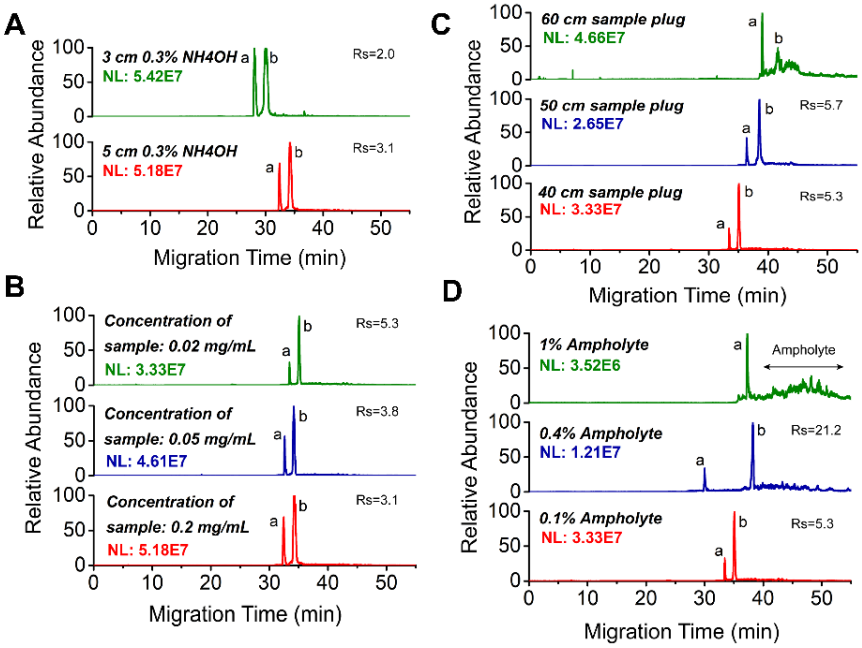

Figure 2.2 Optimization of separation in cIEF-MS/MS analysis with a mixture of cytochrome c (peak a, 12 kDa, pI 10) and myoglobin (peak b, 16.9 kDa, pI 7). Different lengths of catholyte plug (A), lengths of sample plug (B), sample concentration (C), ampholyte concentration (D) were investigated.

We then questioned how we further boosted the number of proteoform identifications from single cIEF-MS/MS run. Inspired by our recent CZE-MS/MS-based TDP work using a 1.5-meter-long LPA-coated capillary,[10] we tried cIEF-MS/MS with a 1.5-meter-long LPA-coated capillary for analysis of the same *E. coli* sample used previously. We loaded roughly 50% of the capillary with sample (80-cm long sample plug) for cIEF-MS/MS in this case. The 1.5-meter capillary offered a higher number of proteoform identifications (449 vs. 281) and peak capacity (92 vs. 77) compared to the 80-cm capillary, Figure 2.1C. In addition, we observed that compared to 0.1% (v/v) FA, the use of 5% (v/v) AA as an anolyte further increased the peak capacity (136 vs. 92) and proteoform identifications (771 vs. 449) by nearly 20% and 50%, respectively, Figure 2.1C. 5% (v/v) AA elongated the protein migration time and achieved a wider separation window, and thereby enhanced the number of proteoform identifications and peak capacity. This is likely because 5% (v/v) AA has a higher viscosity and a lower pH than 0.1% (v/v) FA, which slow down protein migration during the mobilization process. The cIEF-MS/MS using a 1.5-meter-long capillary and 5% (v/v) AA as the anolyte enabled the identification of 711 proteoforms and 177 proteins from

the *E. coli* sample in about 2.5-hours instrument time with a consumption of roughly 480 ng of proteins. We named this method high-capacity cIEF-MS/MS. Interestingly, the high-capacity cIEF-MS/MS method is comparable with the dynamic pH junction-based CZE-MS/MS[10,11] and nanoflow RPLC-MS/MS[37-39] regarding the number of proteoform identifications in a single run. We need to point out that the LPA-coated capillaries prepared in our study are generally durable, which can be continuously used for more than 60 hours for cIEF-MS. All the exciting data render cIEF-MS/MS as another powerful tool for large-scale delineation of proteoforms in complex samples.

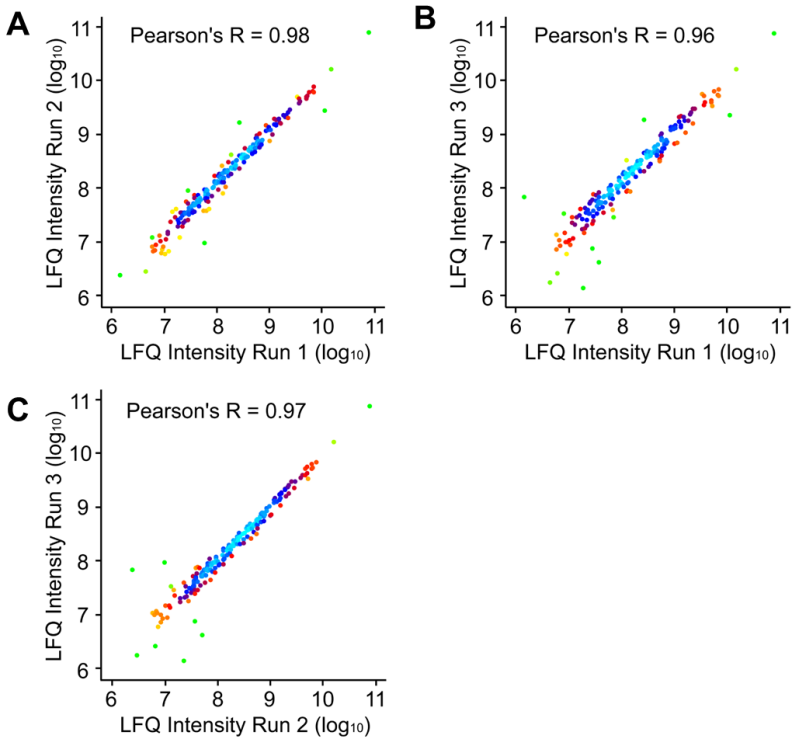

Figure 2.3 Correlations of proteoform label-free quantification (LFQ) intensities between two runs of cIEF-MS/MS analyses of one *E. coli* sample (SEC fraction 3). A: run 1 vs. run 2; B: run 1 vs. run 3; C: run 2 vs. run 3.

2.3.2 Large-scale TDP of *E. coli* cells using SEC-cIEF-MS/MS

2D-PAGE is well known for high-capacity separation of proteoforms based on their molecular weight and pI. Unfortunately, it is challenging to directly couple 2D-PAGE to ESI-MS/MS for TDP due to offline and tedious operations. Here we proposed SEC-cIEF as “gel-free 2D-PAGE” and coupled it to ESI-MS/MS for large-scale TDP for the first time. The *E. coli* proteoforms were first fractionated to six fractions based on their size using SEC, followed by online high-capacity cIEF-MS/MS, Figure 2.4A. Each SEC eluate was further separated into an about 40-minutes separation window by cIEF, indicating good orthogonality of SEC and cIEF for

proteoform separation. The number of identified proteoforms and proteins per SEC fraction ranged from 150 to 711 and 32 to 177, respectively. SDS-PAGE analysis of the SEC fractions showed that SEC offered reasonable separations of proteoforms based on their molecular weights (MWs) with clear MW shift from high to low as the fraction number increased, Figure 2.4B. The mass distribution of identified proteoforms from cIEF-MS/MS analysis of each SEC fraction agreed well with the SDS-PAGE data, Figure 2.4C. Figure 2.4D show the correlations of proteoforms' pIs and migration time from cIEF-MS/MS analyses of two SEC fractions. Basic proteoforms tended to migrate out of the cIEF capillary faster than acidic ones, indicating clear pI-based separations. The data in Figure 2.4D agrees with the cIEF-MS/MS-based BUP data in the literature.[40] Figure 2.4E depicts the cumulative proteoform and protein identifications as a function of the number of SEC fractions with a continuous increase of both protein and proteoform identifications as more SEC fractions were considered.

The SEC-cIEF-MS/MS identified 10,153 proteoform-spectrum matches (PrSMs), 1896 proteoforms and 365 proteins from the E. coli proteome with a 5% proteoform-level FDR, Figure 2.5A and Supplemental Information III. The data represents the first and largest TDP dataset using cIEF-MS/MS. The majority of the identified proteoforms had masses less than 20 kDa, while 83 proteoforms were between 20 and 33 kDa, Figure 2.5B. Although the extracted E. coli proteome consisted of proteins ranging from ~10 kDa to 100 kDa (Figure 2.5B), characterization of proteoforms larger than 30 kDa remains challenging for top-down MS due to dramatic decrease of signal-to-noise ratio with the increase of proteoform's mass, limited mass resolution of mass analyzers, and ion suppression caused by co-eluted small proteins. The number of matched fragment ions of identified proteoforms were in a range of 6 to 92 with the mean at 23, Figure 2.5C. An example of the fragmentation pattern of one proteoform (putative monooxygenase YdhR) is shown in Figure 2.5D. The proteoform was identified with 76 fragment ions, a 1.71e-45 E-value, and a 52% backbone cleavage coverage. On average, we identified about five proteoforms per protein (1896 proteoforms and 365 proteins). For some proteins, the number of proteoforms could be much higher. For instance, we identified 48 proteoforms of the protein Osmotically-inducible protein Y (osmY). All these proteoforms were truncated either at the N-termini (47) or at the C-termini (1). Because TDP directly characterizes intact proteoforms, we were able to determine the distribution of the first amino acid residue position of the truncated proteoforms at the N-termini, Figure 2.5E. For 23 out of the 47 N-terminally truncated proteoforms, the first 28 amino acids residues were cleaved as the signal peptide as reported in the literature.[41] Interestingly, we also identified 3 and 4 proteoforms with the first 27 and 114 amino acid residues truncated, respectively. We then analyzed relative abundance of these proteoforms truncated at different positions based

on the number of PrSMs of each proteoform,[12-42] Figure 2.5F. The 23 proteoforms with the first 28 amino acid residues removed accounted for about 87% of the total number of PrSMs of osmY (248 out of 284). We further examined the 23 proteoforms and discovered that they either had no PTMs or carried various PTMs, e.g., methylation, acetylation, and succinylation. According to their numbers of PrSMs, the proteoform with the first 28 amino acid removed and without any PTMs is the most abundant proteoform of osmY in the *E. coli* cells. The data suggest the power of our SEC-cIEF-MS/MS platform for delineating proteoforms in complex biological samples on a global scale.

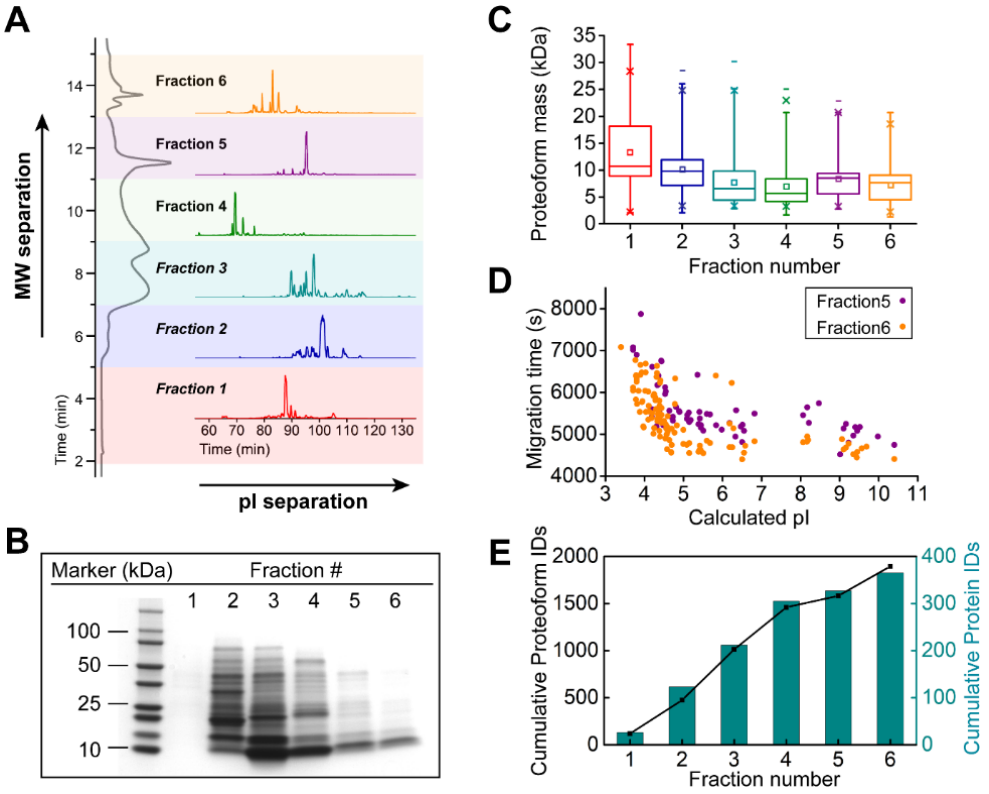

Figure 2.4 Characterization of an *E. coli* proteome using SEC-cIEF-MS/MS. (A) 2D separation of *E. coli* proteome using SEC-cIEF platform. Proteins were fractionated based on molecular weights in SEC dimension (vertical chromatogram) and further be separated according to pI values in cIEF dimension (horizontal electropherograms). (B) SDS-PAGE profiling of proteome in SEC fractions. (C) Box plots of mass distribution of identified proteoforms in SEC fractions. (D) Migration time versus calculated pI value of proteoforms without modifications in SEC fractions 5 and 6. The pI values were calculated using ExPASy (https://web.expasy.org/compute_pi/). (E) Number of proteoform (the black line) and protein identifications (the dark cyan colored bars) cumulated on fractions.

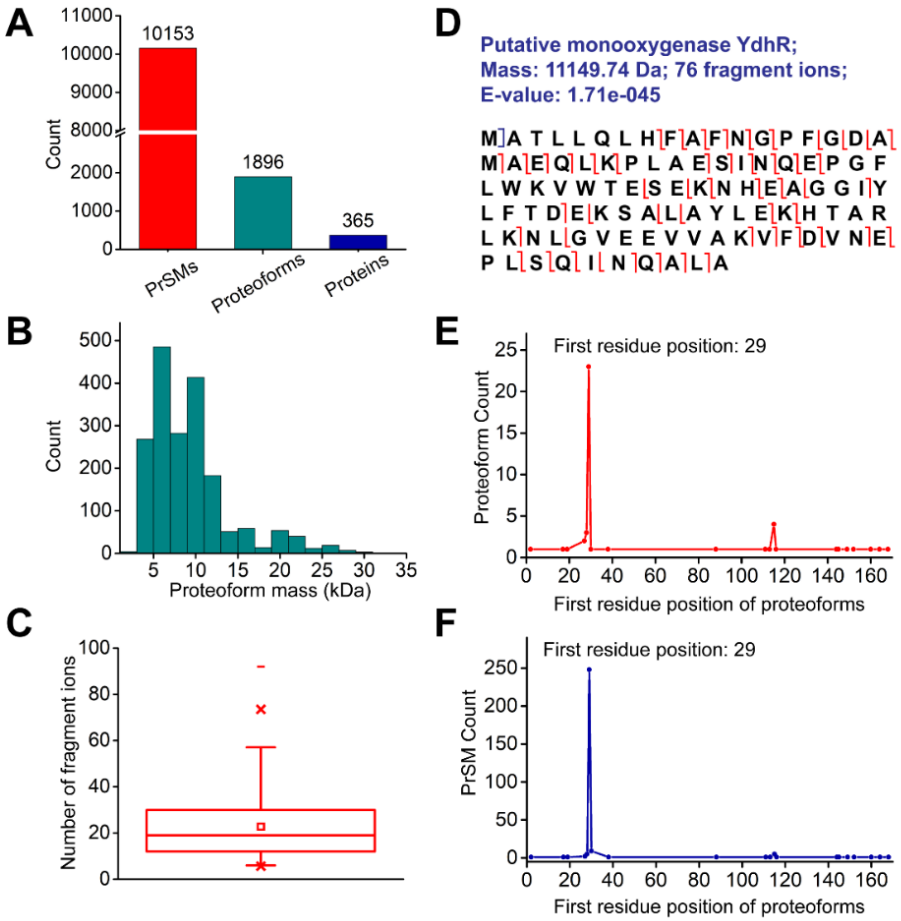

Figure 2.5 Identification results of the *E. coli* proteome from large-scale TDP using SEC-cIEF-MS/MS. (A) Summary of the number of PrSMs, proteoforms, and proteins identified from six SEC fractions of the *E. coli* proteome. (B) Mass distribution of identified proteoforms. (C) Box plot of the number of matched fragment ions of identified proteoforms. (D) Sequence and fragmentation pattern of N-terminal methionine removed *Putative monooxygenase YdhR* with a backbone cleavage coverage of 52%. (E) Proteoform count *versus* the first residue position of truncated proteoforms of *osmY*. (F) PrSM count *versus* the first residue position of truncated proteoforms of *osmY*.

2.4 Conclusions

Top-down proteomics (TDP) requires high-capacity separations of proteoforms before mass spectrometry (MS) and MS/MS. Capillary isoelectric focusing (cIEF)-MS has been recognized as a useful tool for TDP in 1990s because cIEF is capable of high-resolution separation of proteoforms. Previous cIEF-MS studies concentrated on measuring protein's mass without MS/MS, impeding the confident proteoform identification in complex samples as well as

the accurate localization of post-translational modifications (PTMs) on proteoforms. Here, for the first time, we present automated cIEF-MS/MS-based TDP for large-scale delineation of proteoforms in complex proteomes. Single-shot cIEF-MS/MS identified 771 proteoforms from an *E. coli* proteome consuming only nanograms of proteins. Coupling two-dimensional size exclusion chromatography (SEC)-cIEF to ESI-MS/MS enabled the identification of nearly 2000 proteoforms from the *E. coli* proteome. Our study provides the proteome community with a new and powerful tool for the large-scale TDP profiling of complex proteomes.

2.5 Acknowledgments

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's group at the Department of Animal Science of Michigan State University for their help on collecting zebrafish brains for the project. We thank Prof. Xiaowen Liu's group at Indiana University-Purdue University Indianapolis for their help on the top-down proteomics database search using the TopPIC software. We thank the support from the National Institute of General Medical Sciences (NIGMS) through Grant R01GM125991 and the National Science Foundation through Grant DBI1846913 (CAREER Award).

REFERENCES

(1) Smith, L. M.; Kelleher, N. L.Proteoforms as the next proteomics currency. *Science* **2018**, *359*, 1106-1107.

(2) 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*, 499-519.

(3) Cabras, T.; Pisano, E.; Montaldo, C.; Giuca, M. R.; Iavarone, F.; Zampino, G.; Castagnola, M.; Messana, I.Significant modifications of the salivary proteome potentially associated with complications of Down syndrome revealed by top-down proteomics. *Mol. Cell. Proteomics* **2013**, *12*, 1844-1852.

(4) Calligaris, D.; Villard, C.; Lafitte, D.Advances in top-down proteomics for disease biomarker discovery. *J. Proteomics* **2011**, *74*, 920-934.

(5) Li, H.; Nguyen, H. H.; Loo, R. R. O.; Campuzano, I. D.; Loo, J. A.An integrated native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. *Nat. Chem.* **2018**, *10*, 139.

(6) Gomes, F. P.; Yates III, J. R.Recent trends of capillary electrophoresis-mass spectrometry in proteomics research. *Mass Spectrom. Rev.* **2019**, *38*, 445-460.

(7) 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.Identification and quantification of proteoforms by mass spectrometry. *Proteomics* **2019**, *19*, 1800361.

(8) Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L.Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *Trends Anal. Chem.* **2019**, *120*, 115644.

(9) Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavallée-Adam, M.; Yates III, 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.

(10) 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*, 1435-1445.

(11) 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*, 12059-12067.

(12) 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*, 5529-5533.

(13) 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.

(14) Hühner, J.; Lämmerhofer, M.; Neusüß, C.Capillary isoelectric focusing-mass spectrometry: Coupling strategies and applications. *Electrophoresis* **2015**, *36*, 2670-2686.

(15) Shen, Y.; Xiang, F.; Veenstra, T. D.; Fung, E. N.; Smith, R. D.High-resolution capillary isoelectric focusing of complex protein mixtures from lysates of microorganisms. *Anal. Chem.* **1999**, *71*, 5348-5353.

(16) Paša-Tolić, L.; Jensen, P. K.; Anderson, G. A.; Lipton, M. S.; Peden, K. K.; Martinović, S.; Tolić, N.; Bruce, J. E.; Smith, R. D.High throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.* **1999**, *121*, 7949-7950.

(17) Tang, Q.; Harrata, A. K.; Lee, C. S.Capillary isoelectric focusing-electrospray mass spectrometry for protein analysis. *Anal. Chem.* **1995**, *67*, 3515-3519.

(18) Tang, Q.; Harrata, A. K.; Lee, C. S.Two-dimensional analysis of recombinant E. coli proteins using capillary isoelectric focusing electrospray ionization mass spectrometry. *Anal. Chem.* **1997**, *69*, 3177-3182.

(19) Jensen, P. K.; Paša-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*, 2076-2084.

(20) Smith, R. D.; Barinaga, C. J.; Udseth, H. R.Improved electrospray ionization interface for capillary zone electrophoresis-mass spectrometry. *Anal. Chem.* **1988**, *60*, 1948-1952.

(21) Maxwell, E. J.; Zhong, X.; Zhang, H.; van Zeijl, N.; Chen, D. D.Decoupling CE and ESI for a more robust interface with MS. *Electrophoresis* **2010**, *31*, 1130-1137.

(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*, 2312-2321.

(23) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J.Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554-2560.

(24) Mokaddem, M.; Gareil, P.; Varenne, A.Online CIEF-ESI-MS in glycerol–water media with a view to hydrophobic protein applications. *Electrophoresis* **2009**, *30*, 4040-4048.

(25) Zhong, X.; Maxwell, E. J.; Ratnayake, C.; Mack, S.; Chen, D. D.Flow-through microvial facilitating interface of capillary isoelectric focusing and electrospray ionization mass spectrometry. *Anal. Chem.* **2011**, *83*, 8748-8755.

(26) Zhu, G.; Sun, L.; Dovichi, N. J.Simplified capillary isoelectric focusing with chemical mobilization for intact protein analysis. *J. Sep. Sci.* **2017**, *40*, 948-953.

(27) Dai, J.; Lamp, J.; Xia, Q.; Zhang, Y.Capillary isoelectric focusing-mass spectrometry method for the separation and online characterization of intact monoclonal antibody charge variants. *Anal. Chem.* **2018**, *90*, 2246-2254.

(28) Lechner, A.; Giorgetti, J.; Gahoual, R.; Beck, A.; Leize-Wagner, E.; François, Y.-N.Insights from capillary electrophoresis approaches for characterization of monoclonal antibodies and antibody drug conjugates in the period 2016–2018. *J. Chromatogr. B* **2019**, *1122*, 1-17.

(29) Wang, L.; Bo, T.; Zhang, Z.; Wang, G.; Tong, W.; Da Yong Chen, D.High resolution capillary isoelectric focusing mass spectrometry analysis of peptides, proteins, and monoclonal antibodies with a flow-through microvial interface. *Anal. Chem.* **2018**, *90*, 9495-9503.

(30) Wang, L.; Chen, D. D. Y.Analysis of four therapeutic monoclonal antibodies by online capillary isoelectric focusing directly coupled to quadrupole time-of-flight mass spectrometry. *Electrophoresis* **2019**, *40*, 2899-2907.

(31) 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.

(32) 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.* **2013**, *52*, 13661-13664.

(33) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P.ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534-2536.

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

(35) 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*, 207-214.

(36) 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*, 5383-5392.

(37) 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.Identification and characterization of human proteoforms by top-down LC-21 tesla FT-ICR mass spectrometry. *J. Proteome Res.* **2017**, *16*, 1087-1096.

(38) Liu, Z.; Wang, R.; Liu, J.; Sun, R.; Wang, F.Global quantification of intact proteins via chemical isotope labeling and mass spectrometry. *J. Proteome Res.* **2019**, *18*, 2185-2194.

(39) 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*, 8553-8560.

(40) Zhu, G.; Sun, L.; Yang, P.; Dovichi, N. J.On-line amino acid-based capillary isoelectric focusing-ESI-MS/MS for protein digests analysis. *Anal. Chim. Acta* **2012**, *750*, 207-211.

(41) Yim, H. H.; Villarejo, M.osmY, a new hyperosmotically inducible gene, encodes a periplasmic protein in Escherichia coli. *J. Bacteriol.* **1992**, *174*, 3637-3644.

(42) Geis-Asteggiante, L.; Ostrand-Rosenberg, S.; Fenselau, C.; Edwards, N. J.Evaluation of spectral counting for relative quantitation of proteoforms in top-down proteomics. *Anal. Chem.* **2016**, *88*, 10900-10907.

CHAPTER 3.* Improved cIEF-MS for ultrahigh-resolution characterization of charge variants of biotherapeutics

3.1 Introduction

As an emerging class of therapeutic proteins with high specificity, monoclonal antibodies (mAbs) have shown great potential for the treatment of cancers, virus infections, and autoimmune disorders in recent years.[1-6] In the process of mAb manufacturing, the formation of undesired post-translational modifications (PTMs), such as asparagine/glutamine deamidation, glycation, C-terminal proline amidation, iso-Asp modification, and methionine oxidation, are critical quality attributes (CQAs). PTMs can alter the surface charge distribution and conformation of a mAb, which derives charge heterogeneity and influences its pharmacological effects. For example, asparagine deamidation of trastuzumab has been found to reduce antibody charge and change HER2 antigen binding activity.[7] Therefore, routine characterization of antibody charge heterogeneity and PTMs are necessary to guarantee product stability, safety, and efficacy.

Capillary isoelectric focusing (cIEF)[8-10], capillary zone electrophoresis (CZE)[11,12], and ion exchange chromatography (IEX)[13-15], are frequently employed techniques for monitoring charge heterogeneity of mAbs in the pharmaceutical field. CZE with a background electrolyte (BGE) containing ε-amino-caproic acid (EACA) has demonstrated high-resolution separation of mAb charge variants.[12] Conventional imaged cIEF (icIEF) is reliable for examining the relative abundance of charge variants and performing pI measurement. However, the results gathered from optical detections (i.e., UV and fluorescence) remain less informative for charge variant identifications. Moreover, mAb charge variants were typically separated using IEX, followed by fraction collection, buffer exchange, and liquid chromatography (LC)-mass spectrometry (MS) analysis. The method also has limited application as it is labor-intensive, difficult to provide comparable charge profiles with icIEF-UV, and unable to provide pI information.

Developing platforms that directly integrate the separation methods above with MS characterization is highly desirable for industrial use. cIEF-MS is of most interest as it provides ultra-high resolving power for mAb charge variants with subtle isoelectric point (pI) difference and enables sensitive detection. The hyphenation of cIEF with electrospray ionization (ESI)-MS has been initiated since the 1990s for characterizing pIs and masses of standard proteins and

---

* This chapter is adapted with permission from *Xu, T.; Han, L.; Thompson, A. M. G.; Sun, L. Anal. Methods 2022, 14 (4), 383-393.*

complex proteomes.[16-19] However, for a long time, it was not accepted for wide applications due to complicated operations, semi-online coupling, lack of highly sensitive CE-MS interface, and suppression of ionizations by ampholytes.

Microchip-based capillary electrophoresis (CE) devices have been developed and commercialized for the comprehensive characterization of charge variants of mAbs using MS detection.[20-25] The cIEF-MS microchip system of Blaze (Intabio, CA) integrated imaged cIEF feature and ESI tip into a microchip. The system was coupled with MS for real-time monitoring of focusing and mobilization processes of cIEF, and enabled efficient characterization of mAb charge variants in 15 min.[20] The ZipChip system of 908 devices is another type of microfluidic-based technique that uses capillary zone electrophoresis (CZE)-MS to facilitate fast screening of mAb variants,[21-24] antibody-drug conjugate variants,[25] and perform peptide mapping and mAb PTM identification[23,24]. These microchip-based devices are generally easy to operate, and provide fast speed and high-resolution separation, making them highly attractive tools in pharmaceutical applications.

So far, regular CE systems remain dominant in CE-MS characterization of proteins. Compared with microfluidic devices, regular CE systems give more flexibility for separation optimization (e.g. adjusting the capillary length, applying different CE separation modes) according to different subjects. In recent years, substantial progress has been achieved in improving the performance of cIEF-MS system. The new generation of sheath-flow CE-MS interfaces, including flow-through microvial interface and electrokinetically pumped sheath flow interface, enable ultra-low flow rate of sheath liquid at nL per min level (~two magnitudes lower than the traditional interfaces), thereby largely improving the sensitivity of CE-MS.[26-28] Fully automated cIEF-MS has been achieved with the advent of “sandwich” injection configuration.[29-31] Moreover, multiple strategies have been developed to reduce the influence of ampholytes on ionization, including choosing compromised ampholyte concentration, incorporating a microdialysis device into cIEF-MS interface, removal of ampholytes using two-dimensional systems (cIEF-liquid chromatography or cIEF-CZE), and utilization of immobilized pH gradient.[19, 32-37]

The cIEF-MS platforms have recently been used for the characterization of charge variants of mAbs, and large-scale top-down proteomics.[38-41] Dai et al. developed the first online cIEF-MS approach on an Agilent CE system for the characterization of mAb charge variants using an electrokinetically pumped sheath flow interface and “sandwich” injection configuration.[38] Their cIEF-MS platform was successfully applied for resolving the charge variants of trastuzumab, bevacizumab, infliximab, and cetuximab. Later, Wang et al. established a similar cIEF-MS

platform with a flow microvial interface and “sandwich” injection configuration.[39, 40] This platform enabled the separation of mAb charge variants with a pI difference of only 0.02-0.2 pH unit and provided accurate measurement of intact masses. In these studies, capillaries with neutral coating were employed for cIEF separation, which was crucial for maintaining high separation resolution by suppressing electroosmotic flow (EOF) and minimizing protein adsorption to the capillary wall. The additives (e.g., methylcellulose) typically used in conventional icIEF-UV for suppressing EOF and preventing protein adsorption need to be avoided in cIEF-MS.

The stability of capillary coating could be challenged in cIEF-MS by the usage of extremely basic catholyte (ammonium hydroxide, pH >11). Reducing catholyte pH lower than 10 was considered an effective way to eliminate coating degradation. Ramsay et al. compared sodium hydroxide (40mM, pH 12) and CAPS (100 mM, pH 9.8) as catholyte for cIEF- laser-induced fluorescence (LIF) analysis of gastric biopsies.[42] They found the catholyte with lower pH (CAPS, pH 9.8) maintained a better reproducibility of cIEF separation. The catholyte developed in the study was only applicable to optical detection. Based on our best knowledge, there is still no cIEF-MS study that developed MS-compatible catholyte buffer with pH 10 for protein analysis to improve the system’s reproducibility and stability.

Here, we improved the automated cIEF-MS method to achieve reproducible and high-resolution separation of charge variants of mAbs. To address the coating degradation problem, a highly durable linear polyacrylamide (LPA) capillary coating was developed, and an ESI-MS-friendly catholyte (10 mM ammonium bicarbonate, pH~10) was employed. Critical parameters including catholyte length, sample concentration, ampholyte composition as well as ampholyte concentration were systematically investigated to improve separation resolution. Using the optimized separation parameters and a high-resolution quadrupole-time-of-flight (Q-TOF) mass spectrometer, we characterized charge variants of NISTmAb, and mAb1 under denaturing conditions. The methods developed in this work can potentially be adapted for analyzing charging variants of other mAbs.

3.2 Experimental section

3.2.1 Chemicals and materials

Ammonium bicarbonate (ABC), ammonium hydroxide (NH_4OH), 3-(Trimethoxysilyl) propyl methacrylate, glycerol, cytochrome c, Pharmalyte 3-10, 5-8, 8-10.5 (GE healthcare), and Amicon Ultra (0.5 mL, 10 kDa cut-off size) centrifugal filter units were purchased from Sigma-Aldrich (St. Louis, MO). LC/MS grade water, acetonitrile (ACN), HPLC-grade acetic acid (AA), fused silica capillaries (50 µm i.d., 360 µm o.d., Polymicro Technologies) were ordered from Fisher Scientific

(Pittsburgh, PA). Acrylamide was purchased from Acros Organics (Fair Lawn, NJ). Five peptide markers with pI values of 4.1, 5.5, 7.0, 9.8, 10.0 were obtained from Beckman Coulter (Brea, CA). cIEF-MS reagent kit containing anolyte (Buffer A, pH 2.7), catholyte (Buffer B, pH 11.6), solution for buffer exchange and sample dilution (Buffer C), sheath liquid (Buffer SL), and ampholyte buffer (buffer S35) was obtained from CMP Scientific (Brooklyn, NY). mAb1 was provided by AbbVie (North Chicago, IL). NIST mAb was obtained from Sigma-Aldrich (St. Louis, MO).

3.2.2 Sample preparation

100 µg each of NISTmAb and mAb1(10 mg/mL, containing 12.5 mM histidine and 12.5mM histidine HCl, pH 6.0) were diluted in 200 µL Buffer C, and then centrifuged via Amicon Ultra centrifugal filters (0.5 mL, 10 kDa cut-off size) at 14,000 g for 10 min. After spinning down, 200 µL Buffer C was added into the filters and centrifuged. The process was repeated two times to thoroughly replace the sample buffer with Buffer C. After buffer exchange, sample volume was adjusted to 50 µL with Buffer C, and the concentrations of mAbs were roughly 2 mg/mL assuming no sample loss in desalting process. The desalted mAbs were stored in -20 °C. Before sample analysis, the desalted mAbs (2 mg/mL) were further diluted with Buffer C and then mixed with ampholyte stock solution (2×) at a ratio of 1:1.

3.2.3 Preparation of linear polyacrylamide (LPA)-coated capillary

The procedure for preparation of LPA-coated capillary (50 µm i.d., 360 µm o.d.) was slightly modified according to the previous literature.[43] After capillary pretreatment, the capillary was filled with degassed acrylamide-ammonium persulfate mixture and incubated in 50 ˚C water bath for reaction. We extended the incubation time from 45 minutes to 60 minutes to achieve a highly durable capillary coating. The solution in the capillary was pushed out after the 60-min reaction using a high-pressure liquid chromatography pump.

3.2.4 Online and automated cIEF-MS analysis

The online and automated cIEF-MS platform was constructed by integrating an Agilent 7100 CE System with an Agilent 6545XT Q-TOF mass spectrometer via a commercialized electrokinetically pumped sheath-flow CE-MS nano-spray interface (EMASS-II, CMP Scientific). The ESI emitter (orifice size: 25~35 µm) on the interface was filled with sheath liquid, and the emitter orifice was positioned 4~5 mm away from the inlet of the mass spectrometer. The electrospray voltage was carefully adjusted in the range of 2.3 kV to 2.5 kV to generate a stable electrospray.

A 75 cm LPA-coated capillary was used for cIEF separation. Two ends of the capillary were installed in the ESI emitter and CE system, respectively. The “sandwich” injection strategy was adopted for automated cIEF-MS analysis. Precisely, the capillary was filled with a plug of

catholyte (10 mM ABC, 15% glycerol, pH 10), a plug of mAb-ampholyte mixture, and then inserted into a vial with anolyte (Buffer A from cIEF kit or buffer with 0.1%FA and 15% glycerol, pH ~2.7). After sample loading, a voltage of 20 kV was applied to the capillary injection end to facilitate protein focusing and mobilization. After 20 min, a pressure (10 mbar) was applied at the capillary inlet to assist protein mobilization. Finally, the capillary was flushed with anolyte with a pressure of 950 mbar for 5 min. The total analysis time including focusing and mobilization was around 75 minutes.

For the Q-TOF mass spectrometer, a regular ESI spray shield was installed, and drying gas (325 °C) was set to a low flow rate (1 L/min) to maintain the stability of the electrospray. The voltages for VCap, skimmer, and fragmentor were set to 0 V, 300 V, and 380 V, respectively. The collision energy was set at 10 V to assist the transmission of gas-phase mAb ions in the mass spectrometer. Full scan mass spectra were collected in the m/z range of 2000-6000 with an acquisition rate of 0.5 spectra/sec.

3.2.5 Data analysis

The cIEF electropherograms and MS data were analyzed using Agilent Mass-Hunter Qualitative Navigator B.08.00. The intact masses of mAb charge variants were obtained by manual averaging across the base peak electropherogram of a specific peak and performing deconvolution in Agilent Mass-Hunter Bio-Confirm 10.0. Parameters for deconvolution were set as follows: Maximum Entropy algorithm, mass step of 0.05 Da. The other parameters were kept as default.

3.2.6 Capillary equilibrium and cleanup

The new LPA-coated capillaries were equilibrated by flushing with bovine serum albumin (BSA) or mAb samples for 2 minutes to block residual adsorption sites. The capillary cleanup was performed after every three cIEF-MS runs to remove proteins adsorbed on the capillary inner wall. The capillary was flushed with 4 M urea for 10 minutes, water for 5 minutes, and anolyte for 5 minutes successively.

3.3 Results and discussions

3.3.1 Development of a robust automated cIEF-MS method using NISTmAb

A stable neutral coating is critical for achieving high separation resolution in cIEF. Different types of neutrally coated capillaries have been developed previously, including LPA coating, poly (vinyl alcohol) (PVA) coating, and hydroxypropyl cellulose (HPC) coating. The LPA coating can be made easily and is the most widely used one. Here we chose the capillaries with LPA coating (~75 cm in length) for cIEF-MS analysis of mAbs.

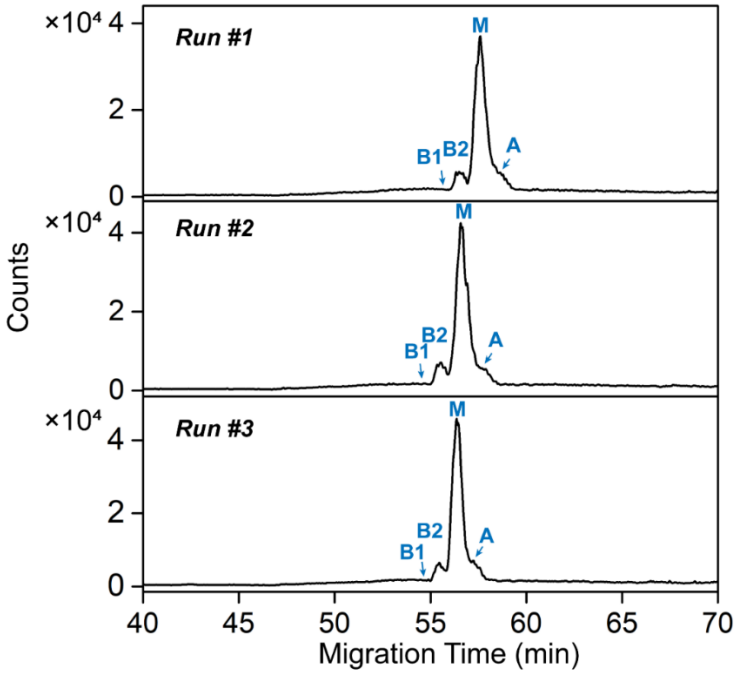

Figure 3.1 Triplicate cIEF-MS runs of NISTmAb with a pH 10.0 catholyte and an LPA-coated capillary (75 cm). Other parameters for cIEF separation were 30 cm catholyte plug, 45 cm sample plug, 0.2 mg/mL sample concentration, 1.5% ampholyte mixture (pH range 3-10 and 8-10.5 with ratio of 1:4), 20 kV separation voltage, 10 mbar pressure applied at 20 min.

A NISTmAb sample and a commercialized cIEF-MS kit (CMP scientific) were used to test separation performance. We observed quick degradation of the capillary coating after one or two cIEF-MS runs (12-cm catholyte plug, 63-cm sample plug) due to the highly basic catholyte (pH 11.6) from the kit. It is a common issue in cIEF as the Si-O bond is susceptible to hydrolysis under high pH, leading to the destruction of capillary coating. To address this issue, we prolonged the reaction time for LPA coating synthesis from regular 45 minutes to 60 minutes to achieve a highly durable capillary coating. Additionally, an ESI-MS-friendly catholyte with lower pH (10 mM ABC, 15% glycerol, pH 10) was employed to mitigate the hydrolysis effect. Using the same separation parameters, however, we found a decrease in catholyte pH had an adverse impact on separation performance because the pH gradient could not hold enough time for protein focusing. This problem was overcome by increasing the catholyte length from 12 cm to 30 cm. Overall, the separation profile of NISTmAb can be reproduced using the optimized catholyte condition (Figure 3.1). For certain pI regions, further optimization is still necessary to achieve more resolved separations. The improved cIEF-MS with the LPA-coated capillary was able to perform more than 25 runs of various mAb samples without significant deterioration of separation performance.

We observed that the electrospray was not stable in some cases. We used sheath liquid containing 25% ACN and 20% AA for our experiment. Changing compositions of the sheath liquid (15% ACN, 10% AA; 15% methanol, 10% AA) did not give significant improvements. We then optimized the position of ESI emitter relative to MS inlet. We found that when the emitter was too close to the MS inlet, it led to a significant change in high vacuum values inside the TOF tube, which was the major reason for signal variations. Therefore, to achieve a stable MS response, the spray emitter orifice should be positioned 4-5 mm away from the MS inlet and then shifted 1~2 mm away from the central zone of the MS inlet.

Furthermore, we investigated the impact of ampholyte composition and sample concentration on separation resolution. Cytochrome c (0.05 mg/mL, pI~10) was spiked into the NISTmAb sample as a pI marker to normalize the migration time and evaluate the separation window. Figure 3.2 indicated that the ampholyte with and without the narrow pI range 5-8 produced a minimal difference in the separation resolution of NISTmAb charge variants. It could be due to that NISTmAb variants are highly basic (pI 9.0~9.4),[44, 45] and the addition of the pI range of 5-8 cannot significantly promote focusing of the acidic charge variants. When the concentration of NISTmAb was increased from 0.2 to 0.8 mg/mL, the MS signal increased by nearly 10-folds without significant reduction of separation resolution (Figure 3.2), indicating the 0.2-0.8 mg/mL should at least be a good range of sample concentration for experiments.

Four charge variants of NISTmAb were observed in Figure 3.2C. Each variant included various glycoforms, which were resolved by MS (Figure 3.3A). The intact masses of the charge variants, Figure 3.3B, were obtained by averaging across each variant peak and performing deconvolution. The modifications that appeared in acidic or basic variants were estimated by comparing the mass difference to corresponding glycoforms in the main peak. For example, 148038.0 Da in the main peak of NISTmAb can be assigned to a mAb form containing glycan pairs G0F/G0F and two lysine clipping on heavy chain C-terminus (theoretical mass: 148037.2 Da). The two basic variants (148165.8 Da in B1, 148294.4 Da in B2) exhibited +127.8 Da and +256.4 Da mass shift relative to the G0F/G0F form in the main peak (148038.0 Da), corresponding to one and two incomplete C-terminal lysine clipping. In addition, the acidic variant (148200.9 Da) showed +162.9 Da mass shift from the main species (148038.0 Da), indicating a glycation modification on lysine. The NISTmAb charge variants detected in our study were consistent with previous reports,[45, 46] which suggests high accuracy of our cIEF-MS system for mAb characterization.

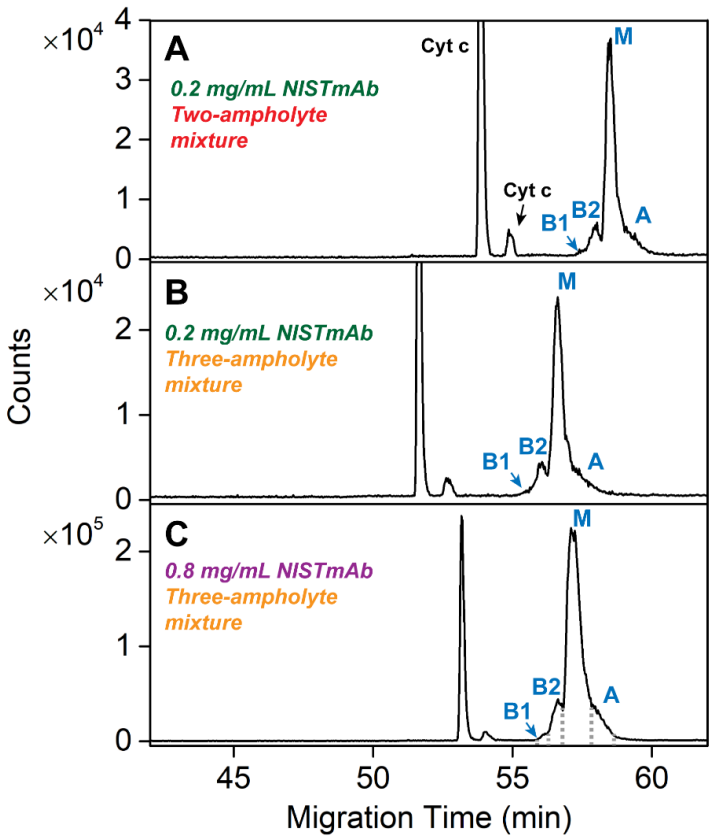

Figure 3.2 Comparison of cIEF separation of NISTmAb charge variants with different ampholyte compositions and sample concentrations. (A, B) Comparison of cIEF separation of NISTmAb between using a two-ampholyte mixture and a three-ampholyte mixture. The two-ampholyte mixture contains 1.5% ampholytes with pH range of 3-10 and 8-10.5 (ratio 1:4). The three-ampholyte mixture comprises 2% ampholytes with pH range of 3-10, 5-8 and 8-10.5 (ratio: 1:2:4). (B, C) Comparison of cIEF separation and MS signal of NISTmAb between using 0.2 mg/mL and 0.8 mg/mL sample concentration. Other parameters for separation: 75 cm LPA-coated capillary, 30 cm catholyte plug, 45 cm sample plug, 0.05 mg/mL cytochrome c, 20 kV separation voltage, 10 mbar pressure applied at 20 min.

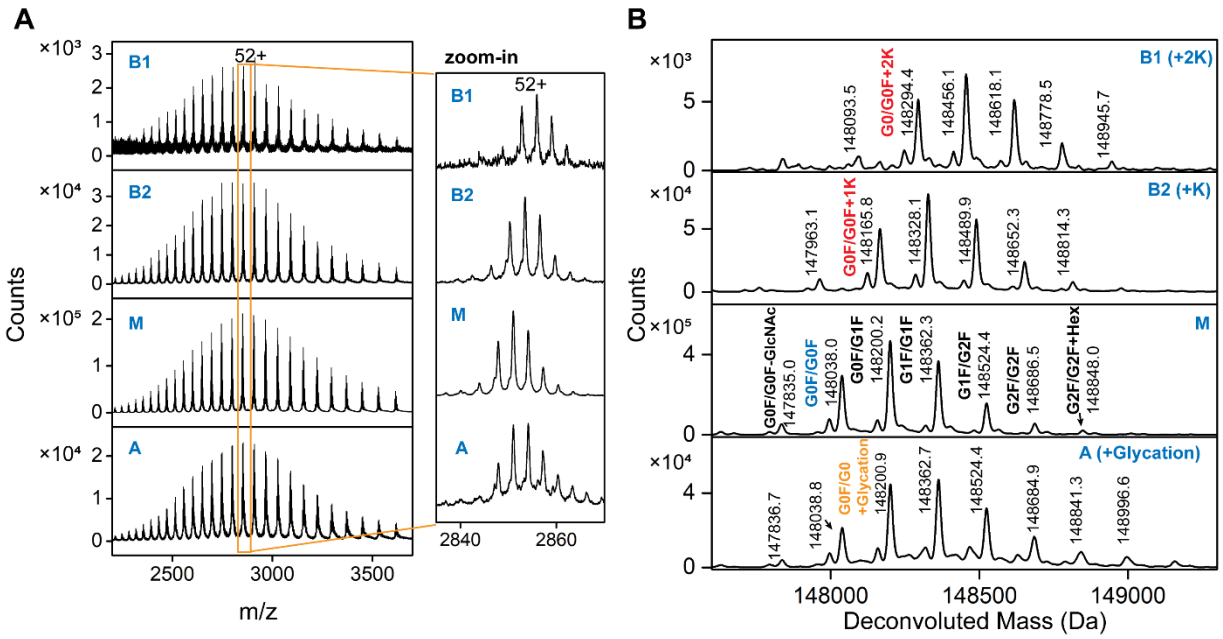

Figure 3.3 Mass spectra (A) and deconvoluted mass spectra (B) of NISTmAb charge variants from two basic peaks (peak B1 and B2), a main peak (peak M) and an acid peak (peak A).

3.3.2 cIEF-MS characterization of mAb1

mAb1 is one model lgG1 mAb manufactured by AbbVie. Here, our improved cIEF-MS platform was applied to understand mAb1 charge variants. In our experiment, the mAb1 sample was directly diluted to 0.8 mg/mL with a sample buffer (buffer C) for cIEF analysis. The cIEF-MS analysis was started by using 0.2 mg/mL mAb1 and a two-ampholyte mixture (pI range of 3-10 and 8-10.5), resulting in three charge variants (a main peak M and two acidic peaks A1, A2), Figure 3.4A. Incorporating pI 5-8 ampholyte and increasing the sample concentration to 0.8 mg/mL further generated an additional acidic variant (A3) and two basic variants (B1, B2), Figure 3.4B and Figure 3.4C. Moreover, we tested direct cIEF-MS analysis of mAb1 sample (11.8 mg/mL mAb1 in 30 mM histidine and 8% sucrose) without desalting to examine if it is possible to simplify the sample preparation process and improve analysis efficiency. In our experiment, the mAb1 sample was directly diluted to 0.8 mg/mL with a sample buffer (buffer C) for cIEF analysis.

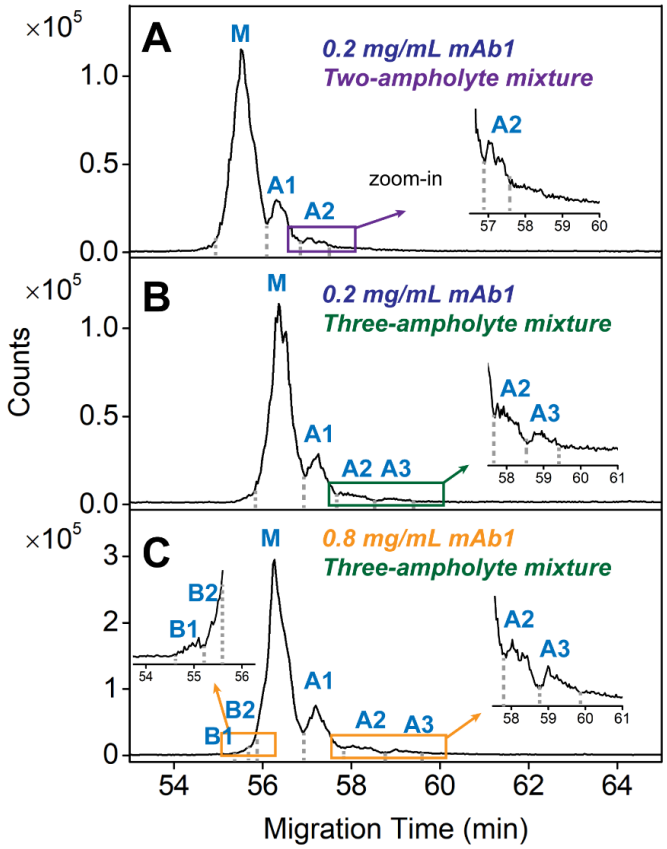

Figure 3.4 Improvement of cIEF separation and MS signal of mAb1 by incorporating a narrow range ampholyte (pH range of 5-8) into a two-ampholyte mixture (pH range of 3-10 and 8-10.5) and using higher sample concentration. (A-C) Base peak electropherograms of mAb1 with two-ampholyte mixture, 0.2 mg/mL sample; three-ampholyte mixture, 0.2 mg/mL sample; and three-ampholyte mixture, 0.8 mg/mL sample. The two-ampholyte mixture used in the experiment contains 1.5% ampholytes with pH range of 3-10 and 8-10.5 (ratio: 1:4). The three-ampholyte mixture comprises 2% ampholytes with pH range of 3-10, 5-8 and 8-10.5 (ratio: 1:2:4). Other parameters for cIEF separation: 75 cm LPA-coated capillary, 30 cm catholyte plug, 45 cm sample plug, 20 kV separation voltage, 10 mbar pressure applied at 20 min.

As a consequence, the signal of mAb increased around 2-fold, and ten charge variants were detected, Figure 3.5, which could be due to less sample loss from the simplified sample preparation workflow. The data here also suggests that antibody samples could be directly diluted for cIEF-MS analysis without desalting or buffer exchange in some cases.

Figure 3.5C presents representative deconvoluted MS spectra of mAb1 charge variants (peak B1, B2, M, A1 and A5). In the main peak of mAb1 (M), the most abundant spectral peak (147610.8 Da) indicated a glycoform with biantennary glycan pairs G0F/G0F (theoretical mass: 147609.4 Da), whereas the other two low abundant peaks (147772.2 Da and 147406.9 Da) represented glycoforms with G0F/G1F and G0F/G0F-GlcNAc, respectively. We compared spectral peaks in the charge variants with G0F/G0F glycoform in the main peak (147610.8 Da). A -2636.2 Da (B1, 144974.6 Da) mass shift was observed in the first basic peak (B1), which is associated with loss of glycan pairs G0F/G0F and two uncleaved C-terminal lysines (theoretical mass shift: 2633 Da). In addition, we identified a -58.9 Da (B2, 147551.9Da) and a +127.3 Da (B2, 147738.1 Da) mass shift in the second basic variant, which might originate from a PGK amidation (theoretical mass shift: -58 Da) and one incomplete C-terminal lysine truncation, respectively. A +1.7 Da mass difference discovered in the first acidic variant (A1) could be assigned as two deamidations. Moreover, the fifth acidic peak (A5) clearly identified three mAb1 variants with a Fab arm missing. Their masses (100697.3 Da, 100458.9 Da, and 100214.6 Da) are well-matched with G0F/G0F glycoform truncated at different sites of the upper hinge region (theoretical mass: 100697.2 Da for the cleavage between C224 and D225, 100453.9 Da for the cleavage between K226 and T227 and 100215.6 Da for the cleavage between H228 and T229), which was previously reported as degradation hotspot of lgG1.[40]

Apart from commonly found PTMs and truncations, our result also presented several unknown mass shifts, such as -199.6 Da mass shift in the first basic variant (B1, 147411.2 Da), a 63.5 Da mass difference in the acidic variant (A1, 147674.3 Da). It is difficult to explain those uncommon mass shifts on an unknown mAb solely based on intact level analysis. In this case, peptide mapping of these variants could be helpful to elucidate PTMs on those variants.

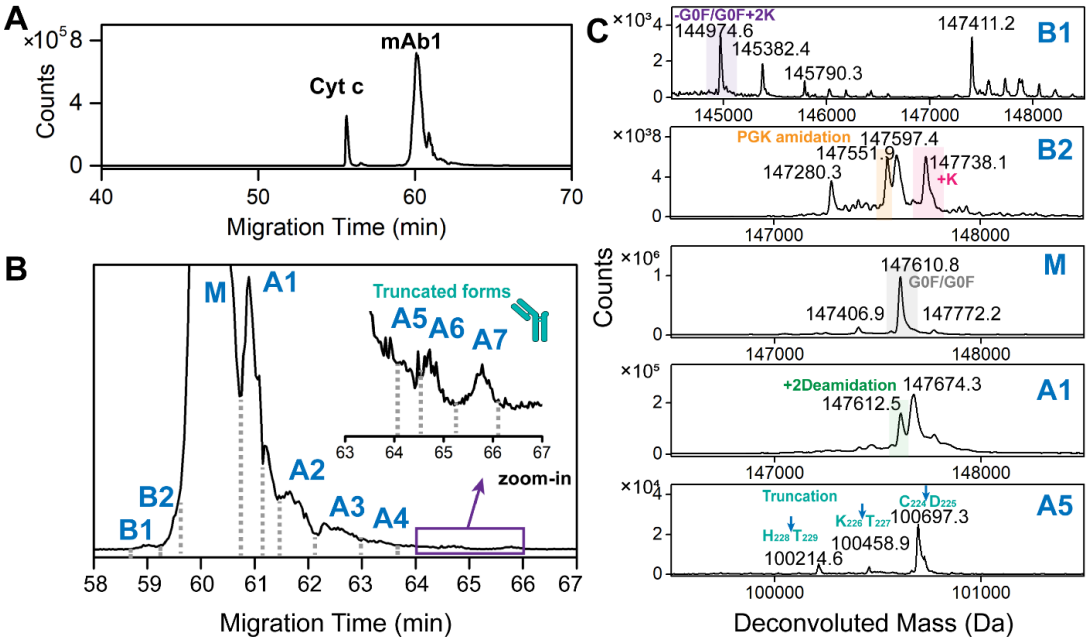

Figure 3.5 Direct cIEF-MS analysis of mAb1 without sample desalting and buffer exchange. (A) Base peak electropherogram of mAb1 and cytochrome c. (B) Zoomed-in base peak electropherogram of mAb1. Eight charge variants of mAb1 were separated and characterized via cIEF-MS platform. (C) Deconvoluted mass spectra of a basic peak (B), a main peak (M), two acidic peaks (A1 and A5). Parameters for cIEF separation: 75 cm LPA-coated capillary, 30 cm catholyte plug, 45 cm sample plug, 0.8 mg/mL mAb1, 0.05 mg/mL cytochrome c, 2% three-ampholyte mixture (pH range of 3-10, 5-8 and 8-10.5, ratio 1:2:4), 20 kV separation voltage, 10 mbar pressure applied at 20 min.

3.3.3 pI determination of charge variants of mAb1

pI is an important property of mAbs, which is known to influence their pharmacokinetic behaviors and half-life.[49, 50] Majority mAbs have pIs in the range of 8-9.5.[44, 49] Under physiological pH (pH~7.4), mAbs are positively charged and can adhere to negatively charged sites of cell surfaces. An increase/decrease of antibody pI by PTMs usually leads to higher/less target tissue uptake and shorter/longer half-time.[49] cIEF has been accepted as a standard method for determining protein pI. Unlike conventional icIEF which performs whole-column imaging and does not require protein mobilization, cIEF-MS needs to migrate focused proteins to MS for analysis by chemical mobilization or pressure. Disturbance of pH gradient can easily occur during this process, which causes a problem for pI determination.

In our study, 15% glycerol was added in catholyte, anolyte, and sample buffer as an anti-convection reagent to maintain pH gradient and separation resolution according to the literature.[38]

Linearity of pH gradient was studied by cIEF-MS analysis of a mixture containing six peptide pI markers (pI 4.05, 5.52, 7.00, 9.50, 9.99), Figure 3.6. Linear regression of pI values of pI markers versus mobilization times exhibited a good correlation coefficient (R^2=0.999) with the wide range ampholyte (pI 3-10). Distortion of linearity at acidic pH range was observed when narrow range ampholyte (pI 8-10.5) was incorporated into pI 3-10 ampholyte. This phenomenon was because much higher concentration of basic ampholytes caused expansion of basic pH gradient to a wider range in the capillary, thereby shrinking and distorting acidic pH gradient. Nevertheless, it still maintained excellent linearity in the basic range (pH 7-10, R^2=0.999). Moreover, the separation window in the pH range of 7-10 was doubled with the addition of narrow pI range ampholyte due to the expansion of basic pH gradient.

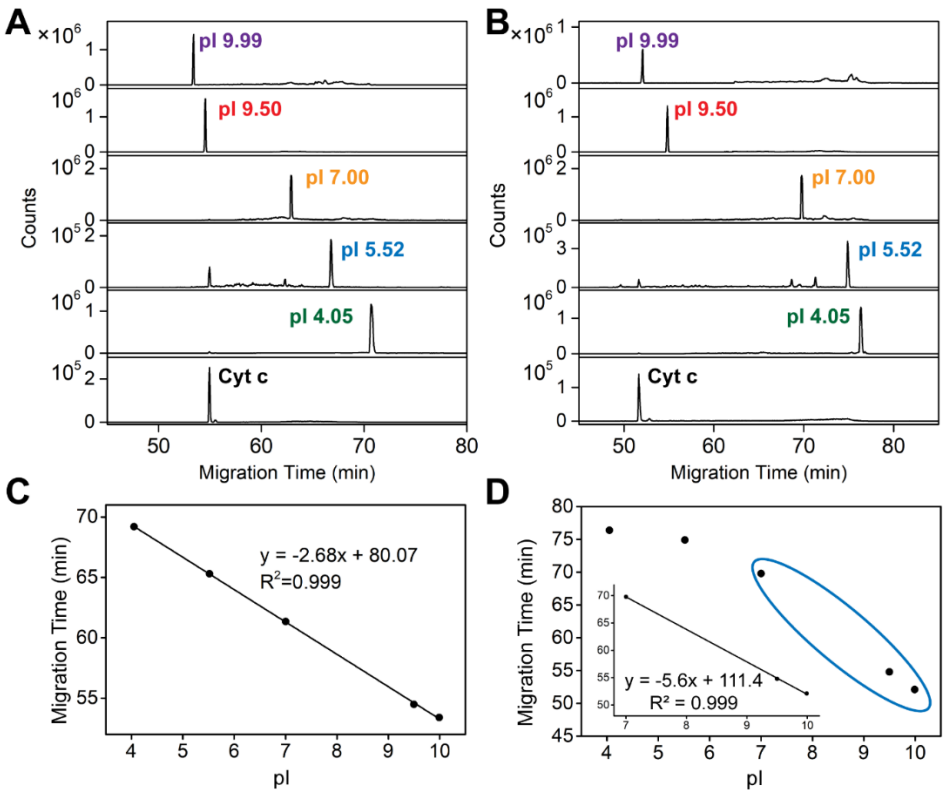

Figure 3.6 Direct cIEF-MS analysis of mAb1 without sample desalting and buffer exchange. (A) Base peak electropherogram of mAb1 and cytochrome c. (B) Zoomed-in base peak electropherogram of mAb1. Eight charge variants of mAb1 were separated and characterized via cIEF-MS platform. (C) Deconvoluted mass spectra of a basic peak (B), a main peak (M), two acidic peaks (A1 and A5).

We evaluated run-to-run variation of migration time in cIEF-MS analysis. For triplicate runs of mAb1, the relative standard deviations (RSDs) of absolute migration time (t1 in Table 3-1) of

mAb charge variants were about 6%. We then aligned the migration time across runs based on the cytochrome c (cyto.c) peak, and the RSDs of migration time (t2 in Table 3-1) became only 1.0%. As cIEF-MS analyses of peptide pI markers and mAb1 were performed with the same separation conditions (ampholyte pI 3-10 and 8-10.5), the linear regression equation (y=-5.6x+111.4) in Figure 3.6D achieved from the analysis of pI markers was used for determining the pIs of mAb1 charge variants. Before pI calculation, the migration times of mAb1 charge variants were normalized to the cyto.c data to reduce migration time variations across runs. As shown in Table 3-1, the calculated pI of the main species of mAb1 is 9.16 ± 0.01, which agrees with the pI value obtained from icIEF-UV analysis (pI 9.23, Figure 3.6). The other three acidic variants (A1, A2, A3) have pIs of 9.01± 0.01, 8.85± 0.00, and 8.72± 0.01, respectively, indicating 0.1~0.3 pI shifts by acidic modifications. The data highlight the potency of our cIEF-MS method for the accurate determination of pIs of proteins in typical top-down MS studies.

Table 3-1 Calculation of pIs of mAb1 charge variants.

	Data 1			Data 2			Data 3			Mean	SD
	t1	t2	pI	t1	t2	pI	t1	t2	pI		
M	59.6	57	9.15	53.3	56.9	9.17	56.3	56.9	9.17	9.16	0.01
A1	60.5	57.9	9	54.2	57.8	9.01	57.2	57.8	9.01	9.01	0.01
A2	61.4	58.8	8.85	55.2	58.8	8.85	58.2	58.8	8.85	8.85	0.00
A3	62.1	59.5	8.73	56	59.6	8.71	59	59.6	8.71	8.72	0.01

t1 and t2 represent migration time (min) before and after normalization based on the cyto.c peak.

3.4 Conclusions

We demonstrated an improved automated cIEF-MS method for the characterization of intact charge variants of mAbs with high resolution and reproducibility. First, improving the quality of LPA-coated capillaries via employing a 60-min reaction time for coating and employing a MS-compatible catholyte containing 10 mM ABC and 15% glycerol (pH ~10) greatly boosted the robustness and reproducibility of the cIEF-MS method by mitigating coating degradation effect. Second, good stability of ESI was achieved by carefully adjusting of the positions of ESI emitter relative to the inlet of the Q-TOF mass spectrometer (4-5 mm away from the MS inlet and 1~2

mm away from the central zone). Third, to gain good separation resolution and MS sensitivity, catholyte length, sample concentration, and ampholyte composition were optimized. The optimized conditions are a 30-cm catholyte plug in a 75-cm capillary, 0.8 mg/mL protein concentration under our current MS condition, a 2% three-ampholyte mixture (pI 3-10, 5-8 and 8-10.5, ratio 1:2:4). Fourth, by comparing different sample preparation methods (sample predilution or sample desalting) with mAb1, we found prediluted sample could be directly used for cIEF-MS analysis. The method can significantly reduce sample loss and improve the detection of low abundant charge variants. Finally, we need to point out that constant capillary clean-up (every three runs) with 4 M urea is necessary for maintaining the reproducibility of the system.

We achieved high-resolution and reproducible characterization of charge variants of mAb1 using the optimized cIEF-MS method. Particularly, cIEF-MS resolved ten charge variants of mAb1. Our cIEF-MS method determined the pIs of charge variants of mAb1 accurately by employing a mixture of pI markers and spiking in cytochrome c for migration time normalization. The results render the cIEF-MS method a powerful tool to the biopharmaceutical community for monitoring and validating the heterogeneity of antibody therapeutics. The improved cIEF-MS method will also be valuable for the top-down proteomics community for top-down MS characterization of proteoforms carrying various PTMs.

While most charge variants are well characterized in our study, it remains challenging to explain some mass shifts in deconvoluted mass spectra solely based on intact masses. Comprehensive characterization of mAb charge variants could be benefited from advanced fragmentation techniques for sequencing mAb and mapping PTMs. For example, recently reported activated ion electron transfer dissociation (AI-ETD) would be a promising method for mAb identification with over 60% sequence coverage in a direct infusion study.[49] Alternatively, integrating intact mass analysis and middle-down or bottom-up strategies for multi-level characterization of charge variants could be effective for a better understanding of charge heterogeneity. The strategy has been successfully applied in a previous study of mAb structural heterogeneity.[50]

3.5 Acknowledgments

The research project is funded by AbbVie. We thank Qunying Zhang (Global NBE, Analytical R&D, AbbVie) for initiating the collaboration on this research, helpful discussion, and manuscript reviewing.

REFERENCES

(1) Breedveld, F.Therapeutic monoclonal antibodies. *The Lancet* **2000**, *355*, 735-740.

(2) Kaplon, H.; Reichert, J. M.Antibodies to watch in 2021. *MAbs* **2021**, *13*, 1860476.

(3) Zahavi, D.; Weiner, L.Monoclonal antibodies in cancer therapy. *Antibodies* **2020**, *9*, 34.

(4) Deb, P.; Molla, M. M. A.; Rahman, K. S.-U.An update to monoclonal antibody as therapeutic option against COVID-19. *Biosafety and Health* **2021**.

(5) Cantoni, S.; Carpenedo, M.; Nichelatti, M.; Sica, L.; Rossini, S.; Milella, M.; Popescu, C.; Cairoli, R.Clinical relevance of antiplatelet antibodies and the hepatic clearance of platelets in patients with immune thrombocytopenia. *Blood* **2016**, *128*, 2183-2185.

(6) Lu, R.-M.; Hwang, Y.-C.; Liu, I.-J.; Lee, C.-C.; Tsai, H.-Z.; Li, H.-J.; Wu, H.-C.Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* **2020**, *27*, 1-30.

(7) Beck, A.; Sanglier-Cianférani, S.; Van Dorsselaer, A.Biosimilar, biobetter, and next generation antibody characterization by mass spectrometry. *Anal. Chem.* **2012**, *84*, 4637-4646.

(8) Felten, C.; Salas-Solano, O.; Michels, D. A.Imaged capillary isoelectric focusing for charge-variant analysis of biopharmaceuticals. *BioProcess Int* **2011**, *9*, 48-53.

(9) Lin, J.; Tan, Q.; Wang, S.A high-resolution capillary isoelectric focusing method for the determination of therapeutic recombinant monoclonal antibody. *J. Sep. Sci.* **2011**, *34*, 1696-1702.

(10) Salas-Solano, O.; Kennel, B.; Park, S. S.; Roby, K.; Sosic, Z.; Boumajny, B.; Free, S.; Reed-Bogan, A.; Michels, D.; McElroy, W.Robustness of i CIEF methodology for the analysis of monoclonal antibodies: an interlaboratory study. *J. Sep. Sci.* **2012**, *35*, 3124-3129.

(11) Moritz, B.; Schnaible, V.; Kiessig, S.; Heyne, A.; Wild, M.; Finkler, C.; Christians, S.; Mueller, K.; Zhang, L.; Furuya, K.Evaluation of capillary zone electrophoresis for charge heterogeneity testing of monoclonal antibodies. *J. Chromatogr. B* **2015**, *983*, 101-110.

(12) He, Y.; Lacher, N. A.; Hou, W.; Wang, Q.; Isele, C.; Starkey, J.; Ruesch, M.Analysis of identity, charge variants, and disulfide isomers of monoclonal antibodies with capillary zone electrophoresis in an uncoated capillary column. *Anal. Chem.* **2010**, *82*, 3222-3230.

(13) Kumar, V.; Leweke, S.; von Lieres, E.; Rathore, A. S.Mechanistic modeling of ion-exchange process chromatography of charge variants of monoclonal antibody products. *J. Chromatogr. A* **2015**, *1426*, 140-153.

(14) Leblanc, Y.; Ramon, C.; Bihoreau, N.; Chevreux, G.Charge variants characterization of a monoclonal antibody by ion exchange chromatography coupled on-line to native mass spectrometry: case study after a long-term storage at+ 5 C. *J. Chromatogr. B* **2017**, *1048*, 130-139.

(15) Duivelshof, B. L.; Fekete, S.; Guillarme, D.; D'Atri, V.A generic workflow for the characterization of therapeutic monoclonal antibodies—application to daratumumab. *Anal. Bioanal. Chem.* **2019**, *411*, 4615-4627.

(16) Tang, Q.; Harrata, A. K.; Lee, C. S.Capillary isoelectric focusing-electrospray mass spectrometry for protein analysis. *Anal. Chem.* **1995**, *67*, 3515-3519.

(17) Tang, Q.; Harrata, A. K.; Lee, C. S.Two-dimensional analysis of recombinant E. coli proteins using capillary isoelectric focusing electrospray ionization mass spectrometry. *Anal. Chem.* **1997**, *69*, 3177-3182.

(18) Jensen, P. K.; Paša-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*, 2076-2084.

(19) Pasa-Tolic, L.; Jensen, P. K.; Anderson, G. A.; Lipton, M. S.; Peden, K. K.; Martinovic, S.; Tolic, N.; Bruce, J. E.; Smith, R. D.High throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.* **1999**, *121*.

(20) Mack, S.; Arnold, D.; Bogdan, G.; Bousse, L.; Danan, L.; Dolnik, V.; Ducusin, M.; Gwerder, E.; Herring, C.; Jensen, M.A novel microchip-based imaged CIEF-MS system for comprehensive characterization and identification of biopharmaceutical charge variants. *Electrophoresis* **2019**, *40*, 3084-3091.

(21) Redman, E. A.; Batz, N. G.; Mellors, J. S.; Ramsey, J. M.Integrated microfluidic capillary electrophoresis-electrospray ionization devices with online MS detection for the separation and characterization of intact monoclonal antibody variants. *Anal. Chem.* **2015**, *87*, 2264-2272.

(22) Carillo, S.; Jakes, C.; Bones, J.In-depth analysis of monoclonal antibodies using microfluidic capillary electrophoresis and native mass spectrometry. *J. Pharm. Biomed. Anal.* **2020**, *185*, 113218.

(23) Sun, Q.; Wang, L.; Li, N.; Shi, L.Characterization and monitoring of charge variants of a recombinant monoclonal antibody using microfluidic capillary electrophoresis-mass spectrometry. *Anal. Biochem.* **2021**, 114214.

(24) Cao, L.; Fabry, D.; Lan, K.Rapid and comprehensive monoclonal antibody Characterization using microfluidic CE-MS. *J. Pharm. Biomed. Anal.* **2021**, *204*, 114251.

(25) Redman, E. A.; Mellors, J. S.; Starkey, J. A.; Ramsey, J. M.Characterization of intact antibody drug conjugate variants using microfluidic capillary electrophoresis–mass spectrometry. *Anal. Chem.* **2016**, *88*, 2220-2226.

(26) Maxwell, E. J.; Zhong, X.; Zhang, H.; van Zeijl, N.; Chen, D. D.Decoupling CE and ESI for a more robust interface with MS. *Electrophoresis* **2010**, *31*, 1130-1137.

(27) Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J.Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2554-2560.

(28) 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*, 2312-2321.

(29) Mokaddem, M.; Gareil, P.; Varenne, A.Online CIEF-ESI-MS in glycerol–water media with a view to hydrophobic protein applications. *Electrophoresis* **2009**, *30*, 4040-4048.

(30) Zhong, X.; Maxwell, E. J.; Ratnayake, C.; Mack, S.; Chen, D. D.Flow-through microvial facilitating interface of capillary isoelectric focusing and electrospray ionization mass spectrometry. *Anal. Chem.* **2011**, *83*, 8748-8755.

(31) Zhu, G.; Sun, L.; Dovichi, N. J.Simplified capillary isoelectric focusing with chemical mobilization for intact protein analysis. *J. Sep. Sci.* **2017**, *40*, 948-953.

(32) Chen, J.; Balgley, B. M.; DeVoe, D. L.; Lee, C. S.Capillary isoelectric focusing-based multidimensional concentration/separation platform for proteome analysis. *Anal. Chem.* **2003**, *75*, 3145-3152.

(33) Zhou, F.; Johnston, M. V.Protein characterization by on-line capillary isoelectric focusing, reversed-phase liquid chromatography, and mass spectrometry. *Anal. Chem.* **2004**, *76*, 2734-2740.

(34) Montealegre, C.; Neusüß, C.Coupling imaged capillary isoelectric focusing with mass spectrometry using a nanoliter valve. *Electrophoresis* **2018**, *39*, 1151-1154.

(35) Hühner, J.; Lämmerhofer, M.; Neusüß, C.Capillary isoelectric focusing-mass spectrometry: Coupling strategies and applications. *Electrophoresis* **2015**, *36*, 2670-2686.

(36) Lamoree, M.; Tjaden, U.; Van der Greef, J.Use of microdialysis for the on-line coupling of capillary isoelectric focusing with electrospray mass spectrometry. *J. Chromatogr. A* **1997**, *777*, 31-39.

(37) Liu, R.; Cheddah, S.; Liu, S.; Liu, Y.; Wang, Y.; Yan, C.A porous layer open-tubular capillary column with immobilized pH gradient (PLOT-IPG) for isoelectric focusing of amino acids and proteins. *Anal. Chim. Acta* **2019**, *1048*, 204-211.

(38) Dai, J.; Lamp, J.; Xia, Q.; Zhang, Y.Capillary isoelectric focusing-mass spectrometry method for the separation and online characterization of intact monoclonal antibody charge variants. *Anal. Chem.* **2018**, *90*, 2246-2254.

(39) Wang, L.; Bo, T.; Zhang, Z.; Wang, G.; Tong, W.; Da Yong Chen, D.High resolution capillary isoelectric focusing mass spectrometry analysis of peptides, proteins, and monoclonal antibodies with a flow-through microvial interface. *Anal. Chem.* **2018**, *90*, 9495-9503.

(40) Wang, L.; Chen, D. D. Y.Analysis of four therapeutic monoclonal antibodies by online capillary isoelectric focusing directly coupled to quadrupole time-of-flight mass spectrometry. *Electrophoresis* **2019**, *40*, 2899-2907.

(41) Xu, T.; Shen, X.; Yang, Z.; Chen, D.; Lubeckyj, R. A.; McCool, E. N.; Sun, L.Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative and Quantitative Top-Down Proteomics. *Anal. Chem.* **2020**.

(42) Ramsay, L. M.; Cermak, N.; Dada, O. O.; Dovichi, N. J.Capillary isoelectric focusing with pH 9.7 cathode for the analysis of gastric biopsies. *Anal. Bioanal. Chem.* **2011**, *400*, 2025-2030.

(43) Chen, D.; Yang, Z.; Shen, X.; Sun, L.Capillary Zone Electrophoresis-Tandem Mass Spectrometry As an Alternative to Liquid Chromatography-Tandem Mass Spectrometry for Top-down Proteomics of Histones. *Anal. Chem.* **2021**, *93*, 4417-4424.

(44) Goyon, A.; Excoffier, M.; Janin-Bussat, M.-C.; Bobaly, B.; Fekete, S.; Guillarme, D.; Beck, A.Determination of isoelectric points and relative charge variants of 23 therapeutic monoclonal antibodies. *J. Chromatogr. B* **2017**, *1065*, 119-128.

(45) Turner, A.; Schiel, J. E.Qualification of NISTmAb charge heterogeneity control assays. *Anal. Bioanal. Chem.* **2018**, *410*, 2079-2093.

(46) Yan, Y.; Liu, A. P.; Wang, S.; Daly, T. J.; Li, N.Ultrasensitive characterization of charge heterogeneity of therapeutic monoclonal antibodies using strong cation exchange chromatography coupled to native mass spectrometry. *Anal. Chem.* **2018**, *90*, 13013-13020.

(47) Bumbaca, D.; Boswell, C. A.; Fielder, P. J.; Khawli, L. A.Physiochemical and biochemical factors influencing the pharmacokinetics of antibody therapeutics. *The AAPS journal* **2012**, *14*, 554-558.

(48) Boswell, C. A.; Tesar, D. B.; Mukhyala, K.; Theil, F.-P.; Fielder, P. J.; Khawli, L. A.Effects of charge on antibody tissue distribution and pharmacokinetics. *Bioconj. Chem.* **2010**, *21*, 2153-2163.

(49) Lodge, J. M.; Schauer, K. L.; Brademan, D. R.; Riley, N. M.; Shishkova, E.; Westphall, M. S.; Coon, J. J.Top-down characterization of an intact monoclonal antibody using activated ion electron transfer dissociation. *Anal. Chem.* **2020**, *92*, 10246-10251.

(50) Zhu, W.; Li, M.; Zhang, J.Integrating Intact Mass Analysis and Middle-Down Mass Spectrometry Approaches to Effectively Characterize Trastuzumab and Adalimumab Structural Heterogeneity. *J. Proteome Res.* **2020**, *20*, 270–278.

CHAPTER 4.* Development of non-denaturing cIEF-MS for ultrahigh-resolution characterization of microheterogeneity of protein complexes

4.1 Introduction

Proteins in cells typically form protein complexes to perform functional processes. Microheterogeneity in protein complexes arising from protein sequence variations, post-translational modifications (PTMs), etc., can result in alteration of their physicochemical properties (e.g., charge) and activities.[1] Therefore, an effective method for characterizing microheterogeneity of protein complexes is highly desired for biological and biopharmaceutical applications.

Native mass spectrometry (nMS) has emerged as a powerful analytical tool for the delineation of protein complexes.[2-6] Due to the heterogeneous nature of protein complexes, coupling online and native liquid-phase separations to nMS is an ideal approach for protein complex analysis. Various liquid-phase separation techniques, such as size exclusion chromatography[7,8], ion-exchange chromatography[9], and capillary zone electrophoresis (CZE)[10-16], have been coupled directly to nMS for extensive characterization of protein complexes in simple to complex samples. However, separation resolution of the liquid chromatography (LC) and CZE techniques for protein complexes still need to be improved. Additionally, for CZE, more effort needs to be made to boost the sample loading capacity, although some progresses have been made recently.[17]

Capillary isoelectric focusing (cIEF) is a powerful electrophoretic technique for proteoform and even protein complex separations based on their isoelectric points (pIs) with extremely high resolution.[18-23] It has a drastically higher sample loading capacity (maximum loading of entire capillary volume) compared to CZE (typical loading of less than 1% of the capillary volume). In addition, cIEF can measure the pIs of proteoforms accurately using pI standards.[24] Because the pIs reflect the surface electrostatic properties of proteoforms/protein complexes and eventually impact their biological activities,[25-28] pI measurements can potentially provide evidence for understanding activities of protein complexes. Further coupling cIEF to MS is attractive for proteoform/protein complex analysis by offering capability of mass detection. Denaturing cIEF-MS methods have been widely employed for top-down characterization of proteoforms, *e.g.,*

---

* This chapter is adapted with permission from *Xu, T.; Han, L.; Sun, L. Anal. Chem. 2022, 94 (27), 9674-9682.*

characterization of monoclonal antibody (mAb) charge variants.[29-34] Only very few reports developed and applied non-denaturing cIEF-MS (ncIEF-MS) for analysis of protein complexes.[18,35] The Smith group reported the first example of ncIEF-MS for the characterization of standard protein complexes over 20 years ago.[18] The ncIEF-MS work was carried out in semi-online manner, which required manual operations, including transferring the capillary outlet to CE-MS interface after offline cIEF focusing and then lifting up capillary inlet for protein mobilization. Later, the Daniel group applied "sandwich" injection configuration to ncIEF-MS and was capable of facilitating online protein focusing.[35] They characterized the cytokine human interferon-gamma (IFN-γ) and its homodimer using ncIEF-MS and determined the pI of the dimerized IFN-γ as 9.95.[35] However, the ncIEF-MS remains challenging on both sensitivity and resolution aspects. First, coaxial sheath-flow interface was used for coupling ncIEF and MS with a high sheath liquid flow rate (1~10 μL/min), leading to severe sample dilution before MS measurement. Moreover, using of pressure-driven hydrodynamic mobilization can sacrifice separation resolution by inducing significant peak broadening. Furthermore, separation parameters still need to be improved to better characterize the microheterogeneity of protein complexes. The Yates group presented high resolving power of ncIEF for differentiating several phosphorylation states of Dam1 complex.[19] Unfortunately, their study was only based on UV detection. Lack of nMS characterization greatly limited the identification of variants in the protein complexes.

Here, we advanced ncIEF-MS on sensitivity and resolution for discovering microheterogeneity (sequence variations, PTMs, conformational variations, and cofactor binding) of protein complexes (up to 150 kDa). We coupled ncIEF to MS with a new generation of electro-kinetically pumped sheath flow CE-MS interface (EMASS-II, ~100 nL/min sheath liquid flow rate)[36,37] for intact protein complex characterization with high sensitivity. Meanwhile, chemical mobilization was employed as a major mobilization method for largely maintaining separation resolution. To further improve the performance of ncIEF-MS, we systematically investigated electrospray ionization (ESI) and MS conditions for boosting sensitivity and optimized ncIEF conditions (such as ampholyte composition and concentration) for better separation resolution.

Using our novel ncIEF-MS technique, for the first time, we delineate the microheterogeneity of streptavidin homotetramer and an interchain cysteine-linked antibody-drug conjugate (ADC1). Particularly, ADCs represent promising antitumor agents to be used as one of the tools in personalized cancer medicine.[38] Cysteine conjugated ADC achieved by controlled partial reduction of mAb interchain disulfide bonds followed by drug linker conjugation is one common form of ADCs, which can be considered as ~150 kDa 'protein complex'. Denaturing methods do not work for characterizing this kind of ADC in its intact form because the intact ADC

easily falls apart during analysis. Our work offers ncIEF-MS as a promising technique for monitoring variations in the interchain cysteine-linked ADCs. Besides, we compared results between ncIEF-MS and conventional nCZE-MS, which provided the research community with a better understanding of the features of the two techniques for protein complex analysis. Furthermore, we determined the pIs of those protein complexes with high accuracy and studied how protein sequence variations/PTMs modulate pIs of protein complexes and how drug loading affects the pI of ADC1.

4.2 Experimental section

4.2.1 Materials

Ammonium bicarbonate (NH4HCO3), ammonium acetate (NH4Ac), Pharmalytes (wide pI range of 3-10, norrow pI range of 5-8 and 8-10.5, GE healthcare), Amicon Ultra centrifugal filter units (0.5 mL, 10 kDa cut-off size), cytochrome c (Cyt c), myoglobin (Myo), carbonic anhydrase II (CA) and recombinant streptavidin (SA) were purchased from Sigma-Aldrich (St. Louis, MO). Water (LC-MS grade), acetic acid (AA, LC-MS grade), formic acid (FA, LC-MS grade), acetonitrile (ACN, LC-grade), fused silica capillaries (50 μm i.d./360 μm o.d., Polymicro Technologies) were bought from Fisher Scientific (Pittsburgh, PA). An IEF marker kit containing five peptide pI markers (pI values of 4.1, 5.5, 7.0, 9.8, 10.0) was ordered from Beckman Coulter (Brea, CA). ADC 1 was provided by AbbVie (North Chicago, IL)

4.2.2 Sample preparation

For cIEF-MS analysis, standard protein complexes [cytochrome c (Cyt c), myoglobin (Myo), carbonic anhydrase II (CA) and recombinant streptavidin (SA)]-ampholyte mixtures, ADC1-ampholyte mixtures, and IEF marker-ampholyte mixtures were prepared in a buffer consisting of 10 mM NH4Ac and 15% glycerol (pH 6.7). For pI determination of SA and ADC1, Cyt c was incorporated into the sample-ampholyte mixtures at a final concentration of 0.1 mg/mL as an internal standard for calibration of migration time. Other reagents for cIEF-MS analysis were prepared, including anolyte [0.1% (v/v) formic acid (FA), 15% glycerol, pH 3.0], catholyte (10 mM NH_4HCO_3, 15% glycerol, pH 10.0), and sheath buffer [10 mM NH_4Ac, 10% (v/v) acetonitrile (ACN), pH 5.0].

For CZE-MS analysis, a standard protein complex mixture and ADC1 sample were prepared in 10 mM NH4Ac (pH 6.7). The background electrolyte was 25 mM NH4Ac (pH 6.7).

4.2.3 ncIEF-MS analysis

The ncIEF-MS analysis was performed on an Agilent 7100 CE system coupled with a 6545XT Q-TOF mass spectrometer via an EMASS-II CE-MS interface (CMP Scientific).[36,37] The

ESI emitter orifice size was 25-35 μm. For stable electrospray, the emitter orifice was positioned at a distance of 4~5 mm to the inlet of the mass spectrometer, and the electrospray voltage was 2.3~2.5 kV. The sheath buffer flow rate in the ESI emitter was about 100 nL/min. A linear polyacrylamide (LPA)-coated capillary was used for cIEF separation.

The automated cIEF separation includes the following steps. First, “sandwich” sample injection method was applied by filling an LPA-coated capillary (75 cm long, 50 μm i.d., 360 μm o.d.) with a plug catholyte (pH~10, 30 cm) followed by a plug of sample-ampholyte mixture (pH~7, 45 cm).[33] Afterwards, the capillary inlet was inserted into the anolyte (pH~3). When a voltage (20 kV) was applied across the capillary, mobilization of protons and hydroxide ions can form a pH gradient in the range of 3 to 10. Meanwhile, protein complexes migrate and focus at different positions in the capillary where pHs are equal to their pIs. Finally, when the pH gradient was gradually disrupted by the protons from anolyte and anions from sheath buffer, the focused protein complexes can be charged and migrate towards MS for detection, which is known as chemical mobilization process. In our experiment, a pressure of 10 mbar was applied after 20 min to assist protein mobilization. Motivated by previous studies,[18,19,35] we expect that protein complexes can be preserved during the cIEF separation.

For Q-TOF mass spectrometer, the flow rate and temperature of drying gas were set to 1L/min and 365 °C. The voltage of VCap was 0V. The voltage of collision energy was 10 V. For analysis of protein complexes, the voltages of the skimmer and fragmentor were set to 150 V and 200 V, and the m/z range was 1800-5000. For ADC1 analysis, the voltages of skimmer and fragmentor were set to 300 V and 380 V, and the m/z range was 2000-10000. The acquisition rate (full scan) for all sample analyses was 0.5 spectrum/sec.

4.2.4 nCZE-MS analysis

Native CZE-MS analysis was performed on the same CE-MS platform used for cIEF-MS analysis. The sheath buffer contained 10 mM NH_4Ac and 10% ACN (pH 5.0). The background electrolyte (BGE) was 25 mM NH4Ac. 30 nL sample was loaded by pressure (100 mbar, 14s) into an LPA-coated capillary (75 cm long). A voltage of 30 Kv was applied for CZE separation. Meanwhile, a pressure of 50 mbar was applied to assist protein mobilization. Parameters for the mass spectrometer were consistent with cIEF-MS analysis.

4.3 Results and discussions

4.3.1 Optimization of nMS conditions for ncIEF-MS

We first studied the impact of MS parameters, sheath buffer, and sample buffer on protein signal and native state of protein complexes via direct infusion of a myoglobin solution (0.5

mg/mL), Figure 4.1. We found that MS parameters with 200 V fragmentor and 150 V skimmer, sample buffer consisting of 10 mM NH_4Ac and 15% glycerol, and sheath buffer comprising 10 mM NH4Ac and 10% ACN (pH 5) produced best protein signals without impacting the integrity of non-covalent heme-apomyoglobin complex, which were considered as optimal conditions for ncIEF-MS analysis of standard protein complex mixtures (12 kDa~53 kDa). However, for ADC1 (150 kDa), MS parameters needed to be adjusted (380 V fragmentor, 300 V skimmer, additional 10 V CID energy applied) to enhance transmission of large protein complex.

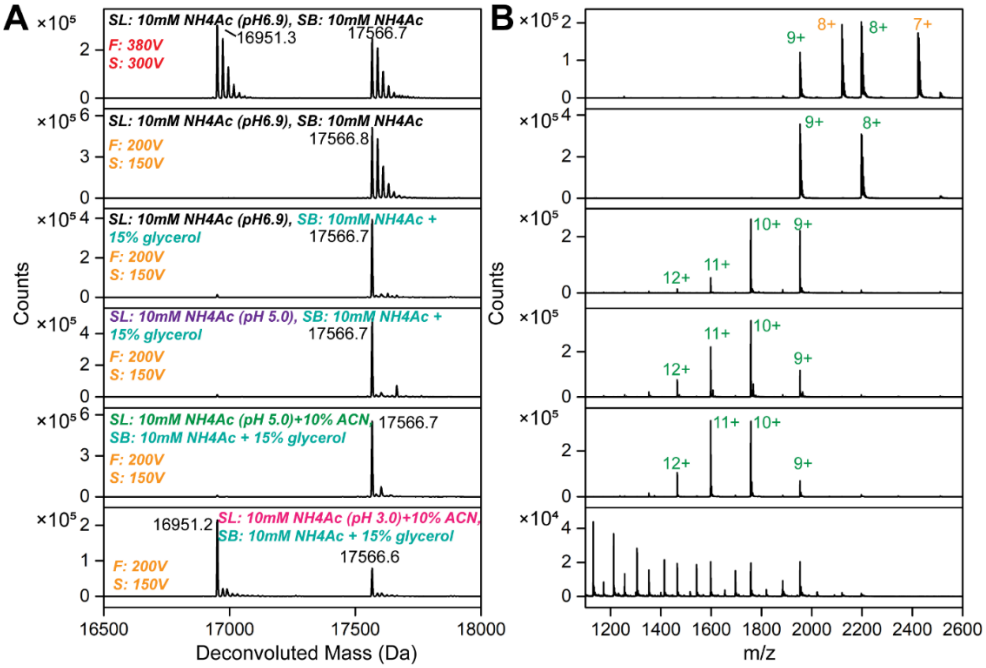

Figure 4.1 Direct infusion of Myo (0.5 mg/mL) for investigating the influence of sheath liquid (SL), sample buffer (SB), and MS settings (voltages of fragmentor (F) and skimmer (S)) on the signal and native state of the Myo. Deconvoluted mass spectra (A) and the corresponding averaged mass spectra (B) of Myo under different conditions.

4.3.2 High-resolution characterization of standard protein complexes using ncIEF-MS

We tested the general performance of ncIEF-MS platform across a wide pI range using a standard protein complex mixture containing cytochrome c (Cyt c, 0.05 mg/mL, pI~10), myoglobin (Myo, 0.2 mg/mL, pI~7), carbonic anhydrase II (CA, 0.4 mg/mL, pI~5), and recombinant streptavidin (SA, 1 mg/mL, pI~7) with 1.5% (v/v) ampholyte (pI range of 3-10). Cyt c and CA were well separated from Myo and SA; SA and Myo were not baseline separated due to their very close pIs, Figure 4.2A. We observed the intact forms of non-covalent protein complexes in the sample, including the holomyoglobin, CA-Zn(II) complex, and homotetramer of SA. The data suggests that the ncIEF-MS condition is capable of maintaining intact protein complexes. Interestingly, we detected two Cyt c peaks (Cyt c 1 and Cyt c 2) with an intensity ratio of roughly 10:1, Figure 4.2A. Both Cyt c variants have the same deconvoluted mass (12231 Da) and similar charge state distribution profiles, which most likely correspond to conformational isomers of holocytochrome c (covalent Cytc-heme complex). In addition, we found CA contained various protein complexes with variations on PTMs, with the main peak and three additional low-abundance variant peaks being characterized by ncIEF-MS, Figure 4.3. The main peak (CA2) was a Zn (II) complex (29086.7 Da), whereas the basic peak [CA1 (29068.8 Da)] showed a -18- Da mass shift relative to the main Zn (II) complex, representing most likely a succinimide formation from aspartic acid on the Zn (II) complex. The two acidic peaks [CA3 (29087.8 Da and CA4 (29058.3 Da)] showed +1-Da and -28-Da mass shifts in comparison with the main Zn (II) complex in CA2, corresponding to a deamidated form of CA2 and a CA variant with an unknown PTM or sequence variation. The data highlights the power of our ncIEF-MS technique for the characterization of microheterogeneity of protein complexes with high separation resolution.

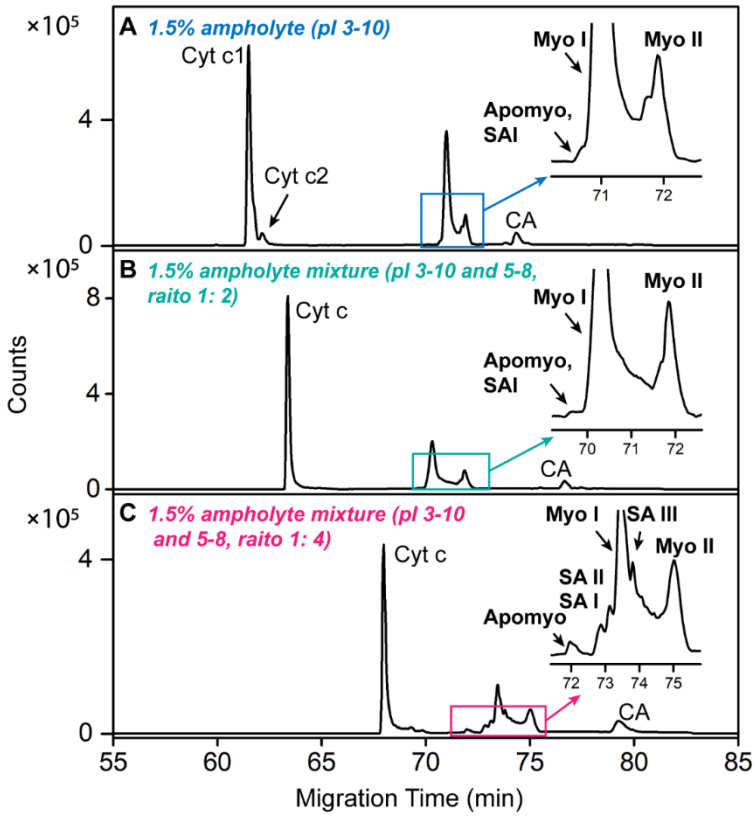

Figure 4.2 ncIEF-MS analysis of a standard protein complex mixture with different ampholyte compositions, including (A) 1.5% (v/v) ampholyte (pI 3-10), (B) 1.5% (v/v) ampholyte (pI 3-10 and 8-10.5 with a ratio of 1:2) and (C) 1.5% (v/v) ampholyte (pI range of 3-10 and 8-10.5 with a ratio of 1:4).

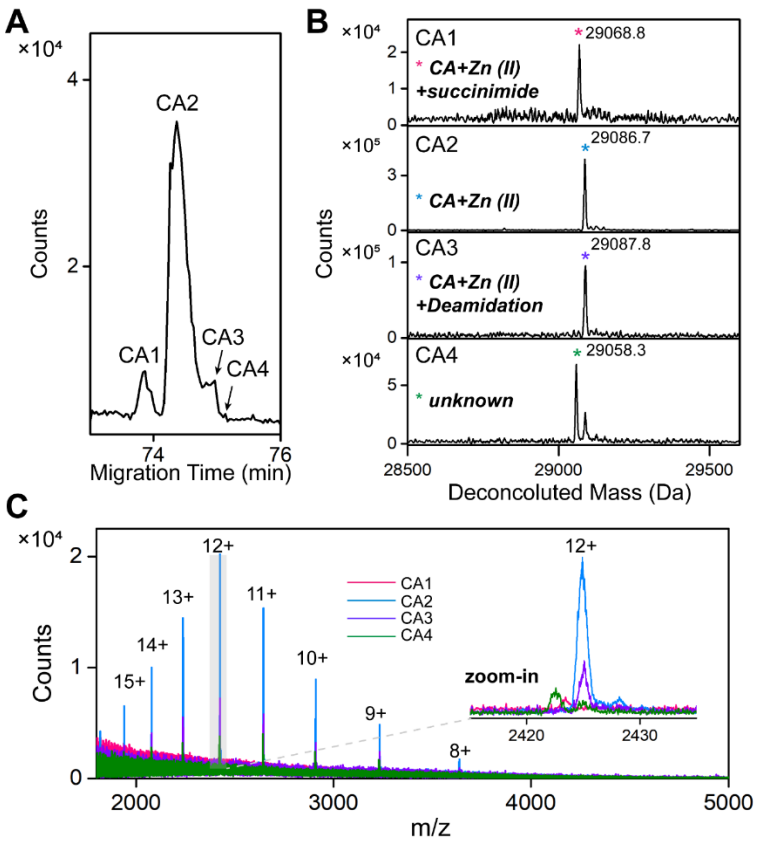

Figure 4.3 ncIEF-MS result of carbonic anhydrase (CA). (A) Base peak electropherogram of CA from ncIEF-MS analysis. Four variant peaks of CA were detected. (B, C) Deconvoluted mass spectra (B) and averaged mass spectra (C) of the four CA variants. The inserted figure in (C) represents the zoomed-in spectra of CA charge variants with a charge state of 12+.

To achieve better separation of Myo and SA, we added a narrow pI range ampholyte (pI 5-8) to the sample-ampholyte (pI 3-10) mixture and maintained the total concentration of ampholyte as 1.5%. For the two-ampholyte mixture (1.5% v/v, pI 3-10 and 5-8), we varied the concentration ratio between the wide (pI 3-10) and narrow range ampholyte (pI 5-8) from 1:2 to 1:4. We improved the separation resolution of SA and Myo slightly with the 1:2 ampholyte mixture and boosted the resolution obviously with the 1:4 ampholyte mixture, Figures 4.2B and 4.2C. For example, we observed three peaks of SA (SA I, II, and III) and three peaks of Myo (Apomyo, Myo I, and Myo II) using the 1:4 ampholyte mixture. The improved separation resolution is due to the fact that the narrow pI range ampholyte (pI 5-8) stabilizes the pH gradient in the 5-8 range. The more stable pH gradient in the two-ampholyte mixture conditions also results in a longer migration time of protein complexes because hydrogen protons need a longer time to titrate the catholyte zone and charge the separated analytes for chemical mobilization, Figures 4.2B and 4.2C.

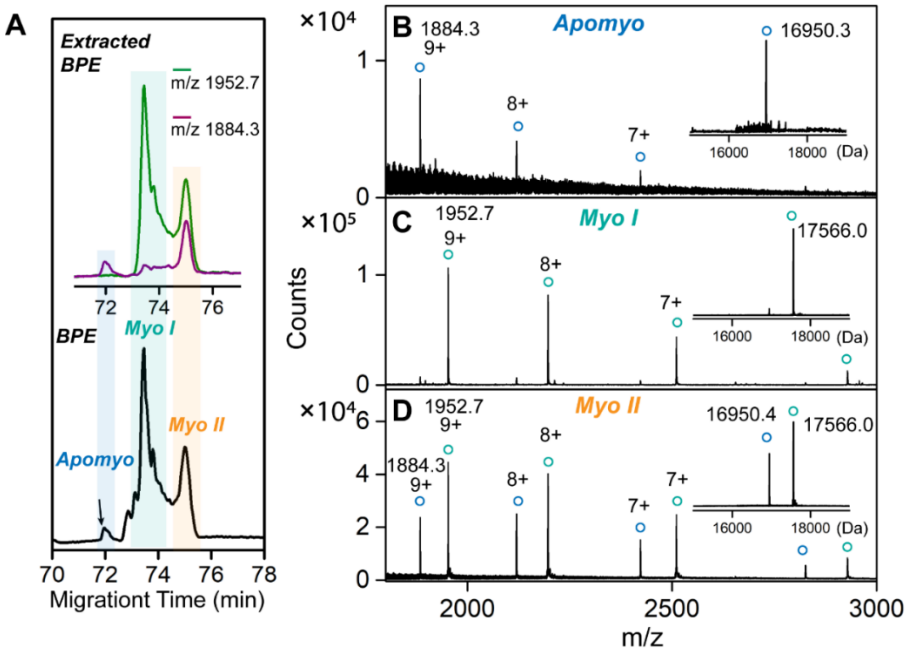

Figure 4.4 ncIEF-MS result of myoglobin (Myo). (A) Base peak electropherogram (BPE) and extracted base peak electropherogram [extracted ion at m/z 1952.7 (contains heme) and 1884. 3 (without heme)] of myoglobin sample from ncIEF-MS analysis. (B-D) Averaged mass spectra and deconvoluted mass spectra (the inserted figures) of three myoglobin variants including apomyoglobin (B), Myo I (C), and Myo II (D) from ncIEF-MS analysis.

As shown in Figures 4.2C and 4.4, we achieved reasonable separations of three Myo peaks, corresponding to Apomyo, Myo I, and Myo II. Myo I (17566.0 Da) was the most abundant and identified as a holomyoglobin containing a non-covalently binding heme group, Figure 4.4C. Compared to Myo I, Myo II appeared to be more acidic and contained both species with heme (17566.0 Da, ~65%, Figure 4.4A) and without heme group (16950.4Da, ~35%, Figure 4.4A), which could be resulted from a conformation isomer/intermediate state of holomyoglobin that is vulnerable to lose heme during ESI, Figure 4.4D. Another low-abundance basic peak (Apomyo) separated by cIEF turned out to be an apomyoglobin (16950.3 Da), Figure 4.4B. The two histidines on Myo, His64 and His93, were known as heme-binding sites.[39] The apomyogbin with free His64 and His93 turned out to be more basic compared to the Myo I. The Apomyo in Figure 4.4B tended to be in a slightly higher charge state (average charge state of 8.4+) compared to the myoglobin losing heme during the ESI process (average charge state of 7.9+) in Figure 4.4D, which suggests a more unfolding structure of Apomyo compared to Myo II.

We also detected three peaks of SA (SA1, SA2, SA3) in Figure 4.2C, corresponding to SA homotetramers with variations in N-terminal methionine removal. It was difficult to sufficiently separate those SA charge variants from the Myo I and Myo II due to their high similarity in pIs. To achieve a better understanding of SA's microheterogeneity, we performed another ncIEF-MS experiment on the SA standard protein complex. Surprisingly, ncIEF-MS resolved seven variants of SA homotetramer, Figure 4.5. SA1, SA2, SA3, SA4 correspond to SA tetramer with remaining N-terminal methionine on zero, one, two, and three monomers, Figure 4.5B. The SA homotetramers preserving more N-terminal methionine had lower pI values. There are two possible reasons for this phenomenon. First, methionine is an acidic amino acid with a pI of ~5.7. Second, the existence of N-terminal methionine changed the structure of SA slightly, leading to minor changes in the surface electrostatic property of SA. Moreover, we detected another three charge variants of SA tetramer (SA5, SA6, and SA7), which contain additional acetylation (Δm~+42Da) PTM on SA1, SA2, and SA3. Interestingly, we detected additional SA variants in SA 5-7 peaks with formylation (Δm ~+28Da) on SA2, SA3, SA4. While missing of formylated SA1, we suspect the formylation is mostly likely to occur on the N-terminal methionine. The formylated methionine on recombinant SA was reported in the literature.[40] Based on our knowledge, this is the first time that the microheterogeneity of SA homotetramer is characterized by such high separation resolution, although SA has been a commonly used standard protein complex in nMS studies. The data provide additional strong evidence about the capability of ncIEF-MS for the delineation of protein complexes.

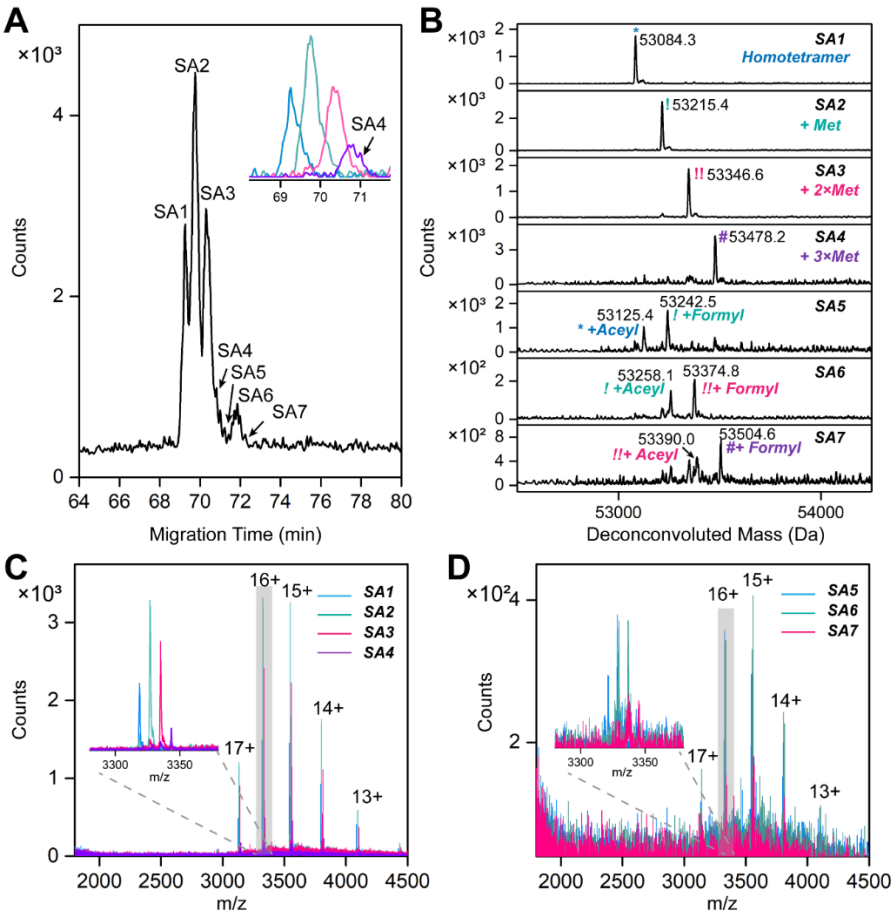

Figure 4.5 ncIEF-MS result of streptavidin (SA). (A) Base peak electropherogram of SA analyzed by ncIEF-MS. (B) Deconvoluted mass spectra of seven SA charge variants separated from ncIEF-MS. (C) Averaged mass spectra of SA1, SA2, SA3 and SA4. (D) Averaged mass spectra of SA5, SA6 and SA7. The mass spectra of SA charge variants were overlaid in (C) and (D) for comparing the subtle difference of peaks at the same charge state. The inserted figures represent the zoomed-in peaks of SA charge variants at a charge state of 16+.

Native CZE-MS (nCZE-MS) has also been reported to characterize protein complexes by our group and several other research groups.[11-16] Here, we compared nCZE-MS and ncIEF-MS for analysis of the protein mixture containing Cyt c, Myo, SA and CA (Figure 4.2C vs Figure 4.6). Overall, ncIEF-MS has a better resolution for delineating the microheterogeneity of protein complexes compared to nCZE-MS. Besides that, ncIEF-MS can provide accurate information on pIs of protein complexes, offering valuable information about their surface electrostatic properties. This point will be well demonstrated in the last part of the “Results and discussions”. We noted that nCZE-MS also has its advantages compared to ncIEF-MS. First, nCZE-MS produced a comparable signal of protein complexes to ncIEF-MS with 30-times lower sample consumption,

most likely due to the ionization suppression of ampholyte in ncIEF-MS. Second, nCZE-MS provides fast analysis compared to ncIEF-MS (30 min vs. over 1 hour). Third, nCZE and ncIEF separate protein complexes according to different principles (size-to-charge ratios vs. pIs). Protein complexes having very close pIs (e.g., Myo and SA) could not be fully resolved by ncIEF but could be separated by nCZE. We expect the combination of nCZE-MS and ncIEF-MS will be important for native proteomics.

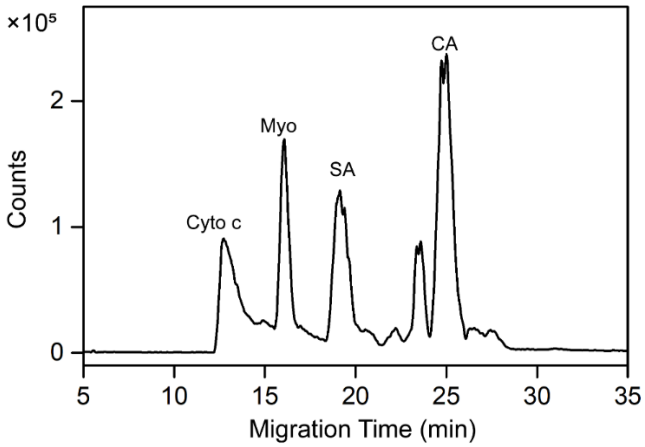

Figure 4.6 Base peak electropherogram of the standard protein mixture from nCZE-MS analysis.

4.3.3 High-resolution characterization of interchain cysteine-linked ADC1 using ncIEF-MS

After validating the ncIEF-MS system with the standard protein complex mixture containing relatively small protein complexes (<60 kDa), we aim to further test it for characterizing the microheterogeneity of a much larger protein complex. We chose one interchain cysteine-linked ADC (ADC1, pI~9.1, ~150 kDa) produced by AbbVie for this purpose. Interchain cysteine-linked ADCs are a type of biotherapeutics for treating cancers.[41] They are manufactured by breaking interchain disulfides of tumor-specific monoclonal antibodies (mAbs) and conjugating drugs/payloads to the free cysteines via linkers.[42] With this process, ADC1 was produced with a broad distribution of varied drug-to-antibody ratio (DAR) species (DAR0, DAR2, DAR4, DAR6). Critical quality attributes (CQAs) of ADCs, such as PTMs, size variants, charge variants and DAR distribution are highly concerned because of their potential impacts on pharmacokinetics, bioactivity, and toxicity.[43] For interchain cysteine-linked ADCs, cIEF separation and MS characterization need to be performed under the non-denaturing condition to maintain the integrity of the ADC structure. Here, we validated our ncIEF-MS methods to characterize CQAs in the ADC1.

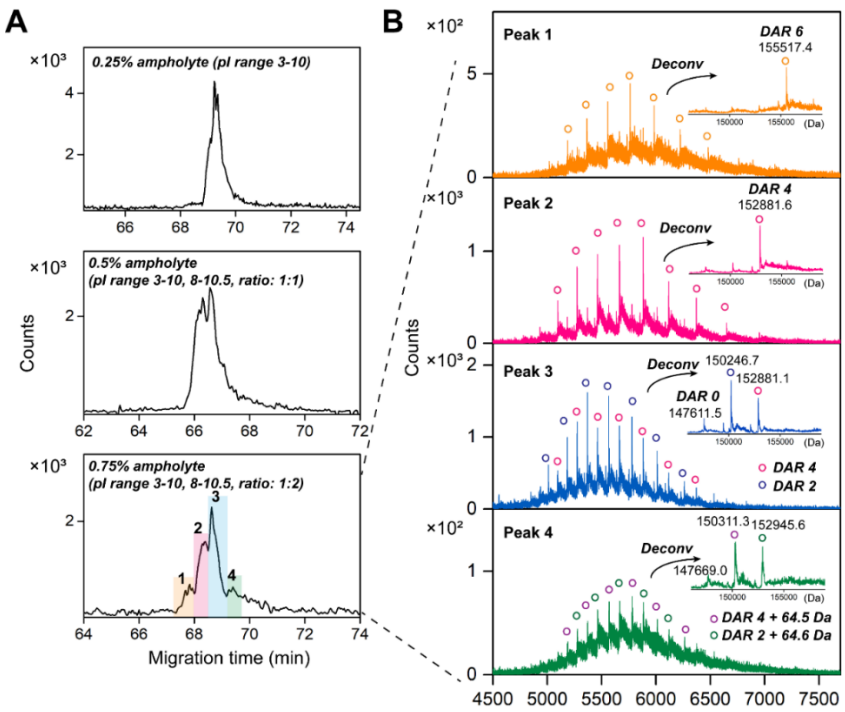

Figure 4.7 Investigation of the influence of incorporating a narrow range ampholyte on separation of ADC1 in ncIEF-MS. (A) Base peak electropherograms of ADC1 with 0.25% (v/v) ampholyte (pI 3-10), 0.5% (v/v) ampholyte (pI 3-10 and 8-10.5 with a ratio of 1:1) and 0.75% (v/v) ampholyte (pI range of 3-10 and 8-10.5 with a ratio of 1:2). ADC1 (3 mg/mL) was used for the experiment to guarantee protein signal. (B) Averaged mass spectra and deconvoluted mass spectra (the inserted figures) of DAR variants (peak 1 to peak 3) and an acidic variant (peak 4) of ADC1 with 0.75% (v/v) ampholyte (pI 3-10 and 8-10.5 with a ratio of 1:2).

As we learned from the standard protein complex mixture, the separation resolution of ADC1 could also be improved by incorporating a narrow pI range ampholyte (pI 8-10.5) into the pI 3-10 ampholyte. As the addition of a narrow pI range ampholyte can raise the concentration of ampholyte in specific region, 3 mg/mL ADC was used to achieve a reasonable protein signal. As shown in Figure 4.7, adding pI 8-10.5 ampholyte into pI 3-10 ampholyte (0.75% (v/v), pI range of 3-10 and 8-10.5 with a ratio of 1:2) improved the separation of different species of ADC1 significantly, resulting in four resolved peaks of ADC1. Deconvoluted mass spectra indicated DAR 6, DAR 4, and a mixture of DAR 4, DAR 2, and DAR 0 in peaks 1, 2 and 3, respectively. Our results also suggest that the high DAR species are more basic than the low DAR species, which could be because the cysteine reduction and drug linker conjugation perturb the local conformation of ADC1, exposing the charged residuals to some extent. Although the DAR 4 was not fully separated from DAR 2 and DAR0 in our study, the coeluted DAR 4 in peak 3 [average

charge state of (27.6 ± 0.0)+] showed slightly higher charge states than the DAR 4 in peak 2 [average charge state of (27.3 ± 0.1)+], Figure 4.7. The predominant charge state shifted to a higher value for DAR4 in peak 3 compared to that in peak 2. We suspect that there is subtle structure heterogeneity in DAR 4 arising from variations in drug binding sites (positional isomers). Optimization of the ampholyte generated a wider separation window, potentially bettering the separation of positional isomers. We also performed nCZE-MS analysis of the ADC1 and could not resolve the different DAR species using nCZE separation, Figure 4.8. The DAR variants separated in ncIEF co-migrated in nCZE, further highlighting the better separation resolution of ncIEF-MS for characterizing the microheterogeneity of protein complexes.

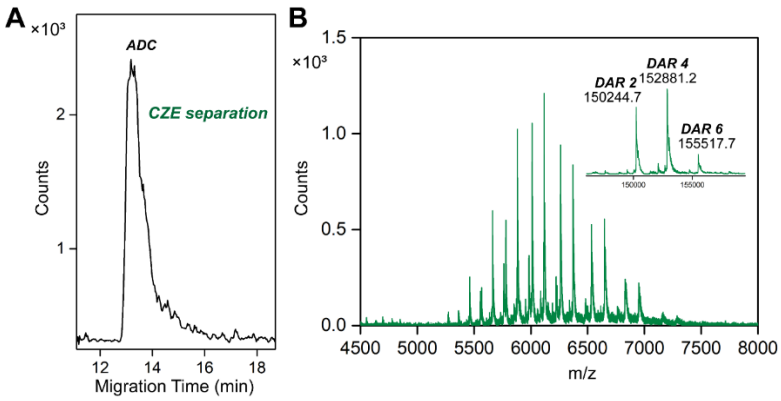

Figure 4.8 CZE-MS result of ADC1. (A) Extracted ion electropherograms (base peak at m/z of 4000-6500) of ADC1 from nCZE-MS analysis. (B) Averaged mass spectrum and deconvoluted mass spectrum (the inserted figure) of ADC1.

4.3.4 Accurate determination of pIs of SA and ADC1 using ncIEF-MS

The pIs of protein complexes could reflect their surface electrostatic properties which are closely associated with protein complex activities.[44,45] cIEF is a powerful tool for determining pIs of proteins via using a mixture of pI markers.[24] Here, we employed ncIEF-MS for accurate determination of pIs of SA and ADC1 via using a mixture of five peptide pI markers (pI 4.05, 5.52, 7.00, 9.50, 9.99). We first investigated the correlation between migration time and pI via ncIEF-MS analysis of the pI markers. A linear correlation (R^2=0.99) was observed using the pI 3-10 ampholyte, Figure 4.9A. For SA tetramer variants, their relative standard deviations (RSDs) of absolute migration time in triplicate runs were around 3%. We normalized their migration time across runs based on the cyt c peak and reduced the migration time RSDs to 0.2%. The pIs of five SA tetramer variants (SA1-5) were determined with high precision based on their normalized migration time and the calibration curve. The pIs of SA tetramer variants range from 6.9-7.4 (Table

4-1), agreeing well with the pI information provided by the manufacturer (pI 6.8~7.5), suggesting that our ncIEF-MS technique can provide an accurate pI determination of SA tetramers.

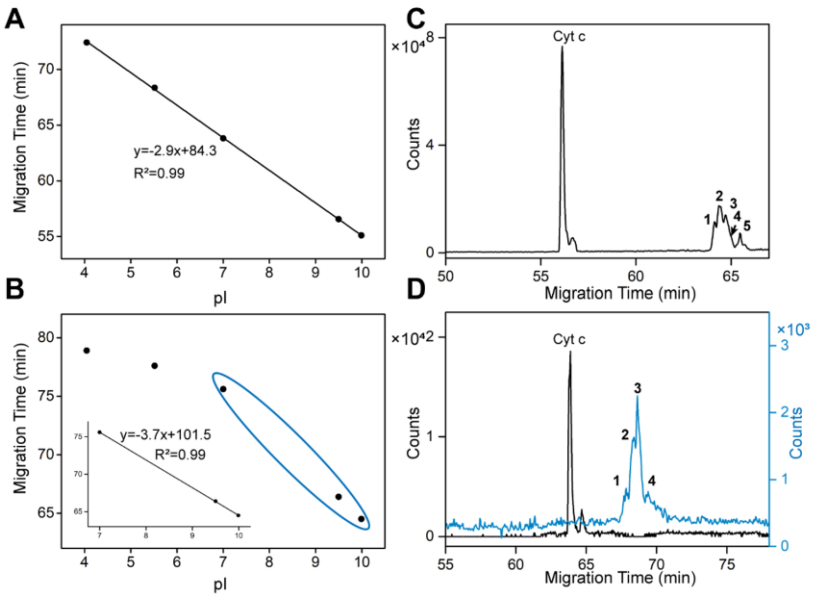

Figure 4.9 pI calculation of SA homotetramer variants and ADC1 variants. Linear regression of pI values of five pI markers (pI 9.99, m/z 624.3; pI 9.50, m/z 950.4; pI 7.00, m/z 627.3; pI 5.52, m/z 471.2; pI 4.05, m/z 591.2) versus mobilization times with (A)1.5% (v/v) ampholyte (pI 3-10) and (B) 0.75% (v/v) ampholyte (pI 3-10 and 8-10.5 with a ratio of 1:2). Addition of narrow range ampholyte led to distortion of acidic pH gradient in (B). However, excellent linear correlation (R^2=0.99) was well maintained in the basic region (pI 7-10) and was used for pI determination of the ADC1 (theoretical pI~9.06).

The data here provide us with a unique opportunity for studying how sequence variations and PTMs of SA affect its pI. From the pIs of SA 1-4, we discovered that keeping the N-terminal methionine residue of one subunit of SA tetramer decreased its overall pI by roughly 0.1, which is most likely due to the acidic feature of methionine (pI ~5.7) and the possibility that the existence of N-terminal methionine changed the structure of SA tetramer slightly. By comparing the pIs of SA1 and SA6, we deciphered that adding one acetylation onto the SA homotetramer reduced its pI by nearly 0.4 and adding one formylation decreased the pI by around 0.3, indicating that acetylation has a slightly higher impact on protein complex's pI compared to formylation. The ncIEF-MS technique enabled the direct connection between sequence variations/PTMs and pIs of protein complexes.

Table 4-1 Calculation of pI values of streptavidin charge variants.

SA variants	Normalized migration time (min)			pI			pI
	Run#1	Run#2	Run#3	Run#1	Run#2	Run#3	
SA1	62.82	62.60	62.70	7.36	7.36	7.40	7.37±0.02
SA2	63.09	62.84	62.90	7.26	7.29	7.33	7.29±0.03
SA3	63.38	63.26	63.22	7.16	7.16	7.22	7.18±0.03
SA4	63.67	63.43	63.50	7.07	7.06	7.12	7.08±0.04
SA2+acetyl	64.14	63.73	63.90	6.90	6.94	6.99	6.94±0.04

We further performed a similar experiment for ADC1 (theoretical pI~9.06). pI determination of ADCs and their charge variants is crucial for understanding the pharmacological properties of ADCs.[27] In this case, we employed the mixture of pI 3-10 and 8-10.5 ampholytes with a ratio of 1:2 (0.75% (v/v) in total). The addition of narrow pI range ampholyte (8-10.5) reduced the overall linearity of migration time and pI across a pI range of 4-10, but we still achieved an excellent linear correlation in the basic region (pI 7-10, R^2=0.99), Figure 4.9B. Based on the calibration curve and normalized migration time of resolved ADC1 variants (Peak 1, 2, 3, and 4), we determined their pIs in a range of 8.8~9.2 with nice precision, Table 4-2. The pIs of ADC1 variants in peaks 1-3 suggest that loading two more drug molecules on one ADC1 molecule increased its overall pI by 0.1, potentially due to the conformational heterogeneity from linker payload conjugation. The PTM (Δm~+64Da) on DAR4 and DAR2 reduced their pIs by about 0.2. The data here clearly render ncIEF-MS as a powerful tool for the delineation of ADCs via providing not only high-resolution separation but also an accurate determination of pIs of ADC variants.

Table 4-2 Calculation of pI values of ADC charge variants.

ADC variants	Normalized migration time (min)			pI			pI
	Run#1	Run#2	Run#3	Run#1	Run#2	Run#3	
Peak1	67.35	67.50	67.40	9.24	9.19	9.22	9.22±0.02
Peak2	67.80	67.89	67.80	9.11	9.09	9.11	9.11±0.01
Peak3	68.21	68.23	68.10	9.00	9.00	9.03	9.01±0.02
Peak4	68.85	69.15	68.95	8.83	8.75	8.80	8.79±0.04

4.4 Conclusions

We developed an automated ncIEF-MS technique, enabling characterization with high-resolution separation of microheterogeneity of intact protein complexes and accurate determination of their pIs. The ncIEF-MS demonstrated advantages over nCZE-MS on separation resolution and the capability for determining pIs of protein complexes. Our method disclosed the various microheterogeneity in protein complexes originating from sequence variations, PTMs, conformational variations, and cofactor binding. In particular, we reported seven different forms of SA tetramer containing variations on N-terminal methionine removal, and PTMs including acetylation and formylation. We also documented the partial separations of different DAR species of an interchain cysteine-linked ADC. The ncIEF-MS methodology enabled precise pI measurements of SA tetramers and ADC1 variants via employing peptide pI markers for a calibration curve and using cyt c for migration time normalization across runs.

We expect that some improvements in the mass spectrometer will further boost the capability of our ncIEF-MS for the delineation of protein complexes. For example, coupling ncIEF to a mass spectrometer with a much higher resolution (i.e., Orbitrap extended mass range (EMR)[46,47] will allow high-resolution characterization of mega-dalton protein complexes. In addition, better declustering and transmission for large protein complexes are still needed, although we largely improved the sheath buffer and some MS parameters. Combining the ncIEF and a mass spectrometer with the capability of fragmenting large protein complexes in the gas phase will provide tremendous potential for native proteomics.[15, 48-50] It will be attractive to couple ncIEF with ion mobility MS to study the conformational heterogeneity of protein complexes. We need to point out that the current ncIEF-MS method requires over 1 hour for a single run, and

improvement of its throughput is critical in the near future. The development of minimized CE systems with much shorter separation capillaries will be a potential solution for the throughput issue.

4.5 Acknowledgments

The research project is funded by AbbVie. We thank Qunying Zhang and Alayna M George Thompson (NBE, Analytical R&D, AbbVie) for initiating the collaboration on this research, helpful discussion, and manuscript reviewing.

REFERENCES

(1) Rolland, A. D.; Prell, J. S. Approaches to Heterogeneity in Native Mass Spectrometry. *Chem. Rev.* **2022**. 122, 7909-7951.

(2) Heck, A. J. Native Mass Spectrometry: A Bridge Between Interactomics and Structural Biology. *Nat. Methods* **2008**, 5, 927-933.

(3) Li, H.; Nguyen, H. H.; Loo, R. R. O.; Campuzano, I. D.; Loo, J. A. An Integrated Native Mass Spectrometry and Top-Down Proteomics Method That Connects Sequence to Structure and Function of Macromolecular Complexes. *Nat. Chem.* **2018**, 10, 139-148.

(4) Tamara, S.; den Boer, M. A.; Heck, A. J. High-Resolution Native Mass Spectrometry. *Chem. Rev.* **2021**.

(5) Fantin, S. M.; Parson, K. F.; Yadav, P.; Juliano, B.; Li, G. C.; Sanders, C. R.; Ohi, M. D.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry Reveals the Role of Peripheral Myelin Protein Dimers in Peripheral Neuropathy. *Proc. Natl. Acad. Sci.* **2021**, 118.

(6) Keener, J. E.; Zhang, G.; Marty, M. T. Native Mass Spectrometry of Membrane Proteins. *Anal. Chem.* **2020**, 93, 583-597.

(7) Muneeruddin, K.; Thomas, J. J.; Salinas, P. A.; Kaltashov, I. A. Characterization of Small Protein Aggregates and Oligomers Using Size Exclusion Chromatography with Online Detection by Native Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **2014**, 86, 10692-10699.

(8) Busch, F.; VanAernum, Z. L.; Lai, S. M.; Gopalan, V.; Wysocki, V. H. Analysis of Tagged Proteins Using Tandem Affinity-Buffer Exchange Chromatography Online with Native Mass Spectrometry. *Biochemistry* **2021**.

(9) Muneeruddin, K.; Nazzaro, M.; Kaltashov, I. A. Characterization of Intact Protein Conjugates and Biopharmaceuticals Using Ion-Exchange Chromatography with Online Detection by Native Electrospray Ionization Mass Spectrometry and Top-Down Tandem Mass Spectrometry. *Anal. Chem.* **2015**, 87, 10138-10145.

(10) 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, 7169-7173.

(11) 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, 2614-2634.

(12) Shen, X.; Kou, Q.; Guo, R.; Yang, Z.; Chen, D.; Liu, X.; Hong, H.; Sun, L. Native Proteomics in Discovery Mode Using Size-Exclusion Chromatography-Capillary Zone Electrophoresis-Tandem Mass Spectrometry, *Anal. Chem.* **2018**, 90, 10095-10099.

(13) Jooß, K.; Schachner, L. F.; Watson, R.; Gillespie, Z. B.; Howard, S. A.; Cheek, M. A.; Meiners, M. J.; Sobh, A.; Licht, J. D.; Keogh, M.-C. Separation and Characterization of Endogenous Nucleosomes by Native Capillary Zone Electrophoresis-Top-Down Mass Spectrometry. *Anal. Chem.* **2021**, 93, 5151-5160.

(14) Mehaffey, M. R.; Xia, Q.; Brodbelt, J. S. Uniting Native Capillary Electrophoresis and Multistage Ultraviolet Photodissociation Mass Spectrometry for Online Separation and Characterization of Escherichia coli Ribosomal Proteins and Protein Complexes. *Anal. Chem.* **2020**, 92, 15202-15211.

(15) Shen, Y.; Zhao, X.; Wang, G.; Chen, D. D. Differential Hydrogen/Deuterium Exchange during Proteoform Separation Enables Characterization of Conformational Differences between Coexisting Protein States. *Anal. Chem.* **2019**, 91, 3805-3809.

(16) Jooß, K.; McGee, J. P.; Melani, R. D.; Kelleher, N. L. Standard Procedures for Native CZE-MS of Proteins and Protein Complexes Up To 800 kDa. *Electrophoresis* **2021**, 42, 1050-1059.

(17) Shen, X.; Liang, Z.; Xu, T.; Yang, Z.; Wang, Q.; Chen, D.; Pham, L.; Du, W.; Sun, L. Investigating Native Capillary Zone Electrophoresis-Mass Spectrometry on a High-End Quadrupole-Time-of-Flight Mass Spectrometer for the Characterization of Monoclonal Antibodies. *Int. J. Mass spectrom.* **2021**, 462, 116541.

(18) Martinović, S.; Berger, S. J.; Paša-Tolić, L.; Smith, R. D. Separation and Detection of Intact Noncovalent Protein Complexes from Mixtures by On-Line Capillary Isoelectric Focusing-Mass Spectrometry. *Anal. Chem.* **2000**, 72, 5356-5360.

(19) Fonslow, B. R.; Kang, S. A.; Gestaut, D. R.; Graczyk, B.; Davis, T. N.; Sabatini, D. M.; Yates III, J. R. Native Capillary Isoelectric Focusing for the Separation of Protein Complex Isoforms and Subcomplexes. *Anal. Chem.* **2010**, 82, 6643-6651.

(20) 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, 3235-3241.

(21) Dou, P.; Liu, Z.; He, J.; Xu, J.-J.; Chen, H.-Y. Rapid and High-Resolution Glycoform Profiling of Recombinant Human Erythropoietin by Capillary Isoelectric Focusing with Whole Column Imaging Detection. *J. Chromatogr. A* **2008**, 1190, 372-376.

(22) Xu, T.; Sun, L. A Mini Review on Capillary Isoelectric Focusing-Mass Spectrometry for Top-Down Proteomics. *Front. Chem.* **2021**, 9, 651757.

(23) Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L. Recent Advances (2019–2021) of Capillary Electrophoresis-Mass Spectrometry for Multilevel Proteomics. *Mass Spectrom. Rev.* **2021**. doi: 10.1002/mas.21714.

(24) Righetti, P. G. Determination of the Isoelectric Point of Proteins by Capillary Isoelectric Focusing. *J. Chromatogr. A* **2004**, 1037, 491-499.

(25) Loell, K.; Nanda, V. Marginal Protein Stability Drives Subcellular Proteome Isoelectric Point. *Proc. Natl. Acad. Sci.* **2018**, 115, 11778-11783.

(26) Sivasankar, S.; Subramaniam, S.; Leckband, D. Direct Molecular Level Measurements of the Electrostatic Properties of a Protein Surface. *Proc. Natl. Acad. Sci.* **1998**, 95, 12961-12966.

(27) Bumbaca, D.; Boswell, C. A.; Fielder, P. J.; Khawli, L. A. Physiochemical and Biochemical Factors Influencing the Pharmacokinetics of Antibody Therapeutics. *AAPS J.* **2012**, 14, 554-558.

(28) Nadendla, K.; Friedman, S. H. Light Control of Protein Solubility Through Isoelectric Point Modulation. *J. Am. Chem. Soc.* **2017**, 139, 17861-17869.

(29) Dai, J.; Lamp, J.; Xia, Q.; Zhang, Y. Capillary Isoelectric Focusing-Mass Spectrometry Method for the Separation and Online Characterization of Intact Monoclonal Antibody Charge Variants. *Anal. Chem.* **2018**, 90, 2246-2254.

(30) Wang, L.; Bo, T.; Zhang, Z.; Wang, G.; Tong, W.; Chen, D. High Resolution Capillary Isoelectric Focusing Mass Spectrometry Analysis of Peptides, Proteins, And Monoclonal Antibodies with a Flow-through Microvial Interface. *Anal. Chem.* **2018**, 90, 9495-9503.

(31) Mack, S.; Arnold, D.; Bogdan, G.; Bousse, L.; Danan, L.; Dolnik, V.; Ducusin, M.; Gwerder, E.; Herring, C.; Jensen, M. A Novel Microchip-Based Imaged CIEF-MS System for Comprehensive Characterization and Identification of Biopharmaceutical Charge Variants. *Electrophoresis* **2019**, 40, 3084-3091.

(32) Xu, T.; Shen, X.; Yang, Z.; Chen, D.; Lubeckyj, R. A.; McCool, E. N.; Sun, L. Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative and Quantitative Top-Down Proteomics. *Anal. Chem.* **2020**, 92, 15890-15898.

(33) Xu, T.; Han, L.; George Thompson, A. M.; Sun, L. An improved capillary isoelectric focusing-mass spectrometry method for high-resolution characterization of monoclonal antibody charge variants. *Anal. Methods* **2022**, 14, 383-393.

(34) He, X.; ElNaggar, M.; Ostrowski M. A.; Guttman, A.; Gentalen, E.; Sperry, J. Evaluation of an icIEF-MS system for comparable charge variant analysis of biotherapeutics with rapid peak identification by mass spectrometry. *Electrophoresis* **2022**. doi: 10.1002/elps.202100295.

(35) Przybylski, C.; Mokaddem, M.; Prull-Janssen, M.; Saesen, E.; Lortat-Jacob, H.; Gonnet, F.; Varenne, A.; Daniel, R. On-Line Capillary Isoelectric Focusing Hyphenated to Native Electrospray Ionization Mass Spectrometry for the Characterization of Interferon-Γ And Variants. *Analyst* **2015**, 140, 543-550.

(36) 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, 2312-2321.

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

(38) Chari, R.V.; Miller, M.L.; Widdison, W.C. Antibody–Drug Conjugates: An Emerging Concept in Cancer Therapy. *Angew. Chem. Int. Ed.* **2014**, 53, 3796-3827.

(39) Enyenihi, A. A.; Yang, H.; Ytterberg, A. J.; Lyutvinskiy, Y.; Zubarev, R. A. Heme Binding in Gas-Phase Holo-Myoglobin Cations: Distal Becomes Proximal? *J. Am. Soc. Mass Spectrom.* **2011**, 22.

(40) Wu, S.C.; Wong, S.L. Structure-Guided Design of An Engineered Streptavidin with Reusability to Purify Streptavidin-Binding Peptide Tagged Proteins or Biotinylated Proteins. *PloS one*, **2013**, 8, e69530.

(41) Behrens, C. R.; Ha, E. H.; Chinn, L. L.; Bowers, S.; Probst, G.; Fitch-Bruhns, M.; Monteon, J.; Valdiosera, A.; Bermudez, A.; Liao-Chan, S. Antibody-Drug Conjugates (ADCs) Derived from Interchain Cysteine Cross-Linking Demonstrate Improved Homogeneity and Other Pharmacological Properties over Conventional Heterogeneous ADCs. *Mol. Pharm.* **2015**, 12, 3986-3998.

(42) Larson, E. J.; Roberts, D. S.; Melby, J. A.; Buck, K. M.; Zhu, Y.; Zhou, S.; Han, L.; Zhang, Q.; Ge, Y. High-Throughput Multi-attribute Analysis of Antibody-Drug Conjugates Enabled by Trapped Ion Mobility Spectrometry and Top-Down Mass Spectrometry. *Anal. Chem.* **2021**, 93, 10013-10021.

(43) Wagh, A.; Song, H.; Zeng, M.; Tao, L.; Das, T. K. Challenges and New Frontiers in Analytical Characterization of Antibody-Drug Conjugates. *mAbs* **2018**, 222-243.

(44) Linse, S.; Brodin, P.; Johansson, C.; Thulin, E.; Grundström, T.; Forsén, S. The Role of Protein Surface Charges in Ion Binding. *Nature* **1988**, 335, 651-652.

(45) Hunter, T.; Karin, M. The Regulation of Transcription by Phosphorylation. *Cell* **1992**, 70, 375-387.

(46) Dyachenko, A.; Wang, G.; Belov, M.; Makarov, A.; de Jong, R. N.; van den Bremer, E. T.; Parren, P. W.; Heck, A. J. Tandem Native Mass-Spectrometry on Antibody–Drug Conjugates and Submillion Da Antibody–Antigen Protein Assemblies on an Orbitrap EMR Equipped with a High-Mass Quadrupole Mass Selector. *Anal. Chem.* **2015**, 87, 6095-6102.

(47) Keener, J. E.; Zambrano, D. E.; Zhang, G.; Zak, C. K.; Reid, D. J.; Deodhar, B. S.; Pemberton, J. E.; Prell, J. S.; Marty, M. T. Chemical Additives Enable Native Mass Spectrometry Measurement of Membrane Protein Oligomeric State within Intact Nanodiscs. *J. Am. Chem. Soc.* **2019**, 141, 1054-1061.

(48) Li, H.; Nguyen, H. H.; Ogorzalek Loo, R. R.; Campuzano, I. D. G.; Loo, J. A. An Integrated Native Mass Spectrometry and Top-Down Proteomics Method That Connects Sequence to Structure and Function of Macromolecular Complexes. *Nat. Chem.* **2018**, 10, 139-148.

(49) Zhou, M.; Wysocki, V. H. Surface Induced Dissociation: Dissecting Noncovalent Protein Complexes in the Gas phase. *Acc. Chem. Res.* **2014**, 47, 1010-1018.

(50) Skinner, O. S.; Haverland, N. A.; Fornelli, L.; Melani, R. D.; Do Vale, L. H. F.; Seckler, H. S.; Doubleday, P. F.; Schachner, L. F.; Srzentić, K.; Kelleher, N. L.; Compton, P. D. Top-Down Characterization of Endogenous Protein Complexes with Native Proteomics. *Nat. Chem. Biol.* **2018**, 14, 36-41.

CHAPTER 5. Using FAIMS to enhance the performance of CZE-MS/MS for TDP

5.1 Introduction

Top-down proteomics (TDP) requires sufficient separation of complex proteome samples to reduce sample coelution prior to tandem mass spectrometry (MS/MS) analysis. Liquid chromatography (LC) and capillary electrophoresis (CE) are predominant techniques compatible with MS/MS for resolving heterogenous proteoforms.[1-3] Constructing orthogonal multidimensional platforms by combining of different separation methods is frequently adopted for further enhancing identification outcomes. Offline sample fractionations in solution or gel, such as Gel-eluted liquid fraction entrapment electrophoresis (GELFrEE), Passively Eluting Proteins from Polyacrylamide gels as Intact species (PEPPI), size exclusion chromatography (SEC), and reverse phase liquid chromatography (RPLC), have been extensively hyphenated to LC-MS/MS or CE-MS/MS analysis, which boosted the proteoform identification to thousands and even tens of thousands level. [4-9] Incorporating multiple dimensions of liquid phase fractionation is beneficial for achieving higher proteome coverage but needs large starting materials (hundred micrograms to milligrams of the samples), potentially has high sample loss, and is labor intensive.

High-field asymmetric waveform ion mobility spectrometry (FAIMS), as a fractionation strategy in the gas phase, provides rapid and online filtering of ions based on their differential mobilities in oscillating high and low electric fields. The detailed mechanism of FAIMS has been explained in various studies.[10-12] Briefly, the FAIMS is composed of cylindrical inner and outer electrodes with dispersion voltage (DV) applied to deliver an asymmetric waveform. The ions which have different mobilities in high and low field segments of waveform eventually end up colliding with the electrodes. The fractionation of ions can be facilitated by applying compensation voltage (CV) on the inner electrode to offset the drift of ions and selectively allow the transmission of specific groups of ions. A variety of applications of FAIMS in bottom-up proteomics (BUP) have been reported.[12-18] By offering gas-phase fractionation and improved sensitivity, the technique greatly benefited the identification of peptides carrying post-translational modifications (PTMs) and enhanced the depth of proteome coverage. Most recently, FAIMS presented attractive performances in protein complex analysis and TDP. [5,11,19-21] In particular, for TDP, controlling the CV of FAIMS showed potency to fractionate proteoforms according to masses.[11,20] The FAIMS has been coupled with RPLC-MS/MS for deep TDP of cells and tissues.[5,11, 20-21] The works generally achieved around 2-fold number of proteoform identifications (IDs) in contrast to conditions without FAIMS.

Capillary zone electrophoresis (CZE) is one of the most popular CE approaches, which enables the differentiation of proteoforms on basis of their charge-to-size ratios. Our group previously demonstrated the advantage of CZE-MS/MS on separation resolution and sensitivity for TDP.[3,7,22] To our best knowledge, there is no report incorporating FAIMS in CE-MS/MS for TDP. FAIMS have the potential to enhance the proteoform identification performance in CZE-MS/MS analysis and provide better characterization of larger proteoforms via mass-based fractionation. Typically, the identification of intact proteoforms (with full protein sequence or only containing N-terminal methionine excision) and larger proteoforms (>20 kDa) has long been challenging in the field of TDP, mainly due to the loss of these proteoforms during buffer exchange/separation, signal suppression from coeluted small proteins, and limitation of mass spectrometers on mass/resolution/fragmentation. CZE is potentially better suitable for large proteoform separation than RPLC, as the separation capillary causes less sample diffusion and sample loss than an LC column with the stationary phase. Following post-fractionation using FAIMS can further benefit the enrichment and identification of proteoforms of higher masses.

In this study, we carried out the first CZE-FAIMS-MS/MS study for TDP. The yeast lysate extracted by ammonium bicarbonate (ABC, pH~8) was used as a model sample to evaluate the performance of the platform. Different CVs were tested by performing a single CV per CZE-MS/MS run. The results were largely compared between FAIMS and no FAIMS conditions to understand the features of FAIMS and how FAIMS benefits the sensitivity of detection and proteoform identifications in CZE-MS/MS analysis. In addition, we examined the CZE and FAIMS separation of proteoforms originating from the same gene and showed how the coupling of CZE to FAIMS can better differentiate these proteoforms. Finally, we conducted CZE-FAIMS-MS/MS experiments on a yeast sample extracted by urea buffer for comparison with the result of ABC-extracted yeast. The work aimed to demonstrate that directly using a CZE sample buffer, such as ABC, for protein extraction can well maintain the large or intact proteoforms and reduce their loss in buffer exchange. By combining ABC protein extraction protocol and CZE-FAIMS-MS/MS platform, we were capable of improving the identifications of larger or intact proteoforms than conventional TDP workflow.

5.2 Experimental section

5.2.1 Sample preparation

Around 0.2g of Barker's yeast (Saccharomyces cerevisiae, strain ATCC 204508 / S288c) was added in 1L of yeast extract peptone dextrose (YPD) medium (autoclaved) and cultured at 37 °C (300 rpm shaking) overnight in an incubator shaker (Thermo Scientific MaxQ 4000). The

yeast was harvested by centrifugation at 3000g for 5 minutes, followed by washing with phosphate-buffered saline (PBS) three times. The yeast pellets were suspended in cell lysis buffer containing 100 mM ammonium bicarbonate (ABC, pH 8.0), protease inhibitor (cOmplete ULTRA Tables, Roche), and phosphatase inhibitor (PhosSTOP, Roche). The yeast cells were lysed (3 minutes, 3 times) using a homogenizer 150 (Fisher Scientific) and sonicated on ice (10 minutes) with Branson Sonifier 250 (VWR Scientific). The supernatant of the cell lysate was collected with centrifugation at 18,000 g for 10 minutes. The same procedure was applied to the protein extraction of yeast using a buffer consisting of 8M urea,100 mM ABC (pH 8.0), protease inhibitor, and phosphatase inhibitor. The concentrations of protein samples were determined using bicinchoninic acid (BCA) kit (Fisher Scientific) according to the manufacturer's instructions. Before CZE-MS/MS and CZE-FAIMS-MS/MS analyses, 150 µg of the yeast lysates in ABC buffer/urea buffer were loaded onto Amicon centrifugal filters (10 kDa molecular weight cut-off, Millipore Sigma) for buffer exchange. The samples were centrifuged at 14,000 g for 15 minutes at 10 °C and then washed four times with 50 mM ABC.

5.2.2 CZE-MS/MS and CZE-FAIMS-MS/MS analyses

CE-MS/MS system was set up by coupling a CESI 8000 Plus CE system (Beckman Coulter) to an Orbitrap Exploris 480 spectrometer (Thermo Fisher Scientific) with an in-house constructed electrokinetically pumped sheath-flow CE-MS nanospray interface. A glass spray emitter with an orifice size of 30~35 µm was installed on the interface and filled with sheath liquid consisting of 0.2% (v/v) formic acid and 10% (v/v) methanol. The spray voltage was adjusted in the range of 2.2~2.4 kV to generate stable electrospray. The capillary (100 cm length, 50 µm i.d, 360 µm o.d) for CZE was coated with linear polyacrylamide (LPA), according to our previous protocol.23 The inlet of the capillary was fixed in the cartridge of the CE system and the outlet was inserted into the emitter of the interface. The distance of the capillary outlet to the emitter orifice was around 0.5 mm.

To carry out CZE, the capillary was flushed with a background electrolyte (BGE, 5% acetic acid) at 10 psi for 10 minutes, followed by loading of 200 ng of yeast lysate (1mg/mL, injection volume of 200nL). Afterward, the inlet of the capillary was inserted into the background electrolyte (5% acetic acid) with a separation voltage of 30 kV applied.

For the mass spectrometer, the temperature of the ion transfer tube was set to 320 °C and RF lens was 60%. The intact protein mode was turned on and the low-pressure mode was selected. The MS/MS experiments were performed using data-dependent acquisition (DDA). Full MS scan was performed with the following parameters: orbitrap resolution of 480,000 (at m/z of 200), m/z range of 500-2500, normalized AGC target of 300%, microscan of 1. The top 6 most

intense precursors in full MS spectra were isolated with a window of 2 m/z and fragmented using HCD collision energy of 25%. Only precursors with charge states in the range of 5-60 and intensities higher than the threshold value of 10000 were included for fragmentation. Other parameters for MS/MS include the resolution of 60,000 (at m/z 200), m/z range of 200-2000, microscan of 3, normalized AGC target of 100%, and auto maximum injection time. The dynamic exclusion was applied with a duration of 30 s and the exclusion of isotopes was enabled.

FAIMS Pro interface (Thermo Fisher Scientific) was installed, had auto DV tune, and was set to standard resolution for CZE-FAIMS-MS/MS analysis. The nitrogen carrier gas was set as default (4.6L/min). Different CV voltage ranging from -50 V to 30 V with 10V intervals was tested for nine individual CZE-MS/MS runs to investigate the fractionation performance of the FAIMS.

5.2.3 Data analysis

All the raw files were converted to mzML files using MSConvert and further deconvoluted to Msalign files using TopFD (version 1.4.7). The converted data were searched against the Uniprot S. cerevisiae database (UP000002311_559292) using TopPIC (1.4.7). Parameters for database search were set as follows: mass error of precursors and fragments of 15 ppm, the maximum number of unexpected modifications of 2, and maximum and minimum mass shifts of unknown modifications of 500 Da and -500 Da. The false discovery rates (FDRs) were estimated using the target-decoy approach. The spectrum level FDR cut-off was 1% and the proteoform level FDR cut-off was 5%.

5.3 Results and discussions

5.3.1 Investigation of features of the CZE-FAIMS-MS/MS system

CZE-FAIMS-MS/MS analysis of a yeast sample was performed to investigate the performance of the system. We directly used the sample buffer for CZE (ABC) to extract yeast proteins to reduce sample loss during the buffer exchange. Different CVs ranging from -50V to 30V were tested with 10 V increments for each CZE-MS/MS run. The same CZE-MS/MS condition was also conducted without FAIMS for comparison. Without FAIMS, most separation peaks appeared in around 25 minutes migration window, Figure 5.1. In contrast, when FAIMS was installed, distinct separation profiles were presented at different CV settings, Figure 5.1. At higher CV values from -10V to 30V, the peaks were mainly fractionated from 30 to 40 minutes. However, those peaks were overwhelmed and had coelution under the no FAIMS condition. The result strongly suggests the importance of employing FAIMS for additional separation of proteoforms with similar electrophoretic motility in solution. Several LC-MS/MS studies previously reported a significant decrease in protein signals using FAIMS compared to no FAIMS owing to the longer

path for ion transmission.[11, 20] In our study, although some ions showed a slightly lower signal intensity under FAIMS conditions, a variety of ions were observed with increased signals at their best CV values compared to no FAIMS, which was likely because FAIMS reduced signal suppression by removing other coeluted ions.

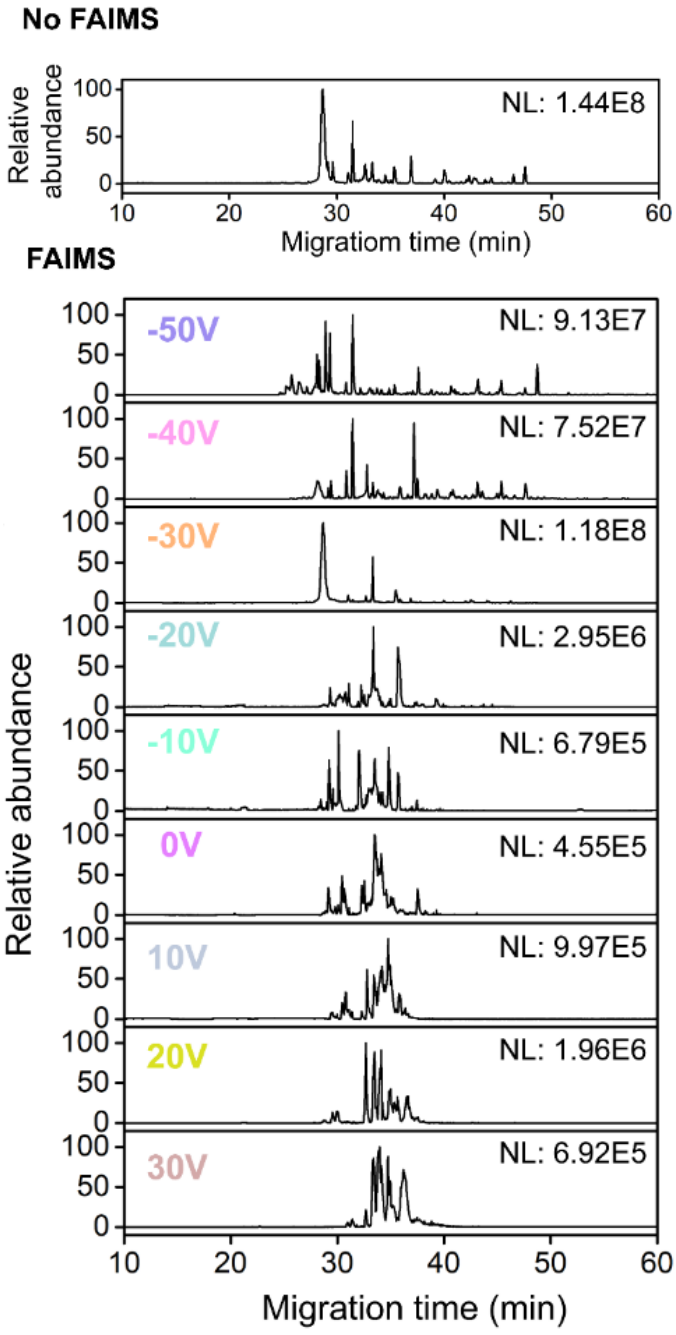

Figure 5.1 Base peak electropherogram of yeast lysate without FAIMS and with FAIMS at nine different CVs.

To better understand the features of FAIMS in CZE-MS/MS analysis, we examined the overlap of proteoform identifications, mass distribution of proteoforms, and sensitivity of detection at different CVs. Figure 5.2 showed larger overlaps between the neighboring CVs than the CVs with longer distances. The highest overlap (46%) occurred between -20 and -10V, whereas the lowest (10%) was between 20 and 30V. The overlaps of other neighboring CVs were around 30~40%. We also performed triplicate CZE-MS/MS analysis without FAMIS and with FAIMS at -40V. The electropherograms presented reproducible CZE separation profiles. The overlap between the triplicate datasets without FAMIS was 54% to 58%, which is reasonable due to the data dependent acquisition (DDA) method. In contrast, the overlap between FAIMS (-40V) triplicate (60 to 73%) was slightly higher than the no FAIMS condition. This could be due to the reduced complexity of mass spectra after fractionation, thereby giving the precursors of the same proteoform a higher chance to be isolated for fragmentation. In addition, the mass distribution of proteoforms indicated a correlation with CV values. As shown in Figure 5.2, the median mass was 7 kDa without FAIMS fractionation. With FAIMS, the median mass was increased from 6 kDa to 30 kDa when raising the CV from -50 to 30 V, revealing the higher CV favors the transmission of larger proteoforms. It is also interesting to see small proteoforms have a wide distribution of migration time in CZE, while the larger proteoforms are more concentrated in earlier migration time (Figure 5.1). For example, at -50V, most identified proteoforms were less than 10 kDa and were separated between 25 to 50 minutes. As CV increases to -10V, the identified proteoforms, mainly 10~20 kDa, were shifted to a migration window between 28 to 40 min. The electrophoretic mobilities of small proteoforms are intrinsically more sensitive to a small difference in charge or size, thereby generating a wide separation window. The large proteoforms usually carry higher charges and are more likely to have faster electrophoretic mobility. Moreover, we found the application of FAIMS significantly improved the sensitivity of detection in CZE-MS/MS analysis. We manually extracted base peak electropherograms of 20 proteoforms identified in both FAIMS and no FAIMS conditions and compared their signal-to-noise ratios (S/N). On average, using FAIMS presented around a 50-fold higher S/N than the no FAIMS condition (Figure 5.2). The improvement was mainly benefited by much-reduced background noise using FAIMS, which was in accordance with the observations in other studies related to FAIMS.[11,18]

Furthermore, comparing the number of proteoform IDs showed a single CV of -50V was capable of identifying 32% more proteoforms than no FAIMS (432 vs. 327). Higher CV value particularly boosted the number of larger proteoforms. For example, the CV of 30V was identified around 3-fold of proteoforms larger than 20 kDa than no FAIMS (58 vs. 20). Combining the results of different CVs, we achieved 940 proteoform identifications from 288 proteins, which is a 3-fold

proteoform-level and 2.2-fold protein-level increase compared to the result without FAIMS (327 proteoforms and 126 proteins), Figure 5.2. 65% (213) of proteoforms identified without FAIMS were included in the results of FAIMS, Figure 5.2. In addition, the combined CVs obtained a nearly 6-fold improvement of indications proteoforms in the range of 20 kDa and 30 kDa, and above 30 kDa, Figure 5.2.

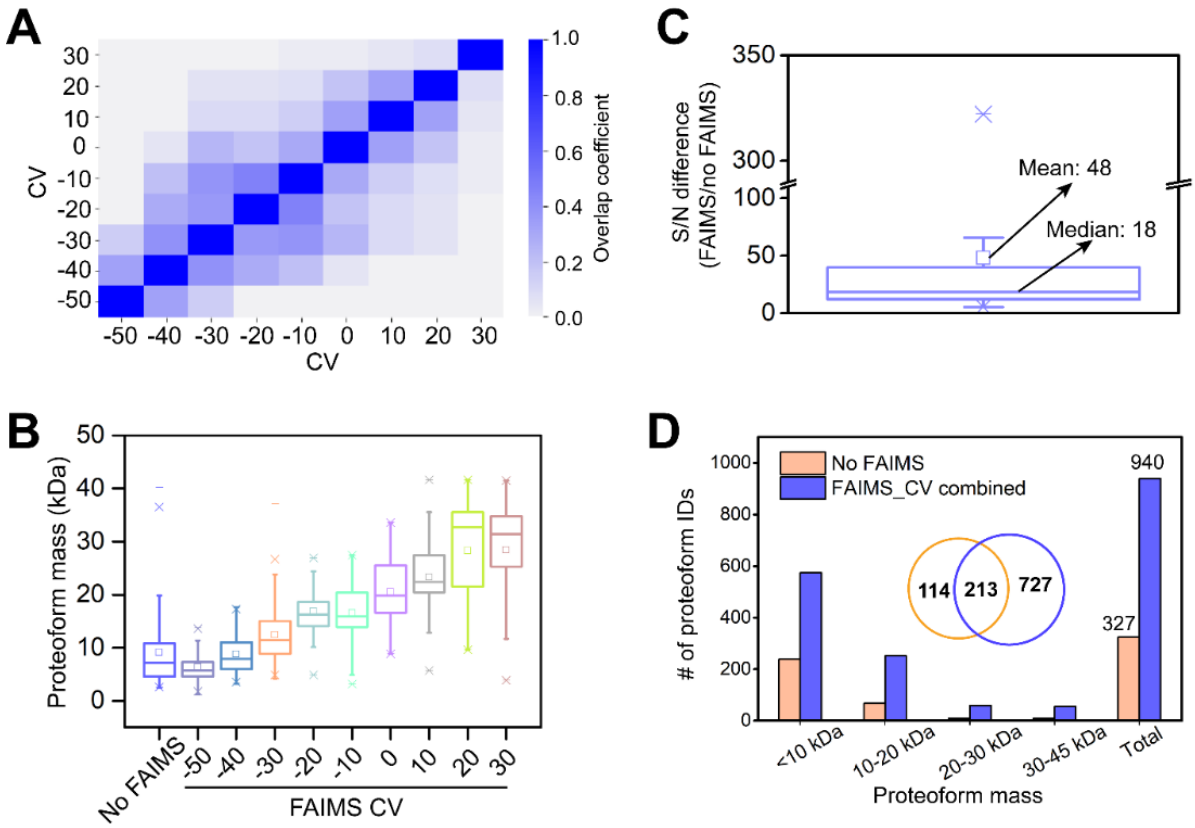

Figure 5.2 Comparison of proteoform identifications without FAIMS and with FAIMS in CZE-MS analysis. (A) Overlap of identified proteoforms between FAIMS CVs. (B) Mass distribution of proteoforms identified without FAIMS and with different FAIMS CVs. (C) Improvement of the signal-to-noise ratio of proteoforms with FAIMS relative to no FAIMS condition. (D) Comparison of the number of proteoform identifications at the different mass ranges between no FAIMS condition and FAIMS condition.

5.3.2 Enhancing proteoform characterization by combining CZE with FAIMS

Proteoforms derived from the same gene can have diverse functions. High-resolution separation and characterization of the proteoforms are vital for understanding their roles in biological processes. For some of the proteoforms, such as proteoforms with different PTMs, it is difficult to achieve good separations using RPLC, because of their high similarities in

hydrophobicity. In this case, CZE could offer better resolutions for those proteoforms on charge/size differences. We previously reported shifts of mobility of proteoforms with phosphorylation based on the prediction of electrophoretic mobility.[24] In this study, we also observed the separation of proteoforms without and with phosphorylation by CZE.

As shown in Figure 5.3, the two peaks baseline separated in CZE-FAIMS (0V)-MS/MS analysis represents two proteoforms of *Nascent polypeptide-associated complex subunit alpha (NAC-α)* with N-terminal methionine excision and N-terminal acetylation but containing no phosphorylation (proteoform 1) and phosphorylation (proteoform 2), respectively. The fragmentation pattern (Figure 5.3B) showed the phosphorylation of *NAC-α* is located between Pro92 and Ala113. The modification site can be further confirmed to Ser93 by the information on Uniprot (https://www.uniprot.org/uniprot/P38879). *NAC-α* can either be tethered to the cytoplasmic ribosome and function as a complex component *of nascent polypeptide-associated complex (NAC)* to modulate co-translational processes and protein translocation or accumulate in nuclei to participate in transcriptional coactivation.[25-29] The phosphorylation of *NAC-α* was found to be regulated by the proteasome pathway and associated with its degradation.[29] Our result showed that non-phosphorylated *NAC-α* has a much higher abundance than phosphorylated proteoform, indicating protein depletion only occurs in a small portion of *NAC-α* (Figure 5.3A).

Furthermore, we found that the phosphorylated proteoform presented slower migration relative to the non-phosphorylated proteoforms in CZE. The phosphorylation can reduce the protein charge, therefore leading to decreases in the charge-to-size ratio and mobility. Although the CZE has the potential to resolve some proteoforms with PTM heterogeneity, for a complex proteome, however, it remains challenging to achieve sufficient separation for every species. For *NAC-α,* none of the two proteoforms shown in CZE-FAIMS-MS/MS above were identified in the triplicate CZE-MS/MS runs without FAIMS, due to signal suppression from other coeluted species. The result strongly highlights incorporating FAIMS to CZE-MS/MS can provide better proteoform characterization of the complex proteome.

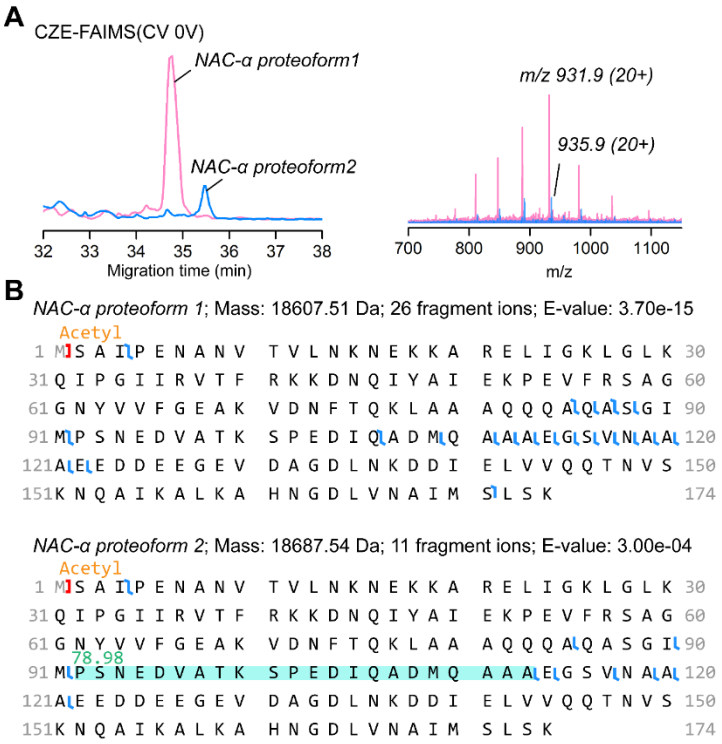

Figure 5.3 Two intact proteoforms of nascent polypeptide-associated complex subunit alpha (*NAC-α*) identified in CZE-FAIMS-MS analysis at a CV value of 0V. (A) Overlapped base peak electropherograms of two *NAC-α* proteoforms separated by CZE. (B) Fragmentation patterns of *NAC-α* proteoform 1 (unphosphorylated intact *NAC-α*) and *NAC-α* proteoform 2 (phosphorylated *NAC-α*).

We were also interested in the performance of FAIMS for fractionating the proteoforms originating from the same genes in CZE-MS/MS analysis. Here, the proteoforms identified without FAIMS and at different CVs were merged according to their protein accession numbers (Figure 5.4A). We found 51% of genes had their proteoforms distributed into more than three different CVs. For the genes covered by both FAIMS and no FAIMS conditions, 56% of those genes had more proteoforms identified at a single CV condition than no FAIMS, and the total number of proteoforms IDs per gene was 2.2-fold improved in combined CVs relative to no FAIMS (Figure 5.4B). Figure 5.4C showed two examples of proteoform identifications from the same gene without FAIMS and with FAIMS. The proteoforms of *TPIS* were detected in a wide CV range from -30V to 30V. Four out of six CVs detected more proteoforms than no FAIMS. Similarly, *CYPH* has three out of four CVs (from -40V to -10V) identified more proteoforms. For both genes, the total number of proteoform IDs was increased 6.3 times by FAIMS.

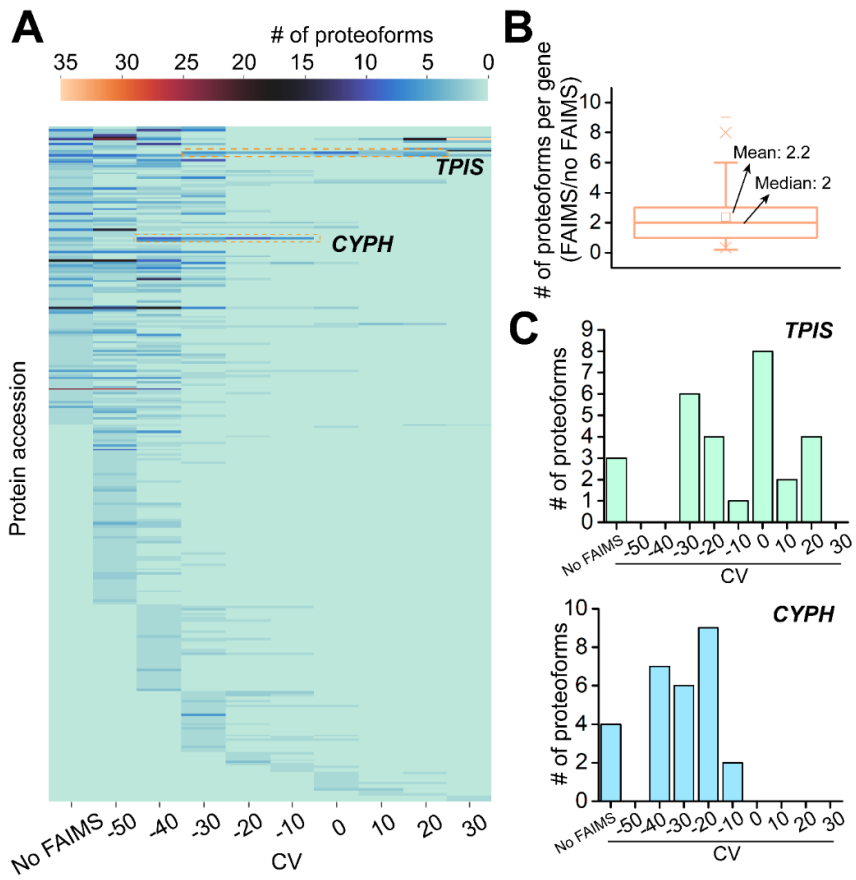

Figure 5.4 Gas-phase fractionation of proteoforms originating from the same genes in CZE-MS analysis. (A) Heatmap of proteoform IDs at different CVs merged by their protein accessions. (B) Improvement of the number of proteoforms per gene using FAIMS relative to no FAIMS condition (C)The number of proteoform IDs of CYPH and TPIS without FAIMS and at different CVs.

5.3.3 Impact of protein extraction buffers on proteoform IDs

Current TDP works mostly focus on proteoforms with lower mass (<20 kDa), due to the technical limitation mentioned in the introduction section. Improvement of separation is one important solution to expand the MW of proteoform IDs. Therefore, different multidimensional separation platforms have been developed by coupling different LC methods or combining LC with CE to achieve this goal. In this study, we offered CZE-FAIMS as a new 2D separation platform to promote proteoform IDs with higher mass. Besides separation, the sample preparation method could be another factor that impacts the mass distribution of proteoforms. The conventional TDP typically uses buffers containing detergent or urea for cell lysis and protein extraction, which requires desalting/buffer exchange before MS/MS analysis. Protein precipitation could occur in this process due to a dramatic change in salt condition and hydrophobic effect,

leading to the loss of large proteins. For CZE-MS/MS, the sample needs to be exchanged to a relatively basic buffer (50 mM ABC, pH~8) to enable dynamic pH junction.[22,31] We speculate that direct using ABC buffer for protein extraction could reduce the loss of protein in buffer exchange and benefit larger proteoform identification. To understand how the protein extraction buffers influence protein identifications, we carried out CZE-FAIMS-MS/MS analysis of the urea-extracted yeast and compared the sample with ABC.

CZE-FAIMS-MS/MS experiment of the urea sample was performed at three single CVs (-50V, -40V, and -30V) for better throughput. Based on our experience with the ABC sample, 80% of proteoforms could be detected in those three CVs. We found that the overlaps of proteoforms between the urea and yeast samples at the same CVs or without FAIMS are generally small (20%~28%) (Figure 5.5A), suggesting the two buffers favor the extraction of different proteoforms.

Moreover, we found more proteoforms were identified in urea than ABC either without FAIMS (1217 vs. 327) or with 3-CV combined (2070 vs. 769) (Figure 5.5B). However, a very low percentage of these proteoforms (~4%) in urea have masses higher than 10kDa. For the ABC sample, the proteoforms above 10 kDa account for 22% without FAIMS and 27% for three-combined CVs (-50V, -40V, -30V). The absolute number of proteoforms larger than 10kDa is also higher in ABC than in urea (211 vs. 103, 3 CV-combined) (Figure 5.5B). The result strongly suggests that ABC can better preserve the proteoform above 10kDa than urea buffer. Using nine CVs for the ABC sample furtherly boosted larger proteoform identification (>20 kDa) from 14 (3 CVs) to 114, which is 12% of total proteoform IDs. Furthermore, the ABC sample tends to conserve more “intact” proteoforms that either cover the full protein sequences or only have N-terminal methionine removed. Our result showed 78 (23%) and 268 (29%) of proteoforms in the ABC sample were intact without and using FAIMS, respectively (Figure 5.5C). In contrast, the majority of proteoforms (96%) in urea were truncated forms.

Improvement of intact proteoform identification allows us to decipher their functions and properties in cells. We identified 8 intact proteoforms of *isoform cytoplasmic of glutaredoxin 2* (*Grx2c*, P17695-2) in the ABC sample. *Grx2c* is an important glutathione-dependent oxidoreductase in the cytosol, participating in the reduction of protein disulfide bonds.[32, 33] Our result showed intact Crx2c proteoforms were either N-terminal methionine removed or reserved and contained different PTMs. For the forms without N-terminal methionine, one proteoform has two free thiols at Cys 27 and Cys30 (E-value: 1.77e-11), which could be associated with their oxidoreductase activity; one contains phosphorylation at Ser57 (E-value: 6.49e-7), one has a mass shift of 129 Da in the range of Leu51 to Leu53 (4.05e-10), which might be glutamylation on Glu52; the other one has a mass shift of 210 Da from Leu51 to Ser57 (E-value: 1.54e-6), which

could be a combination of phosphorylation and glutamylation. The glutamylation of Glu was previously found in one other protein[34] but never reported in Grx2c. We suspect the two glutamylated Grx2 proteoforms might result from other biological processes related to glutathione. In addition, we identified 33 proteoforms related to isomers of acidic ribosomal P proteins (*RPLP1-α, RPLP1-β, RPLP2- α, RPLP2-β*). Interestingly, while intact proteoforms were found in both *RPLP2- α and RPLP2-β,* only truncated proteoforms were detected in *RPLP1- α and RPLP1-β.* The result is in good agreement with the report that P1 proteins generally have much lower half-lives than P2 in yeast cells[30].

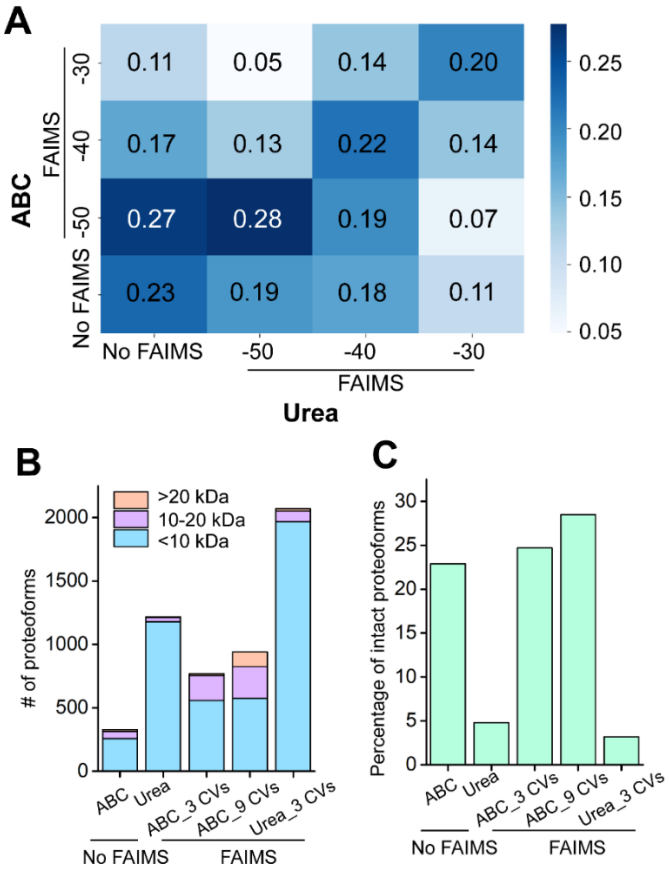

Figure 5.5 Comparison of proteoform IDs of yeast lysate prepared from different protein extraction buffers (ABC vs urea). (A) Overlap of proteoforms between yeast lysates extracted by ABC and urea. (B, C) Mass distribution of proteoforms (B) and percentage of the number of intact proteoforms (C) extracted by ABC and urea, and under no FAIMS and FAIMS conditions.

5.4 Conclusions

We incorporated FAIMS to CZE-MS/MS for TDP and demonstrated the features of the system. The gas-phase fractionation by FAIMS greatly reduced the complexity of CZE

electropherograms and improved the sensitivity of CZE-MS/MS by around 50-fold on average. The CVs of FAIMS showed strong fractionation dependence on the proteoform masses. Overall, CZE-FAIMS-MS/MS increased the number of proteoform IDs 3-fold compared to no FAIMS. To have a deeper understanding of how the combination of CZE-MS/MS and FAIMS benefit proteoform separation and identification, we typically focused on several examples in which CZE showed superior separation performance on proteoforms with heterogeneity on PTMs (e.g. phosphorylation) and the addition of FAIMS further assisted identification of these proteoforms. Our result highlighted the potential of CZE-FAIMS-MS/MS for better characterizing proteoforms derived from the same genes. Furthermore, considering FAIMS has the capability to fractionate large proteoforms from complex proteome, we were interested in whether we can make full use of CZE-FAIMS-MS/MS system to improve the identification of larger and intact proteoforms. We compared two yeast samples extracted from different buffers (ABC and urea) at 3 combined CVs (-50V, -40V, -30V). We found that ABC better preserved proteoforms above 10 kDa as well as intact proteoforms compared to urea (~25% vs. 4%). CZE-FAIMS-MS/MS analysis of ABC-extracted yeast at 9 combined CVs identified more than 100 proteoforms above 20 kDa, which accounts for more than 10% of total proteoform IDs. We need to note that ABC tends to extract more hydrophilic proteoforms, which results in the missing identification of plasma membrane proteins. This defect might be compensated by combining a membrane protein extraction protocol with CZE-FAIMS-MS/MS to target membrane protein analysis in the future.

Our current CZE-FAIMS-MS/MS method using 9 CVs consumed only around 2 μg of yeast in total and required around 10 hours for sample analysis. While this study focuses on investigating fractionation and proteoform identification outcomes of CZE-FAIMS-MS/MS at different CVs, internal CV stepping, which uses multiple CVs in a single run, is worth being applied to CZE-MS/MS for better throughput in the future. A further coupling of other liquid phase separation approaches such as SEC with CZE-FAIMS-MS/MS can serve as a promising platform for deep TDP. Moreover, although an improvement on larger proteoform IDs (>20 kDa) was obtained in the study, there remains room to upgrade database search software to achieve better results. The software we currently used remain favors analysis of proteoforms below 30 kDa. Besides data acquisition using high-resolution MS1 (480,000) and MS2 (6,000) settings, we tested low-resolution MS1 (7,500) and high-resolution MS2 (12,000) at the CV of 30V. The low-resolution MS1 presented various proteoforms above 40 kDa by manual deconvolution. However, the current software remains difficult in interpreting this type of data. We expect the advancement of bioinformatics tools can greatly boost the number of larger proteoform IDs.

REFERENCES

(1) 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.

(2) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-Down Proteomics: Ready for Prime Time? *Anal. Chem*. **2018**, *90* (1), 110-127.

(3) Chen, D.; McCool, E.N.; Yang, Z.; Shen, X.; Lubeckyj, R.A.; Xu, T.; Wang, Q.; Sun, L. Recent Advances (2019–2021) of Capillary Electrophoresis-Mass Spectrometry for Multilevel Proteomics. *Mass Spectrom. Rev.* **2021**

(4) Melani, R. D.; Gerbasi, V. R.; Anderson, L. C.; Sikora, J. W.; Toby, T. K.; Hutton, J. E.; Butcher, D. S.; Negrão, F.; Seckler, H. S.; Srzentić, K.; Fornelli, L.; Camarillo, J. M.; LeDuc, R. D.; Cesnik, A. J.; Lundberg, E.; Greer, J. B.; Fellers, R. T.; Robey, M. T.; DeHart, C. J.; Forte, E.; Hendrickson, C. L.; Abbatiello, S. E.; Thomas, P. M.; Kokaji, A. I.; Levitsky, J.; Kelleher, N. L. The Blood Proteoform Atlas: A Reference Map of Proteoforms in Human Hematopoietic Cells. *Science* **2022**, *375* (6579), 411-418.

(5) Takemori, A.; Kaulich, P. T.; Cassidy, L.; Takemori, N.; Tholey, A. Size-Based Proteome Fractionation through Polyacrylamide Gel Electrophoresis Combined with LC-FAIMS-MS for In-Depth Top-Down Proteomics. *Anal. Chem.* **2022**, *94* (37), 12815-12821.

(6) 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.

(7) 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.

(8) Xu, T.; Shen, X.; Yang, Z.; Chen, D.; Lubeckyj, R.A.; McCool, E.N.; Sun, L. Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative And Quantitative Top-Down Proteomics. *Anal. Chem.* **2020**, *92*(24),15890-15898.

(9) McCool, E.N.; Xu, T.; Chen, W.; Beller, N.C.; Nolan, S.M.; Hummon, A.B.; Liu, X.; Sun, L. Deep Top-Down Proteomics Revealed Significant Proteoform-Level Differences between Metastatic and Nonmetastatic Colorectal Cancer Cells. *Sci. Adv.* **2022**, *8* (51), 6348.

(10) Cooper, H. J. To What Extent Is FAIMS Beneficial in the Analysis of Proteins? *J. Am. Soc. Mass Spectrom.* **2016**, *27* (4), 566-577.

(11) Gerbasi, V. R.; Melani, R. D.; Abbatiello, S. E.; Belford, M. W.; Huguet, R.; McGee, J. P.; Dayhoff, D.; Thomas, P. M.; Kelleher, N. L. Deeper Protein Identification Using Field Asymmetric Ion Mobility Spectrometry in Top-Down Proteomics. *Anal. Chem.* **2021**, *93* (16), 6323-6328.

(12) Hebert, A. S.; Prasad, S.; Belford, M. W.; Bailey, D. J.; McAlister, G. C.; Abbatiello, S. E.; Huguet, R.; Wouters, E. R.; Dunyach, J.-J.; Brademan, D. R.; Westphall, M. S.; Coon, J. J. Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Anal. Chem.* **2018**, *90* (15), 9529-9537.

(13) Swearingen, K. E.; Hoopmann, M. R.; Johnson, R. S.; Saleem, R. A.; Aitchison, J. D.; Moritz, R. L. Nanospray FAIMS Fractionation Provides Significant Increases in Proteome Coverage of Unfractionated Complex Protein Digests. *Mol. Cell Proteomics* **2012**, *11* (4).

(14) Schnirch, L.; Nadler-Holly, M.; Siao, S.-W.; Frese, C. K.; Viner, R.; Liu, F. Expanding the Depth and Sensitivity of Cross-Link Identification by Differential Ion Mobility Using High-Field Asymmetric Waveform Ion Mobility Spectrometry. *Anal. Chem.* **2020**, *92* (15), 10495-10503.

(15) Adoni, K. R.; Cunningham, D. L.; Heath, J. K.; Leney, A. C. FAIMS Enhances the Detection of PTM Crosstalk Sites. *J. Proteome Res.* **2022**, *21* (4), 930-939.

(16) Greguš, M.; Kostas, J.C.; Ray, S.; Abbatiello, S.E.; Ivanov, A.R. Improved sensitivity of ultralow flow LC–MS-based proteomic profiling of limited samples using monolithic capillary columns and FAIMS technology. *Anal. Chem.* **2020**, *92* (21),14702-14712.

(17) Fang, P.; Ji, Y.; Silbern, I.; Viner, R.; Oellerich, T.; Pan, K.T.; Urlaub, H. Evaluation and optimization of High-Field Asymmetric Waveform Ion-Mobility Spectrometry for multiplexed quantitative site-specific N-glycoproteomics. *Anal. Chem.* **2021**, *93*(25), 8846-8855.

(18) Johnson, K. R.; Greguš, M.; Ivanov, A. R. Coupling High-Field Asymmetric Ion Mobility Spectrometry with Capillary Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry Improves Protein Identifications in Bottom-Up Proteomic Analysis of Low Nanogram Samples*. J. Proteome Res.* **2022**, *21* (10), 2453-2461.

(19) Hale, O. J.; Illes-Toth, E.; Mize, T. H.; Cooper, H. J. High-Field Asymmetric Waveform Ion Mobility Spectrometry and Native Mass Spectrometry: Analysis of Intact Protein Assemblies and Protein Complexes. *Anal. Chem.* **2020**, *92* (10), 6811-6816.

(20) Fulcher, J.M., Makaju, A., Moore, R.J., Zhou, M., Bennett, D.A., De Jager, P.L., Qian, W.J., Paša-Tolić, L. and Petyuk, V.A. Enhancing Top-Down Proteomics of Brain Tissue with FAIMS. *J. Proteome Res.* **2021**, *20* (5), 2780-2795.

(21) Kaulich, P. T.; Cassidy, L.; Winkels, K.; Tholey, A. Improved Identification of Proteoforms in Top-Down Proteomics Using FAIMS with Internal CV Stepping. *Anal. Chem.* **2022**, *94* (8), 3600-3607.

(22) 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.

(23) Xu, T.; Han, L.; Thompson, A. M. G.; Sun, L. An Improved Capillary Isoelectric Focusing-Mass Spectrometry Method for High-Resolution Characterization of Monoclonal Antibody Charge Variants. *Anal. Methods* **2022**, *14* (4), 383-393.

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

(25) Andersen, K. M.; Semple, C. A.; Hartmann-Petersen, R. Characterisation of the Nascent Polypeptide-Associated Complex in Fission Yeast. *Mol. Biol. Rep.* **2007**, *34* (4), 275-281.

(26) Ott, A.K.; Locher, L.; Koch, M.; Deuerling, E. Functional Dissection of the Nascent Polypeptide-Associated Complex in Saccharomyces Cerevisiae. *PLoS One* **2015**, *10* (11).

(27) Raue, U.; Oellerer, S.; Rospert, S. Association of Protein Biogenesis Factors at the Yeast Ribosomal Tunnel Exit is Affected by the Translational Status and Nascent Polypeptide Sequence. *J. Biol. Chem.* **2007**, *282* (11), 7809-7816.

(28) George, R.; Walsh, P.; Beddoe, T.; Lithgow, T. The Nascent Polypeptide-Associated Complex (NAC) Promotes Interaction of Ribosomes with the Mitochondrial Surface in Vivo. *FEBS Lett.* **2002**, *516* (1), 213-216.

(29) Quélo, I.; Akhouayri, O.; Prud'homme, J.; St-Arnaud, R. GSK3β-Dependent Phosphorylation of the αNAC Coactivator Regulates Its Nuclear Translocation and Proteasome-Mediated Degradation. *Biochemistry* **2004**, *43* (10), 2906-2914.

(30) Tchórzewski, M. The Acidic Ribosomal P Proteins. *Int. J. Biochem. Cell Biol.* **2002**, *34* (8), 911-915

(31) 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.

(32) Collinson, E.J.; Grant, C.M. Role of Yeast Glutaredoxins as Glutathione S-Transferases. *J. Biol. Chem.* **2003**, *278* (25), 22492-22497.

(33) Porras, P.; McDonagh, B.; Pedrajas, J.R.; Bárcena, J.A.; Padilla, C.A. Structure and Function Of Yeast Glutaredoxin 2 Depend on Postranslational Processing and Are Related to Subcellular Distribution. *Biochim Biophys Acta Proteins Proteom BBA-PROTEINS PROTEOM* **2010**, *1804* (4), 839-845.

(34) Eddé, B.; Rossier, J.; Le Caer, J.-P.; Desbruyères, E.; Gros, F.; Denoulet, P. Posttranslational Glutamylation of α-Tubulin. Science **1990**, *247* (4938), 83-85.

CHAPTER 6.* Application of SEC-cIEF-MS/MS for studying sexual dimorphism of brains

6.1 Introduction

Sexual dimorphism of brains, which is mainly generated from the expression of sex chromosome genes and effects of hormones secreted from gonads, determines phenotypic differences on memory, cognition, emotion, stress responsivity, and reproductive behaviors.[1] Only several works employed quantitative BUP to study sexual dimorphism of brains.[2-4] Based on our knowledge, no quantitative TDP studies have been done to compare male and female brain proteomes in a proteoform-specific manner. Zebrafish is an important model organism in developmental biology for both embryogenesis studies and drug development.[5] Here, we performed a label-free quantitative TDP study using SEC-cIEF-MS/MS developed in Chapter 2 to investigate the sex-related proteoforms in zebrafish brains. The quantitative proteomics datasets of zebrafish brains from TDP were compared with bottom-up proteomics (BUP) data to better understand the features the two proteomics strategies.

6.2 Experimental section

6.2.1 Sample preparation

The zebrafish (AB/Tuebingen line, 11 months old) brain tissues (5 males and 5 females) were provided by Professor Jose Cibelli's laboratory at Department of Animal Science of Michigan State University. The whole protocol for collecting zebrafish brains were operated in compliance with national and institutional guidelines for animal research. All the experimental procedures were approved by Institutional Animal Care and Use Committee of Michigan State University. The collected brains were gently washed by PBS to remove the blood, and stored at -80 °C. The zebrafish brains were lysed with mammalian cell-PE LB™ buffer (G-Biosciences) containing protease and phosphatase inhibitors (Roche). The protein extraction, denaturation, reduction, alkylation, and buffer exchange experiments were performed using the same protocol applied to the *E. coli* sample in Chapter 2.

6.2.2 SEC fractionation of zebrafish proteome

The same LC system and SEC column in Chapter 2 was applied for SEC fractionation of zebrafish proteome. 500 μg (2 mg/mL, 50 μL × 5 injections) of male or female brain proteins were

---

* This chapter is adapted with permission from *Xu, T.; Shen, X.; Yang, Z.; Chen, D.; Lubeckyj, R. A.; McCool, E. N.; Sun, L. Anal. Chem. 2020, 92 (24), 15890-15898.*

loaded onto the SEC column and separated using 0.05% trifluoroacetic acid (TFA) at a flow rate of 0.25 mL/min. Three fractions were collected from 9.30 to 15.3 minute with 2 minutes per fraction (500 µL per fraction) and the final fraction (1000 µL) were obtained from 15.3 to 19.3 minute. The fractions were lyophilized in the speed vacuum and re-dissolved in 20 µL of 10 mM ammonium acetate (pH 6.9).

6.2.3 cIEF-MS/MS analysis

For quantitative TDP analysis of zebrafish brain proteome, a high-throughput cIEF-MS/MS method developed in Chapter 2 was used, which utilized an 80-cm-long LPA-coated capillary, a 5-cm plug of 0.3% (v/v) NH3·H2O (pH 11.8) as catholyte, a 40-cm sample plug containing brain proteins, 0.1% ampholyte, and 0.1% (v/v) FA as anolyte (pH ~3.0). The same MS/MS parameters in Chapter 2 were applied for zebrafish samples.

6.2.4 Database search for proteoform identification and quantification

The same database search software and parameters were applied for the identification of proteoforms of zebrafish brains. UniProt database of zebrafish (UP000000437) was used. The data of eight zebrafish fractions including four fractions of male brains and four fractions of female brains were also filtered with a 5% proteoform-level FDR.

When performing label-free quantification of proteoforms expressed in male and female brains of zebrafish, the feature intensity of a proteoform was calculated as sum of intensities of its corresponding peaks from all scans and charge states as described in our previous work.[6] For each SEC fraction, only the proteoforms which were reported with feature intensities in all triplicate cIEF-MS/MS runs were further considered for relative quantification.

6.2.5 Quantitative bottom-up proteomics (BUP) of zebrafish male and female brains

Aliquots of the zebrafish brain lysates prepared in the “sample preparation” section (100 µg of total proteins each gender) were precipitated with four times volumes of cold acetone overnight. After centrifugation at 14,000 g for 10 minutes, the supernatants were removed, and the pellets containing the extracted proteins were washed with cold acetone twice and air-dried at room temperature to remove residual acetone. The extracted proteins were dissolved in 100 µL of 100 mM ammonia bicarbonate buffer (pH 8.0) containing 8 M urea. The proteins were denatured at 37 °C for 30 minutes, reduced by adding 2 µL of 100 mM DTT at 37 °C for 30 minutes, and alkylated by adding 5 µL of 100 mM IAA for 20 minutes in dark at room temperature. The reaction was quenched by adding 2 µL of 100 mM DTT. The protein samples were diluted by five times using 100 mM ammonia bicarbonate, followed by trypsin (8 µg, Bovine pancreas TPCK-treated) digestion at 37 °C overnight. The digestion was finally terminated by adding 8 µL of 20% (v/v) formic acid. The samples were desalted with Sep-Pak C18 Cartridge (Waters) according to

manufacturer's protocol. The eluates were lyophilized in a vacuum concentrator, and then re-dissolved in 100 µL of 50 mM HEPES buffer (pH 8.0).

The tryptic digest of each gender was equally divided into three aliquots (33 µg peptide each aliquot). The TMT labeling experiment (6 channels) was performed by mixing each aliquot with a TMT reagent and incubating them at room temperature for 1 hour, according to manufacturer's instruction. Specifically, the three aliquots of male brains were labeled with 126, 127, 128 isobaric tags, while the aliquots of female brains were labeled with 129, 130, 131 isobaric tags, respectively. After TMT labeling, the six aliquots were combined into one sample and lyophilized. The sample was finally re-dissolved in 100 µL of 0.1% (v/v) formic acid in water.

A C18 reversed-phase column (Zobax 300 Extend-C18, 2.1mm i.d. × 150 mm length, 3.5 µm particles, Agilent Technologies) was used for fractionation of the TMT labeled peptide sample. A gradient elution with a flow rate of 0.3 mL/min was applied as follows: 0-5 min, 2% B (0.1% FA in 80% ACN- 20% Water mixture); 5-7 min, 2-5% B; 7-67 min, 5-50% B; 67-69 min, 50-100% B; 69-79 min, 100% B; 79-81 min, 100-2% B; 81-90 min, 2%B. The mobile phase A was a 2% ACN-98% water mixture containing 0.1% FA (v/v). 200 µg of TMT labeled peptides were loaded onto the column. 54 fractions were collected from 11.4 min to 65.4 min with 1 min each fraction. Subsequently, the fraction N was combined with fraction N+27 to generate 27 fractions in total. The fractions were dried in a speed vacuum and then re-dissolved in 20 µL of 20 mM ammonia bicarbonate (pH 8.0) for CZE-MS/MS analysis.

CZE-MS/MS analysis was performed on the same platform and coupling strategy as applied to cIEF-MS/MS. A 1-meter LPA-coated capillary (50 µm i.d., 360 µm o.d.) was applied for CZE separation. The sample (500 ng peptides) was loaded into the capillary using dynamic pH junction stacking strategy. After sample loading, the inlet of the capillary was inserted into 5% acetic acid background electrolyte (BGE) and a voltage of 30 kV was applied on the sample injection end to carry out separations. The separated peptides were detected by Q-Exactive HF mass spectrometer. A Top10 data-dependent acquisition (DDA) method was applied with an isolation window of m/z 2, NCE of 28 for HCD fragmentation. Full scan spectra were acquired with scan range of m/z 300-2000, mass resolution of 60000 (at m/z 200), 1 microscan, AGC of 3E6, and maximum injection time of 50 ms. For the MS/MS, the mass resolution was 60000 (at m/z 200), the number of microscans was 1, AGC target value was 1E5, and the maximum injection time was 100 ms. The ion intensity threshold was 2.5E4 and the dynamic exclusion window was 30 s.

6.2.6 Experimental design and statistical rationale

For the quantitative proteomic experiments, five female brains and five male brains were obtained from the zebrafishes which were at same age (11 months old) and from different mothers. The brain lysates of each gender were pooled into one sample (600 μg proteins) to minimize measurement bias caused by individual difference and manual collection of brains. Both of brain samples (male and female) were divided into two portions: 500 μg for TDP and 100 μg for BUP. For label-free quantitative TDP, each SEC fraction was measured using cIEF-MS/MS in triplicate. The quantitative TDP analysis was conducted to compare proteoform abundance in the matched SEC fractions (i.e., Male SEC fraction 1 vs. Female SEC fraction 1). Considering the proteoform overlaps between adjacent SEC fractions, the quantitative results of specific proteoforms appeared in multiple fractions were manually examined. The proteoforms showed inconsistence on expression changes across SEC fractions were not considered in data analysis. The data normalization and t-test analysis were performed using the Perseus software to indicate proteoforms with statistically significant abundance difference between female and male brains of zebrafish (FDR < 0.05, S0=1). For data normalization, the intensities of each proteoform from the six cIEF-MS/MS runs (3 runs for male and 3 runs for female) were normalized to the intensity of the first cIEF-MS/MS run, converting the proteoform intensity to proteoform ratio. Proteoform ratios of each cIEF-MS/MS run were divided by the corresponding median to make sure the ratios of each run center at 1. The proteoforms with significantly altered expression in male or female brains were sorted out for gene ontology (GO) enrichment analysis using database DAVID Bioinformatics 6.8[7] and enriched terms associated with biological process (BP), molecular function (MF) and cellular components (CC) were identified.

For quantitative BUP, after database search, the reporter ion intensity of the TMT channel (channel 129) was used to normalize the rest of reporter ion intensities of other channels for fold change calculation. Briefly, each individual reporter ion intensity was divided by the corresponding reporter ion intensity of the channel 129, converting the reporter ion intensity to protein ratio. Protein ratios of each TMT channel were divided by the corresponding median to make sure the ratios of each channel center at 1. The Perseus software was employed to generate volcano plot and perform t-test analysis. The differentially expressed proteins between the male and female brains were determined with FDR 0.05 and s0 0.4 using the Perseus software. The DAVID Bioinformatics (6.8) was used to perform GO enrichment analysis.

6.3 Results and discussions

6.3.1 Quantitative top-down proteomics of zebrafish male and female brains

Five male zebrafish brains were pooled and homogenized to reduce heterogeneity between fishes and the extracted protein sample was fractionated by SEC into four fractions. The female fish brains were prepared with the same protocol. The eight SEC fractions (four fractions each gender) were analyzed by the *high-throughput cIEF-MS/MS* in technical triplicate. The relative abundance of proteoforms were compared between female and male brains for each pair of SEC fractions (*i.e.,* male SEC fraction 1 *vs.* female SEC fraction 1) to simplify the quantitative TDP data analysis since the SEC separation was highly reproducible. 171, 1268, 1260, and 741 proteoforms corresponding to 51, 211, 216, and 192 proteins were identified from SEC fraction 1, 2, 3, and 4, respectively. Proteoforms with N-terminal methionine excision, N-terminal truncation or signal peptide cleavage, and several common PTMs, including acetylation (+42 Da), phosphorylation (+80 Da), and methylation (+14 Da) were identified. For instance, we identified a proteoform of *calmodulin* containing an N-terminal methionine excision, an N-terminal acetylation and K115 trimethylation, Figure 6.1A, which was also reported in our previous study of zebrafish brains using CZE-MS/MS.[8] In addition, we identified an N-terminal truncated proteoform of *caveolae-associated protein 4a* with the sequence ranged from Lys273 to Asp329 and it is phosphorylated at Thr292, Figure 6.1B. The phosphorylation at Thr292 was further confirmed by PTM information in the UniProt (https://www.uniprot.org/uniprot/A1L260).

When performing label-free quantification (LFQ), only the proteoforms having reported intensities across the six cIEF-MS/MS runs (triplicate runs per gender) were considered for further abundance comparisons between genders. The feature intensity of selected proteoforms were normalized and compared based on the t-test analysis using an FDR threshold of 0.05 and s0 of 1, as depicted in Figures 6.1C-F**.** Out of the 109, 814, 1089, 569 quantified proteoforms in SEC fractions 1 to 4, we discovered 2, 92, 34, 40 proteoforms showing higher abundance in the corresponding SEC fractions of the female brain sample, while 3, 54, 37, 21 proteoforms presented higher abundance in relevant fractions of the male brain sample. In total, 263 proteoforms showed statistically significant difference in abundance between the male and female brains.

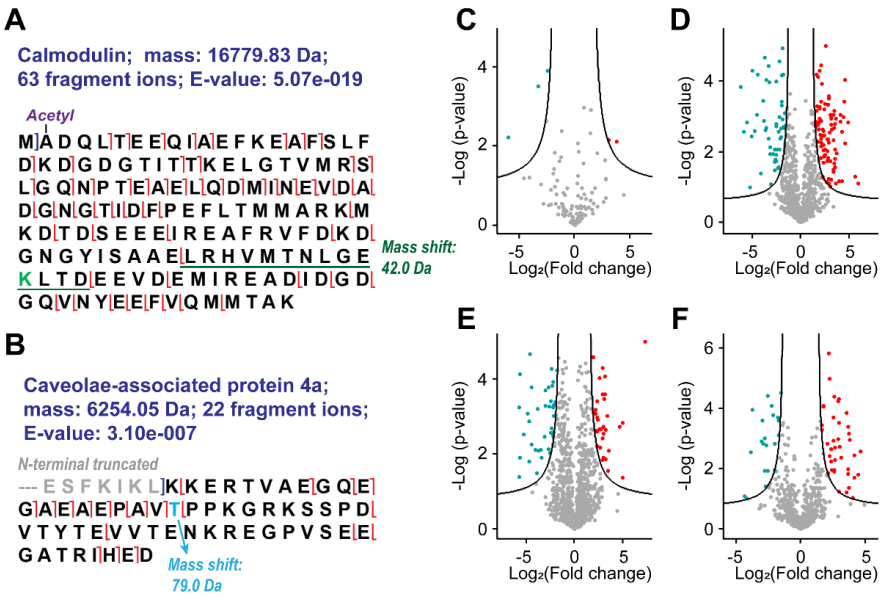

Figure 6.1 Quantitative TDP of four SEC fractions of female and male zebrafish brains using cIEF-MS/MS. (A) Sequence and fragmentation pattern of a proteoform of calmodulin. The sequence underlined with green line has a mass shift of 42.0 Da corresponding to trimethylation at K115. (B) Sequence and fragmentation pattern of a proteoform of cavelolae-associated protein 4a. A mass shift of 79.0 Da at T292 corresponds to a phosphorylation modification. (C)-(F) Volcano plots of -log (p-value) versus $\log_2$ (Fold change, female/male) of quantified proteoforms in SEC fractions 1, 2, 3 and 4 of female and male brains, respectively. The differentially expressed proteoforms were determined by t-test using Perseus with cut-off settings of FDR=0.05 and S0=1. The proteoforms with higher abundance in the female and male brains are highlighted in red and dark cyan color, respectively.

To understand the biological significance of these differentially expressed proteoforms, we performed Gene Ontology (GO) enrichment analysis of genes whose proteoforms showed significantly higher abundance in female and male brains, respectively. We focused on examining the enriched biological process (BP) from 29 annotated genes of female (Figure 6.2) and 34 genes of male (Figure 6.2). In female brains, the enriched BP categories consist of sequestering of actin monomers, histone exchange, neuron projection development, cell proliferation, and actin filament organization, suggesting that these proteoforms involve in neurite outgrowth and neuronal development. Sequestering of actin monomers, as the most enriched BP category, includes *thymosin beta 2* (*Tβ 2*) and *beta thymosin-like protein*. Two proteoforms of the *Tβ 2* and five proteoforms of the *beta thymosin-like protein* showed significantly higher abundance in

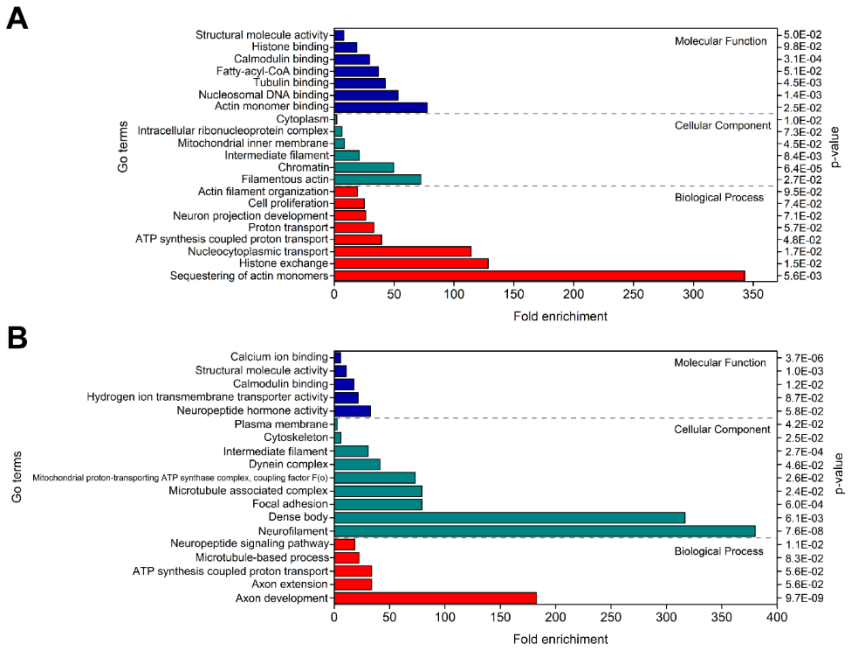

Figure 6.2 GO enrichment analysis of differentially expressed proteoforms in female and male brains of zebrafish. Enriched GO terms of biological process, cellular component, and molecular function associated with the proteoforms with highly expressed in the female brains (A) and the male brains (B) of zebrafish. The values on the right of y-axis denote the p-value of each enriched GO term.

female brains. Studies on *beta-thymosin* of zebrafish have revealed that the protein has monomeric actin binding ability and regulates neuronal growth and differentiation.[9,10] However, the mechanism of how specific proteoforms of *beta-thymosin* involves in sex specific functions of the brains remains unknown. The category of histone exchange includes *acidic leucine-rich nuclear phosphoprotein 32 family member A (ANP32A)* and *acidic leucine-rich nuclear phosphoprotein 32 family member E (ANP32E)*. *APN32A* plays a role in inhibiting acetyltransferase complex in the nucleus, regulating initiation of transcription.[11] *APN32E* is implicated in the removal histone variant H2A.Z via inhibiting protein phosphatase 2A, promoting synaptogenesis.[12-14] Overexpression of N-terminal truncated proteoforms of *APN32A and APN32E in* female brain might play some roles in sex-related regulation of transcription and neuron cell proliferation. *Prothymosin alpha-A* (*PTα-A*) and *prothymosin alpha-B (PTα-B)*, which are enriched in cell proliferation category, have both N-terminal and C-terminal truncated proteoforms identified in our study. *PTα* is an essential nuclear protein, which regulates cell proliferation and protects brain from stroke or traumatic damage by inhibiting cell apoptosis and

neuronal necrosis.[15] In breast cancer MCF7 cells, *PTα* was found to be upregulated by estradiol at both mRNA and protein levels, and gene transcription activity of *PTα* can be altered by estrogen receptor α.[16] Similar data has been observed in neuroblastoma cell, in which the synthesis of *PTα* can be promoted via estradiol treatment.[17] These evidences indicate that the overexpressed proteoforms of *PTα* in the female brains may be associated with estrogen-regulated neural cell proliferation and differentiation.

In male brains, axon development and axon extension were enriched in BP categories, Figure 6.2. Several proteoforms are overexpressed in male brains and their corresponding genes involve in neuronal development. For example*, growth associated protein 43 (Gap43)*, a membrane bound protein, is responsible for axonal outgrowth and elongation.[18] We found a fragment of *Gap43* which was highly expressed in the male brains but not in the female brains, suggesting the expression of *Gap43* might be regulated by hormones. This hypothesis was consistent with previous studies, which showed that the mRNA of *Gap43* was regulated by gonadal hormones and had sex dimorphism.[18,19] Interestingly, we identified several overexpressed proteoforms in the male brains from *pro-opiomelanocortin (POMC), prodynorphin (PDYN),* and *prepronociceptin a (PPNOC),* which are relevant with neuropeptide signaling pathway. Particularly, *POMC* and *PDYN* are important neuropeptide precursors that can be proteolytically cleaved at either paired (such as Lys-Arg or Arg-Arg) or single basic residues to generate endogenous hormone peptides.[20,21] We identified two proteoforms of *POMC* located in the region of N-terminal peptide of *POMC* (NPP, Gln29 to Ser73), which is a potential adrenal growth factor.[22] A proteoform of *POMC* (Ser54 to His105), which contains cleavage sites at His-Lys at C-terminus and Arg-Ser at N-terminus, was identified with 4.6 times higher abundance in the male brains than in the female brains (p-value: $10^{-3.9}$). The other proteoform (Gln29 to Arg53) with N-terminal signaling peptide cleaved was found 2.9 times higher abundance in male brains compared to that in the female brains (p-value: $10^{-2.3}$). Additionally, a proteoform (Asp20 to Val100) of *PDYN* generated from excision of N-terminal signaling peptide and cleavage at Val-Lys at C-terminus showed statistically higher abundance in male brain. A mass shift of +55.06 Da localized in range of Gly81 to Ala85 might be Thr83 to Arg83 mutation (mass shift +55.05 Da). We also identified another *PDYN* proteoform having the same sequence without any mass shift, which showed no statistically significant difference in abundance between male and female brains. Further study will be needed to investigate hormone related biological processes regulated by overexpression of the proteoform of *PDYN* with Thr83-to-Arg83 mutation in male brains.

We noted that ten and four phosphorylated proteoforms showed significantly higher abundance in female and male brains, respectively, including but not limited to proteoforms of

beta thymosin-like protein, *MARCKS-related protein 1-B, thymosin beta 2, calmodulin, and microtubule-associated protein.* The data suggests the potential role of protein phosphorylation in sexual dimorphism.

In summary, we discovered drastic differences in proteoform abundance between male and female zebrafish brains using SEC-cIEF-MS/MS-based label-free TDP. A variety of differentially expressed proteoforms are associated with neuronal development. For example, proteoforms of Tβ 2, beta thymosin-like protein, APN32A, APN32E, PTα-A, PTα-B, stathmin, and microtubule-associated protein were highly expressed in female brains, while proteoforms of neurofilament (medium polypeptide), Gap43, trafficking regulator of GLUT4 (SLC2A4) 1a, and tubulin polymerization-promoting protein family member 2 were highly expressed in the male brains. It has been found that hormones can regulate most of gene expression corresponding to the proteoforms above, and affect multiple cellular processes such as neurogenesis, cell death, and cell differentiation.[23] We speculate that the sex-dependent proteoform expression profile in zebrafish brains could be closely associated with hormone regulation in different genders. Discovering these differentially expressed proteoforms will help us pursue a better understanding of the sex-related neuronal developmental process. Our data demonstrate the value of quantitative TDP in studying sexual dimorphism of brains.

6.3.2 Quantitative bottom-up proteomics of zebrafish male and female brains

We also performed tandem Mass Tag (TMT)-based quantitative BUP of male and female zebrafish brains. We have two goals. First, acquire a comprehensive picture of sex-dependent gene expression outcomes in brain at the protein group level. Second, compare and combine the quantification results of TDP and BUP to pursue a better understanding of the sexual dimorphism of brain. The workflow of TMT quantification is shown in Figure 6.3. In our experiment, we quantified 3811 protein groups from 30738 peptides. The volcano plot was generated with t-test cut-off settings of FDR 0.05 and S0 0.4. We discovered that 67 protein groups were overexpressed in female brains, while 221 protein groups were overexpressed in male brains, Figure 6.3A. GO enrichment analysis of highly expressed protein groups in female indicated several categories associated with neuron growth and brain development, including histone exchange, translational initiation, translation, and cell proliferation, which is consistent with our findings in the top-down study. Overexpressed proteins such as *APN32A, APN32E, PTα-A and PTα-B,* have also been identified to be highly expressed in proteoforms using TDP. Some other highly expressed proteins not annotated in enrichment analysis also drew our attention because they showed drastically higher abundance in female brains. For example, *vitellogenin 1* and *vitellogenin 5* from *vitellogenin* gene family are typical estrogenic biomarkers and showed 10.6

(p-value: $10^{-4.3}$) and 4.6-fold (p-value: $10^{-4.8}$) higher abundance in the female brains. *Coagulation factor XIII (A1 polypeptide a, tandem duplicate 1*), which exhibited 3.3-fold (p-value: $10^{-3.6}$) higher level in female brains, was reported to be greatly upregulated by 17 β-estradiol during embryonic development process.[24] We particularly found multiple hormone-regulated proteins showing significantly higher abundance in male brains than in female brains. These proteins include *hemopexin, antithrombin*, and *lectin (mannose-binding, 1)*, which are associated with cellular response to estrogen stimulus based on GO enrichment analysis. For example, hemopexin, as heme scavenger, maintains iron homeostasis in neurons and prevents heme-mediated oxidative damage.[25] Treatment of zebrafish embryos with estrogen downregulated the expression of hemopexin in the liver at various developmental stages.[24] In our study, the hemopexin showed 2.8 times (p-value: $10^{-5.3}$) higher abundance in male brains, which may be associated with lower level of estrogen.

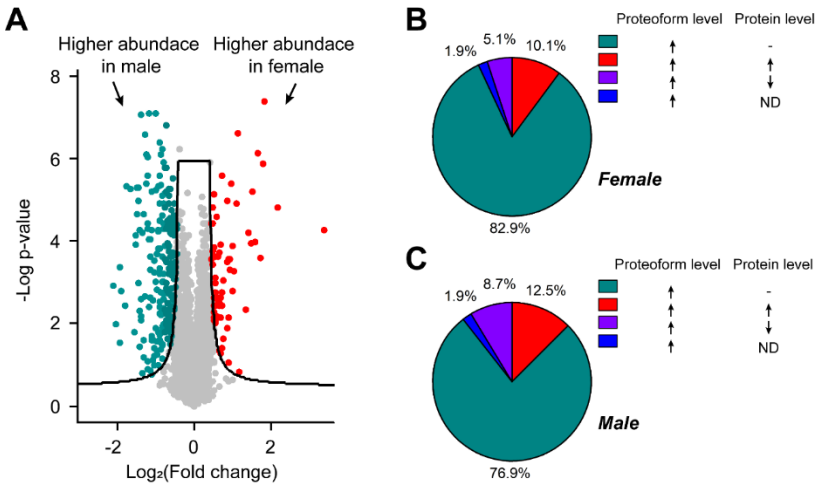

Figure 6.3 Comparison of quantitative BUP and TDP data for achieving overview of gene expression outcome at the protein group and proteoform levels. (A) Volcano plot of protein groups quantified in female and male brains of zebrafish from BUP. The cut-off settings for t-test were FDR=0.05 and S0=0.4. Comparison of quantitative results of female (B) and male (C) brains between TDP and BUP. "ND" means not detected; "-" suggests no significant change in expression level.

When comparing quantitation results of TDP and BUP, we extracted protein accession numbers from the differentially expressed proteoforms from TDP and used them to match with protein groups quantified by BUP to examine whether they were upregulated, downregulated, not differentially expressed, or not identified. Our data revealed that the majority of proteoforms having statistically higher abundance in the female (82.9%) or male brains (77.9%) were not

differentially expressed at the protein group level (Figure 6.3B and 6.3C). For instance, several proteoforms of *beta thymosin-like protein*, *beta-synuclein, thymosin beta 2, calmodulin, pro-opiomelanocortin* and *prodynorphin* with various PTMs from TDP have showed statistically significant difference in abundance between male and female brains. However, the BUP failed to catch these differences. We further analyzed the quantified proteoforms of *calmodulin* and discovered that the summed intensity of the two differentially expressed *calmodulin* proteoforms accounted for only approximately 20% of the total intensity of all the quantified proteoforms. For only 10.1% and 12.5% of the differentially expressed proteoforms, the TDP and BUP data agree. Interestingly, for 5.1% and 8.7% of the differentially expressed proteoforms, TDP and BUP data show opposite expression pattern.

The data of comparing BUP and TDP datasets are very important. First, the results show that combining two quantitative strategies is potentially valuable for generating comprehensive information regarding sexual dimorphism of zebrafish brains since TDP and BUP can provide complementary information on gene expression products. Second, the discrepancies between the BUP and TDP data clearly indicate the importance of delineating proteins in a proteoform-specific manner with TDP for accurately understanding protein function in various biological processes.

6.4 Conclusions

Label-free quantitative TDP of zebrafish male and female brains using the SEC-cIEF-MS/MS quantified thousands of proteoforms and revealed sex-dependent proteoform profiles in brains. We discovered several proteolytic proteoforms of pro-opiomelanocortin and prodynorphin with significantly higher abundance in male brains as potential endogenous hormone proteoforms. Multi-level quantitative proteomics (TDP and BUP) of the brains revealed that majority of proteoforms having statistically significant difference in abundance between genders showed no abundance difference at the protein group level.

6.5 Acknowledgments

We thank Prof. Jose Cibelli's group at the Department of Animal Science of Michigan State University for their help on collecting zebrafish brains for the project. We thank Prof. Xiaowen Liu's group at Indiana University-Purdue University Indianapolis for their help on the top-down proteomics database search using the TopPIC software. We thank the support from the National Institute of General Medical Sciences (NIGMS) through Grant R01GM125991 and the National Science Foundation through Grant DBI1846913 (CAREER Award).

REFERENCES

(1) Arnold, A. P.Sex chromosomes and brain gender. *Nat. Rev. Neurosci.* **2004**, *5*, 701-708.

(2) Di Domenico, F.; Casalena, G.; Sultana, R.; Cai, J.; Pierce, W. M.; Perluigi, M.; Cini, C.; Baracca, A.; Solaini, G.; Lenaz, G.Involvement of Stat3 in mouse brain development and sexual dimorphism: a proteomics approach. *Brain Res.* **2010**, *1362*, 1-12.

(3) Martins-de-Souza, D.; Schmitt, A.; Röder, R.; Lebar, M.; Schneider-Axmann, T.; Falkai, P.; Turck, C. W.Sex-specific proteome differences in the anterior cingulate cortex of schizophrenia. *J. Psychiatr. Res.* **2010**, *44*, 989-991.

(4) Ogata, Y.; Charlesworth, M. C.; Higgins, L.; Keegan, B. M.; Vernino, S.; Muddiman, D. C.Differential protein expression in male and female human lumbar cerebrospinal fluid using iTRAQ reagents after abundant protein depletion. *Proteomics* **2007**, *7*, 3726-3734.

(5) Howe, K.; Clark, M. D.; Torroja, C. F.; Torrance, J.; Berthelot, C.; Muffato, M.; Collins, J. E.; Humphray, S.; McLaren, K.; Matthews, L.The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **2013**, *496*, 498-503.

(6) 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*, 1435-1445.

(7) 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*, 44-57.

(8) McCool, E. N.; Chen, D.; Li, W.; Liu, Y.; Sun, L.Capillary zone electrophoresis-tandem mass spectrometry with ultraviolet photodissociation (213 nm) for large-scale top–down proteomics. *Anal. Methods* **2019**, *11*, 2855-2861.

(9) Roth, L.; Bormann, P.; Bonnet, A.; Reinhard, E.Beta-thymosin is required for axonal tract formation in developing zebrafish brain. *Development* **1999**, *126*, 1365-1374.

(10) van Kesteren, R. E.; Carter, C.; Dissel, H. M.; van Minnen, J.; Gouwenberg, Y.; Syed, N. I.; Spencer, G. E.; Smit, A. B.Local synthesis of actin-binding protein β-thymosin regulates neurite outgrowth. *J. Neurosci.* **2006**, *26*, 152-157.

(11) Wang, S.; Wang, Y.; Lu, Q.; Liu, X.; Wang, F.; Ma, X.; Cui, C.; Shi, C.; Li, J.; Zhang, D.The expression and distributions of ANP32A in the developing brain. *Biomed Res. Int.* **2015**, *2015*.

(12) Costanzo, R. V.; Vilá-Ortíz, G. J.; Perandones, C.; Carminatti, H.; Matilla, A.; Radrizzani, M.Anp32e/Cpd1 regulates protein phosphatase 2A activity at synapses during synaptogenesis. *Eur. J. Neurosci.* **2006**, *23*, 309-324.

(13) Obri, A.; Ouararhni, K.; Papin, C.; Diebold, M.-L.; Padmanabhan, K.; Marek, M.; Stoll, I.; Roy, L.; Reilly, P. T.; Mak, T. W.ANP32E is a histone chaperone that removes H2A. Z from chromatin. *Nature* **2014**, *505*, 648-653.

(14) Shin, H.; He, M.; Yang, Z.; Jeon, Y. H.; Pfleger, J.; Sayed, D.; Abdellatif, M.Transcriptional regulation mediated by H2A. Z via ANP32e-dependent inhibition of protein phosphatase 2A. *Biochim. Biophys. Acta - Gene Regul. Mech.* **2018**, *1861*, 481-496.

(15) Fujita, R.; Ueda, M.; Fujiwara, K.; Ueda, H.Prothymosin-α plays a defensive role in retinal ischemia through necrosis and apoptosis inhibition. *Cell Death Differ.* **2009**, *16*, 349-358.

(16) Bianco, N. R.; Montano, M. M.Regulation of prothymosin α by estrogen receptor α: molecular mechanisms and relevance in estrogen-mediated breast cell growth. *Oncogene* **2002**, *21*, 5233-5244.

(17) Ciana, P.; Ghisletti, S.; Mussi, P.; Eberini, I.; Vegeto, E.; Maggi, A.Estrogen receptor α, a molecular switch converting transforming growth factor-α-mediated proliferation into differentiation in neuroblastoma cells. *J. Biol. Chem.* **2003**, *278*, 31737-31744.

(18) Lustig, R. H.; Sudol, M.; Pfaff, D. W.; Federoff, H. J.Estrogenic regulation and sex dimorphism of growth-associated protein 43 kDa (GAP-43) messenger RNA in the rat. *Mol. Brain Res.* **1991**, *11*, 125-132.

(19) Shughrue, P. J.; Dorsa, D. M.The ontogeny of GAP-43 (neuromodulin) mRNA in postnatal rat brain: Evidence for a sex dimorphism. *J. Comp. Neurol.* **1994**, *340*, 174-184.

(20) Benjannet, S.; Rondeau, N.; Day, R.; Chretien, M.; Seidah, N.PC1 and PC2 are proprotein convertases capable of cleaving proopiomelanocortin at distinct pairs of basic residues. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 3564-3568.

(21) Day, R.; Lazure, C.; Basak, A.; Boudreault, A.; Limperis, P.; Dong, W.; Lindberg, I.Prodynorphin processing by proprotein convertase 2 cleavage at single basic residues and enhanced processing in the presence of carboxypeptidase activity. *J. Biol. Chem.* **1998**, *273*, 829-836.

(22) Bicknell, A. B.60 YEARS OF POMC: N-terminal POMC peptides and adrenal growth. *J. Mol. Endocrinol.* **2016**, *56*, T39-T48.

(23) Cooke, B.; Hegstrom, C. D.; Villeneuve, L. S.; Breedlove, S. M.Sexual differentiation of the vertebrate brain: principles and mechanisms. *Front. Neuroendocrinol.* **1998**, *19*, 323-362.

(24) Hao, R.; Bondesson, M.; Singh, A. V.; Riu, A.; McCollum, C. W.; Knudsen, T. B.; Gorelick, D. A.; Gustafsson, J.-Å.Identification of estrogen target genes during zebrafish embryonic development through transcriptomic analysis. *PLoS One* **2013**, *8*, e79020.

(25) Hahl, P.; Davis, T.; Washburn, C.; Rogers, J. T.; Smith, A.Mechanisms of neuroprotection by hemopexin: modeling the control of heme and iron homeostasis in brain neurons in inflammatory states. *J. Neurochem.* **2013**, *125*, 89-101.

CHAPTER 7.* Application of CZE-MS/MS-based multidimensional platforms for uncovering proteoform-level differences between metastatic and nonmetastatic colorectal cancer cells

7.1 Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and has a high mortality rate even with recent improvements in therapies.[1,2] CRC metastasis is the main cause of CRC-related death. New insights into the molecular mechanisms of CRC metastasis will undoubtedly be beneficial for developing more effective drugs.[3-5] Extensive studies have been completed with the goal of understanding CRC metastasis at the transcriptome level, generating tremendous information about the landscape of mRNA across different stages of CRC. [6,7] However, nucleic-acid–based measurements do not correlate well with protein abundance, which are the primary effectors of function in biology.[8] Quantitative bottom-up proteomics (BUP) studies of metastatic and non-metastatic CRC cell lines have discovered new protein regulators involved in CRC metastasis.[4,9,10] BUP usually provides limited information on the proteoforms, which represent all possible protein molecules derived from the same gene resulting from genetic variations, RNA alternative splicing, and protein post-translational modifications (PTMs).[11,12] Mass spectrometry (MS)-based top-down proteomics (TDP) directly measures intact proteoforms and provides opportunities to study functions of specific proteoforms.[13,14] Unfortunately, there is still no report in the literature about studying CRC metastasis using TDP, and this study will help to fill that gap.

Here, we performed the first deep TDP study of metastatic (SW620) and non-metastatic (SW480) human CRC cell lines, aiming to produce a comprehensive proteoform-level view of the two isogenic CRC cell lines and discover novel proteoform biomarkers of CRC metastasis. We employed four different capillary zone electrophoresis (CZE)-tandem MS (MS/MS) approaches, 1-D CZE-MS/MS, 2-D size exclusion chromatography (SEC)-CZE-MS/MS, 2-D reversed-phase liquid chromatography (RPLC)-CZE-MS/MS, and 3-D SEC-RPLC-CZE-MS/MS analyses of the two cell lines for proteoform identification (ID) and label-free quantification (LFQ), Figure 7.1. For 1-D CZE-MS/MS, each sample was analyzed by CZE-MS/MS in technical triplicate. For 2-D SEC-

---

* This chapter is adapted with permission from *McCool, E. N.†; Xu, T.†; Chen, W.†; Beller, N. C.; Nolan, S. M.; Hummon, A. B.; Liu, X.; Sun, L. Science Advances 2022, 8 (51), eabq6348.*

CZE-MS/MS, each sample was fractionated by SEC into 6 fractions, followed by CZE-MS/MS in technical triplicate. For 2-D RPLC-CZE-MS/MS, we fractionated each sample to 6 or 13 fractions by RPLC and analyzed each LC fraction by single-shot CZE-MS/MS (RPLC 13 fractions) or triplicate CZE-MS/MS measurements (RPLC 6 fractions). For 3-D SEC-RPLC-CZE-MS/MS, 52 LC fractions were collected for each sample, followed by CZE-MS/MS in technical triplicate. From 1-D separation to 3-D separations, the required amount of starting protein materials increased (from 100 μg to 2 mg) due to the unavoidable sample loss during sample collections and transfers. The TopPIC (version 1.4.0) software was used for data analysis,[15] and a 1% proteoform-level false discovery rate (FDR) was used to filter the database search results.

7.2 Experimental section

7.2.1 Materials and reagents

MS-grade water, acetonitrile (ACN), methanol (MeOH), formic acid (FA) and HPLC-grade acetic acid (AA) were purchased from Fisher Scientific (Pittsburgh, PA). Ammonium bicarbonate (NH4HCO3), urea, dithiothreitol (DTT), iodoacetamide (IAA) and 3-(trimethoxysilyl)propyl methacrylate were from Sigma-Aldrich (St. Louis, MO). Hydrofluoric acid (HF, 48-51% solution in water) and acrylamide were purchased from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail (EASYpacks) was from Roche (Indianapolis, IN).

7.2.2 Sample preparation

SW480 (catalogue number CCL-228) and SW620 (catalogue number CCL-227) original cell lines were both purchased from ATCC (Manassas, VA) and were cultured in RPMI 1640 cell culture medium (Life Technologies Corporation, Grand Island, NY) supplemented with 10% fetal bovine serum (Thermo Scientific, Gaithersburg, MD) and 2mM L-glutamine (Invitrogen, San Diego, CA). The cells were incubated at 37°C with 5% CO2 and were passaged every 3-4 days. Both cell lines were last verified by Short Tandem Repeat (STR) sequencing in 2016 and were used within two months after resuscitation from frozen aliquots at -80°C.

Upon growing to confluency, cells were harvested and cleansed of remaining cell culture medium via subsequent washing with HPLC grade water (Fisher Scientific, Pittsburgh, PA) and centrifugation for 5-minute intervals at 15000 × g until supernatant was clear. Proteins were then extracted using mammalian cell lysis buffer. Cell lysis buffer consisted of 8 M urea, 50 mM Tris (pH 8.2), 1 mM β-glycerophosphate, 1 mM phenylmethylsulfonyl fluoride, 75 mM sodium chloride, 1 mM sodium fluoride, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate, and one protease inhibitor cocktail. The reagents for cell lysis buffer were purchased from Sigma-Aldrich

and complete EDTA-free protease inhibitor cocktail tablet was purchased from Roche. Lysis buffer was added to the harvested cells which then underwent sonication on ice three times for 1-minute intervals at 15% amplitude. The resulting extracted proteins were then clarified of cellular debris by centrifugation at 15,000 rpm for 10 minutes. Proteins were quantified using a bicinchoninic acid (BCA) protein assay (Thermo Scientific Pierce, Rockford, IL) and then stored at -80°C until preparation for MS analysis.

SW480 and SW620 proteins were denatured at 37 °C for 30 minutes, reduced at 37 °C for 30 minutes using DTT, and then alkylated at room temperature in the dark for 20 minutes using IAA. The excess IAA were quenched by adding DTT and reacting for 5 min at room temperature.

For the experiment 1 (RPLC-CZE-MS/MS), 200 µg of proteins from SW480 and SW620 cells were reduced, alkylated, and acidified, followed by RPLC fractionation into 13 fractions and CZE-MS/MS. For the experiment 2 (SEC-RPLC-CZE-MS/MS), 2 mg of proteins from SW480 and SW620 cells were reduced and alkylated before fractionated by SEC-RPLC and analyzed by CZE-MS/MS. For the experiment 3 (RPLC-CZE-MS/MS), 420 µg of proteins from SW480 and SW620 cells were reduced and alkylated prior to fractionation by RPLC into 6 fractions and analyses by CZE-MS/MS. For the experiment 4 (SEC-CZE-MS/MS), the samples were desalted after reduction and alkylation using a C4 trap column (4×10 mm, 3 µm particles, 300 Å pore size). Specifically, 500 µg of proteins from SW480 and SW620 cells was loaded onto the column and flushed with mobile phase A (2% (v/v) ACN, 0.1% FA) for 10 minutes at a flow rate of 1 mL/min. The proteins were eluted with mobile phase B (80% ACN, 0.1% FA) for 3 minutes at flow rate of 1 mL/min. The eluates were lyophilized with a speed vacuum and redissolved in 150 µL 0.1% formic acid (FA). Then proteins from SW480 and SW620 cells were fractionated by SEC into 6 fractions, followed by CZE-MS/MS analyses. For the experiment 5 (1D-CZE-MS/MS), 100 µg of proteins from SW480 and SW620 cells were desalted using two methods. In one case, both samples were desalted by a C4 trap column as described in the experiment 4. In the other case, both samples were desalted by Amicon Ultra centrifugal filters with a molecular weight cutoff of 10 kDa. Desalting with centrifugal filter was performed by loading 100 µg of proteins onto the filter and washing the sample four times with 50 mM NH4Ac at 14,000 × g. Finally, the sample was recovered in 30 µL of 50 mM NH4Ac. The samples desalted with the C4 trap column and centrifugal filters were analyzed by 1D-CZE-MS/MS in technical triplicate.

7.2.3 Fractionation of the SW480 and SW620 proteome

All separations were performed on a 1260 Infinity II HPLC system from Agilent (Santa Clara, CA). Detection was performed using a UV-visible detector at a wavelength of 254 nm. Data was collected and analyzed using OpenLAB software. RPLC (C4, 2.1 × 250 mm, Sepax

Technologies) and SEC (4.6 × 300 mm, 500 Å pores, Agilent) were performed offline (Agilent HPLC) for prefractionation. Fractions from SW620 and SW480 from experiment 1 (13 fractions × 2 samples), experiment 2 (84 fractions × 2 samples), experiment 3 (6 fractions × 2 samples), and experiment 4 (6 fractions × 2 samples) were analyzed by CZE-MS/MS, respectively.

In experiment 1, RPLC was used for sample fractionation with a 0.25 mL/min flow rate and gradient of 0-80% mobile phase (MP) B over 90 minutes (MPA: 2% ACN, 0.1% FA in water; MPB: 80% ACN, 0.1% FA in water). Fractions were collected from 15 to 22 minutes (fraction 1) and 22 to 70 minutes (12 fractions, 4 minutes per fraction). For experiment 2, both SEC and RPLC were used for fractionation prior to CZE-MS/MS. For SEC, the flow rate was 0.35 mL/min with a 0.05% TFA mobile phase. 2 mg of proteins in 800 µL solution was fractionated by SEC. Fractions were collected from 5-8 minutes (fraction 1) and 8-12.5 minutes (3 fractions, 1.5 minutes per fraction). One RPLC run was performed for each SEC fraction with a flow rate of 0.25 mL/min and gradient of 0-80% MPB (MPA: 2% ACN, 0.1% TFA in water; MPB: 10% IPA, 0.1% TFA in ACN) over 90 minutes with a 10-minutes equilibration with 100% MPA at the beginning of the separation. Fractions were collected from 20 to 25 minutes (fraction 1) and 25 to 65 minutes (20 fractions, 2 minutes per fraction). In experiment 3, RPLC fractionation was carried out using the same mobile phases as in experiment 1, and a 90-minute gradient was used with a 10-minute equilibration with 100% MPA at the beginning of the separation. Fractions were collected from 25 to 55 minutes (fraction 1), 50 to 70 minutes (4 fractions, 5 minutes per fraction), and 70 to 95 minutes (fraction 6). In experiment 4, SEC fractionation was performed with an Agilent Bio SEC-5 column (4.6 × 300 mm, 5 µm particles, 500 Å pore size). 220 µg of SW480 and SW620 proteins (1.5 mg/mL, 75 µL×2 injections) were loaded into the SEC column and separated isocratically at the flow rate of 0.3 mL/min with 0.1% FA as mobile phase. The first fraction is collected from 5.6 to 8.6 minutes. The second to the fifth fraction was from 8.6 to 14.6 minutes with 1.5 minutes per fraction. The final fraction was collected from 14.6 to 19.0 min. In the experiments 1-4, samples were dried down and redissolved in 50 mM NH4HCO3 (pH 8.0, ~2 mg/mL) for CZE-ESI-MS/MS.

7.2.4 CZE-MS/MS analysis

CZE separation was performed using a CESI 8000 Plus CE system (Beckman Coulter). A commercialized electrokinetically pumped sheath-flow CE-MS nanospray interface (CMP Scientific Corp) was applied for online coupling the CE system and mass spectrometer.[64,65] A glass emitter (orifice size: 20~30 µm) installed on the interface was filled with sheath buffer (0.2% FA, 10% methanol) to generate electrospray at voltage of 2-2.3 kV.

A 100 cm LPA coated fused silica capillary (50 µm i.d., 360 µm o.d.) was used for CZE separation in experiments 1, 2, 4 and 5, while a 70 cm LPA coated capillary (50 µm i.d., 360 µm

o.d.) was employed for separation in experiment 3. The inner wall of the capillary was coated with LPA. One end of the capillary was etched with HF to reduce the outer diameter of the capillary to about 70-80 μm. (Caution: use appropriate safety procedures while handling hydrofluoric acid solutions)

In experiments 1, 2, 4 and 5, the capillary (100 cm) was loaded with 500 nL of sample. In experiment 3, the capillary (70 cm) was loaded with ~350 nL of sample. After sample loading, the capillaries were inserted into background electrolyte, containing 5% acetic acid (pH 2.4), and 30 kV voltage was applied at the sample injection end to carry out separations.

MS1 and MS2 data were collected on a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) under data-dependent acquisition (DDA) mode. The temperature of ion transfer tube was set to 320 °C and s-lens RF was 55. MS1 spectra were collected with following parameters: m/z range of 600-2000, mass resolution of 120,000 (at m/z 200), a microscan number of 3, AGC target value of 1E6, and maximum injection time of 100 ms. The top 5 most abundant precursor ions (charge state higher than 5, or charge state unassigned and intensity threshold 2E4) in the MS1 spectra were isolated with a window of 4 m/z and fragmented via HCD with NCE of 20%. The settings for MS2 spectra were resolution of 120,000 (at m/z 200), a microscan number of 3, AGC target value of 1E5, and maximum injection time of 200 ms. The dynamic exclusion was set to a duration of 30s and the isotopic peaks were excluded.

In experiments 2, 3, 4 and 5, each LC fraction was analyzed by CZE-MS/MS in triplicate. In experiment 1, each LC fraction was analyzed by a single CZE-MS/MS run. In total, 410 MS raw files with good protein signals were produced from experiments 1, 2, 3, and 4 for database search, including 26 MS raw files from experiment 1 (13 fractions × 2 samples), 312 MS raw files from experiment 2 (52 fractions × 2 samples × 3 replicates), 36 MS raw files from experiment 3 (6 fractions × 2 samples × 3 replicates), and 36 MS raw files from experiment 4 (6 fractions × 2 samples × 3 replicates). We need to note that we collected 84 fractions × 2 samples in the experiment 2. However, we only observed good protein signals from 52 LC fractions per sample. 12 MS RAW files were collected from the experiment 5 using CZE-MS/MS.

7.2.5 Data analysis for proteoform identification

All RAW files were analyzed with the TopPIC Suite (version 1.4.0) pipeline. The RAW files were converted into mzML files with msconvert. Then spectral deconvolution was performed with TopFD (version 1.4.0), which converts precursor and fragment isotope clusters into neutral monoisotopic masses and finds proteoform features by combining precursor isotope clusters with similar monoisotopic masses and close migration times in MS1 scans. The resulting mass spectra with monoisotopic neutral masses were stored in msalign files and the proteoform feature

information was stored in text files. The human proteome database was downloaded from UniProt (UP000005640, 20350 entries, version October 23, 2019, only reviewed protein sequences were included) and concatenated with a random decoy database of the same size. Each msalign file was searched against the concatenated targe-decoy database using TopPIC (version 1.4.0). Cysteine carbamidomethylation was set as a fixed modification, and the maximum number of unexpected modifications was 1. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da. TopPIC reported a list of target and decoy proteoform-spectrum-matches (PrSMs) for each msalign file.

The proteoforms identified from all msalign files were merged and filtered with a proteoform-level FDR. First, the target and decoy PrSMs reported from all the msalign files were combined and filtered with a 5% spectrum-level FDR. The PrSMs were then clustered by grouping PrSMs into the same cluster if they were from the same protein and their precursor mass differences were not large than 2.2 Da. The PrSM with the best E-value was selected for each cluster and its proteoform was reported as the representative one for the cluster. The representative target and decoy proteoforms were finally filtered with a 1% proteoform-level FDR. (The database search was performed with the help of Prof. Xiaowen Liu lab)

7.2.6 Proteoform quantification

There were 18 MS raw files from triplicate CZE-MS/MS analyses of the 6 SEC fractions for the SW480 or SW620 sample in experiment 4. The TopPIC suite pipeline reported a list of targe and decoy PrSM identifications for each raw file. Using the methods in the previous section, the PrSM identifications of the 36 MS raw files were merged and a list of proteoform identifications with a 1% proteoform-level FDR were reported. The abundance of a proteoform was computed as the sum of the proteoform abundances in the six SEC fractions, which were reported by TopFD. Proteoform identifications and their abundances were reported for each replicate using this method. Finally, TopDiff (version 1.4.0), a tool in TopPIC Suite, was used to match proteoform identifications across the three SW480 replicates and three SW620 replicates.

The quantitative results were further analyzed using Perseus software. The intensities of each proteoform in triplicate CZE-MS/MS runs of SW480 and SW620 were normalized to the intensity of corresponding proteoform from the first run of SW480, converting proteoform intensity to proteoform ratio. Then, proteoform ratios of each run were divided by the corresponding median to make sure the ratios center at 1. After log2 transformation of all the data, the significantly differentially expressed proteoforms were determined by performing t-test analysis (FDR threshold: 0.05, S0: 1) using the Perseus software. The volcano plot [-log(p-value) vs. log2(fold change)] was generated.

7.2.7 Proteogenomic analysis

To generate sample-specific protein sequence databases with genetic variations for SW480 and SW620 cells, two RNA-Seq data sets (SRR8616059 for SW480 and SRR8615459 for SW620) were downloaded from the Sequence Read Archive (SRA). The GATK pipeline was employed to align short reads in the RNA-Seq data with the hg38 human genome to call single nucleotide variants (SNVs) and indels, which were further annotated using the gene-based annotation of ANNOVAR (April 16, 2018). The annotated nonsynonymous SNVs and indels in exons were chosen for generating sample-specific protein sequence databases based on the basic annotation of the hg38 human genome in GENCODE. Two sample-specific protein sequence databases were generated using TopPG (version 1.0): one for SW480 cells and the other for SW620 cells. Each protein sequence database contained both reference protein sequences in the basic annotation of GENCODE and protein sequences with sample-specific variants. There were 74887 entries with 51485 reference sequences and 23402 sequences with variants in the database for SW480 cells and 75665 entries with 51432 reference sequences and 24233 sequences with sample-specific variants in the database for SW620 cells. The SW480 and SW620 mass spectra in experiments 3 and 4 were searched against their corresponding sample-specific database using TopPIC (version 1.4.0) with the same parameter setting in Section “Data analysis for proteoform identification”. Using the methods in Section “Data analysis for proteoform identification”, PrSMs identified in each cell line were combined and clustered, and proteoform identifications were filtered by a 5% proteoform-level FDR. Identifications with single amino acid variant (SAAV) sites were manually inspected. If a proteoform with SAAV sites contained no unexpected mass shifts or had at least three matched fragment ions between each SAAV site and the unexpected mass shift, it was reported as a confident proteoform identification with SAAV sites. (Proteogenomic analysis was performed with the help of Prof. Xiaowen Liu lab)

7.2.8 QIAGEN ingenuity pathway analysis (IPA)

The cancer-related network analysis results shown in Figure 7.4F, Figure 7.5E, and Figure 7.5F were generated through the use of QIAGEN IPA (QIAGEN Inc., https://digitalinsights.qiagen.com/IPA). Permissions have been granted by QIAGEN to use those copyrighted figures in this publication.

7.2.9 Statistical analysis

Data are presented as mean±standard deviations when available. For the statistical analysis of LFQ data of SW480 and SW620 cell lines, we performed both side t-test using the Perseus software to determine the proteoforms with statistically significant abundance difference between the two cell lines with the following settings, S0=1 and FDR = 0.05.

7.3 Results and discussions

7.3.1 Identification of over 23,000 proteoforms from CRC cells using CZE-MS/MS

One long-term goal of TDP is to characterize all the millions of proteoforms in the human body.[16,17] During the last decade, because of the improvement of proteoform sample preparation, LC and CZE separations, MS and MS/MS, 3,000-5,000 proteoforms corresponding to roughly 1,000 genes can be identified from one human cell line using LC-MS/MS-based platforms,[18-22] and up to 6,000 proteoform IDs corresponding to 850 genes have been reported from an *E. coli* sample using a CZE-MS/MS-based workflow.[23] Only one TDP study of a human cell line using CZE-MS/MS was reported with the identification of about 500 proteoforms.[24] Recently, the Kelleher group reported the identification of ~30,000 proteoforms of 1,690 human genes from 21 human cell types and plasma using RPLC-MS/MS-based strategies, representing a milestone in large-scale TDP.[21] On average, nearly 3,000 proteoforms were identified from one of the 21 human cell types.

In this work, we performed the first global TDP study of a pair of isogenic human non-metastatic and metastatic CRC cell lines (SW480 and SW620). Four different strategies were employed, Figure 7.1. We first compared the four different CZE-MS/MS strategies listed in Figure 7.1B in terms of the number and efficiency of proteoform IDs from the SW480 cells, Figure 7.2A. SEC-RPLC-CZE-MS/MS outperformed SEC-CZE-MS/MS, RPLC-CZE-MS/MS, and CZE-MS/MS in terms of the number of proteoform IDs due to better LC fractionation (2-D LC vs. 1-D or no LC) and much more CZE-MS/MS runs (52 vs. 6 and 13). In terms of the proteoform identification efficiency (the number of proteoform IDs per CZE-MS/MS run), the SEC-CZE-MS/MS (6 LC fractions) produced nearly 700 proteoform IDs per run, which is nearly 6-fold and 4-fold higher than those from SEC-RPLC-CZE-MS/MS and CZE-MS/MS, respectively. We drew two conclusions from the data. First, multi-dimensional separation is crucial for large-scale TDP analysis of human cell lysates due to their extremely high complexity. Second, SEC-CZE-MS/MS and RPLC-CZE-MS/MS under an optimized condition are powerful techniques for deep TDP of human cell lysates with high throughput.

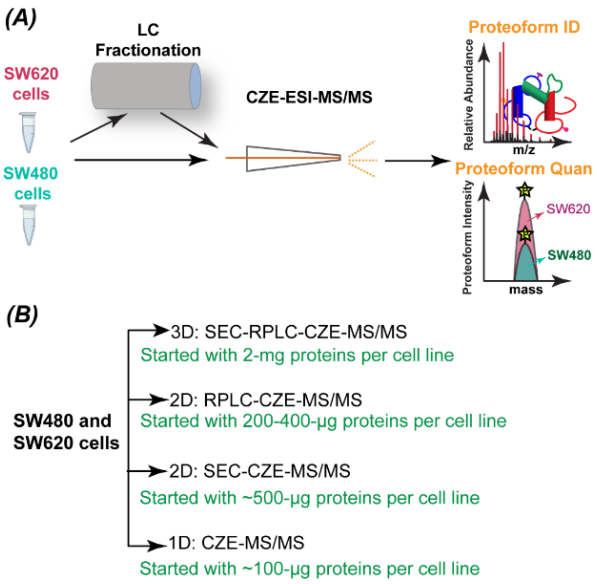

Figure 7.1 Schematic of the experimental design. (A) Schematic design of the TDP study of metastatic (SW620) and non-metastatic (SW480) CRC cells using CZE-ESI-MS/MS and LC-CZE-ESI-MS/MS for proteoform identification and label-free quantification. (B) Four CZE-MS/MS-based strategies in this work with the amounts of protein starting materials.

In total, we collected over 400 MS raw files using the four CZE-MS/MS-based strategies and identified 23,622 proteoforms of 2,332 proteins from the SW480 and SW620 cell lines with a 1% proteoform-level FDR. The number of proteoform IDs from the CRC cells is about 5-8 fold higher than that reported in previous TDP studies of human cancer cells (23,622 vs. 3,000-5,000 proteoforms).[18-20] 17,316 and 14,504 proteoforms (on average 15,910 proteoforms) were identified from SW480 and SW620 cell lines, respectively, representing about 3-fold improvement in the number of proteoform IDs per human cell line compared to previous LC-MS/MS-based TDP datasets. The number of proteoform IDs is about 30-fold higher than previous human cell TDP datasets by CZE-MS/MS (~16,000 vs. ~500).[24] Figure 7.2B shows the number of proteoform IDs per complex sample using TDP in previous works and this study.[18-23]

We need to point out that the nearly 16,000 proteoform IDs from SW480 or SW620 cells combine the results of four different CZE-MS/MS-based strategies and about 200 CZE-MS/MS runs. The previous literature studies typically employ one LC-MS/MS or CZE-MS/MS-based approach.[18-23] We also included the data of SW480 and SW620 cells from only SEC-CZE-MS/MS in Figure 7.2B. A total of 5,855 and 6,273 proteoforms (mean±standard deviation: 6,064±296) were identified from SW480 and SW620 cells, respectively, by SEC-CZE-MS/MS, via 18 CZE-MS/MS runs (6 SEC fractions × 3 CZE-MS/MS runs/fraction). The SEC-CZE-MS/MS produced significantly higher proteoform IDs (6,000 vs. 3,000-5,000) from a single human cell line than LC-

MS/MS-based approaches in the literature with a drastically lower number of MS runs (18 vs. 40-800).

The data clearly demonstrate the power of our CZE-MS/MS-based TDP strategy for comprehensive characterization of proteoforms in complex proteome samples. We attribute the drastic improvement of proteoform IDs to the high separation efficiency of CZE for proteoforms,[25] high sensitivity of CZE-MS for proteoform detection,[25-27] and high orthogonality of LC and CZE for biomolecule separations.[23,28] The features of CZE-MS/MS for TDP have been systematically reviewed recently.[29,30]

We further compared the proteoforms and proteins identified from the SW480 and SW620 cells using the SEC-CZE-MS/MS data. Figure 7.2C shows the heat map of proteoform overlaps among technical replicates of SW480 and SW620 cells. About 60-70% of proteoforms identified in one technical replicate of SW480 or SW620 cells were also identified in another replicate of the same cell line, indicating reasonable reproducibility of proteoform ID using SEC-CZE-MS/MS and the data-dependent acquisition mode. Interestingly, only about 40-50% of proteoforms identified in one replicate of SW480 cells (e.g., SW480_1) were identified in one replicate of SW620 cells (e.g., SW620_1). The proteoform overlaps in Figure 7.2C between the two cell lines are statistically significantly lower than that within each cell line (44±4% vs. 67±4%, $p<10^{-14}$, two-tailed student's t-test). The data clearly demonstrate that the pair of isogenic human non-metastatic (SW480) and metastatic (SW620) CRC cell lines have significantly different proteoform profiles. The two cell lines are also significantly different at the protein level. The difference in protein overlaps between the two cell lines and within each cell line is statistically significant (69±8% vs. 83±3%, $p<10^{-6}$, two-tailed student's t-test).

TDP has some technical challenges for the identification of large proteoforms (i.e., >30 kDa). In this work, we focused on the characterization of proteoforms smaller than 30 kDa using a Thermo Q-Exactive HF mass spectrometer. The majority of identified proteoforms in SW480 and SW620 cells are 10 kDa or smaller, which is one main limitation of this study. It is worth noting that 1600-2200 proteoforms have masses larger than 10 kDa. Figure 7.2D shows the sequences and fragmentation patterns of two example proteoforms. Those two proteoforms were identified with high confidence and were also well characterized with N-terminal methionine removal and N-terminal acetylation.

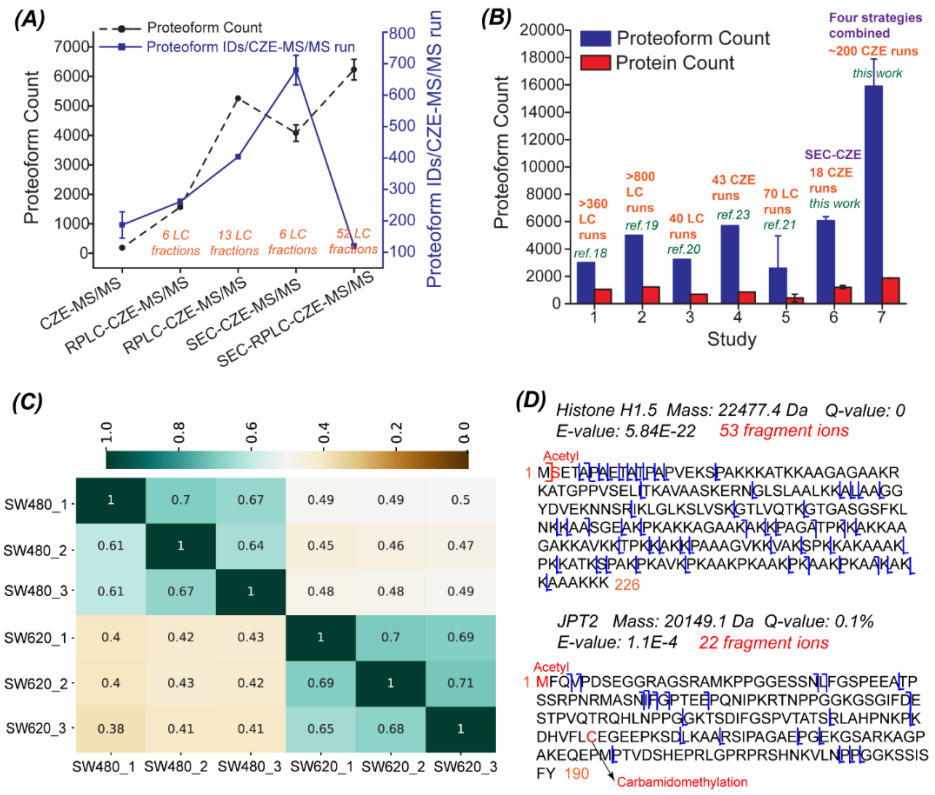

Figure 7.2 Summary of proteoform identification results of this study. (A) Proteoform IDs from SW480 cells using different CZE-MS/MS-based strategies. The error bars represent the standard deviations of the number of proteoform IDs from technical triplicates. (B) The number of proteoform and protein IDs per complex proteome sample using RPLC- or CZE-MS/MS-based TDP strategies. The data of studies 5, 6 and 7 are shown as mean ± standard deviations from various proteome samples. (C) Heat map of proteoform overlaps from technical triplicates of SW480 and SW620 cells using SEC-CZE-MS/MS. Each number in the figure represents a ratio between the number of shared proteoforms in two conditions (e.g., SW480_1 (x-axis) and SW620_1 (y-axis)) and the total number of identified proteoforms in one of the two conditions listed on the y-axis (e.g., SW620_1). For example, the proteoform overlap between SW480_1 (x-axis) and SW620_1 (y-axis) is 0.4, which indicates the ratio between the number of shared proteoforms in those two conditions and the total number of identified proteoforms in SW620_1. (D) Sequences and fragmentation patterns of identified example proteoforms in the study.

7.3.2 Proteoforms of important genes in well-known CRC-related pathways

We further performed QIAGEN Ingenuity Pathway Analysis (IPA) analysis of the genes identified in this work by the four CZE-MS/MS-based strategies and determined several significantly enriched and well-known CRC-related pathways, including WNT/β-catenin Signaling (p-value: 10-3), PI3K/AKT Signaling (p-value: 10-4), mTOR Signaling (p-value: 10-14), and

ERK/MAPK Signaling pathways (p-value: 10-4).[31,32] Those pathways play critical roles in CRC progression via regulating cell proliferation, apoptosis, survival and etc. We identified hundreds of proteoforms from dozens of genes for each pathway, Figure 7.3A. The lists of proteoforms are shown in Supplementary Material II. Comparable numbers of proteoforms were identified from SW480 and SW620 cells for PI3K/AKT Signaling, mTOR Signaling, and ERK/MAPK Signaling pathways. An obviously higher number of proteoforms was obtained from SW480 cells compared to SW620 cells for the WNT/β-catenin Signaling pathway (511 vs. 340). Combination of the data from SW480 and SW620 cells produced about 40% more proteoforms related to the four CRC pathways compared to one cell line alone, indicating the potential differences in proteoform profiles for the well-known CRC-related pathways between the non-metastatic and metastatic CRC cell lines. As shown in Figure 7.3B, the shared proteoforms between SW480 and SW620 cells for each pathway is only about 21%-38% of the total proteoforms identified from the two cell lines. The data suggest that proteoforms in those pathways could potentially play important roles in driving CRC progression and metastasis.

We highlighted some proteoforms of important genes (MARK2, SOX9, EIF4B, and EIF4EBP1) related to the WNT/β-catenin Signaling, mTOR Signaling, and PI3K/AKT Signaling pathways in Table 7-1. MARK2 plays vital roles in modulating directional cancer cell migration, which is crucial for cancer metastasis.[33] SOX9 is a high mobility group (HMG) box transcription factor and plays essential roles in regulating CRC progression.[34] Expression of SOX9 is closely associated with the 5-year overall survival rate of CRC patients.[34] EIF4B regulates cancer cell proliferation and has been reported as a potential target for developing anti-cancer therapies.[35] Phosphorylation of EIF4EBP1 has been reported as an important regulator of cancer progression.[36]

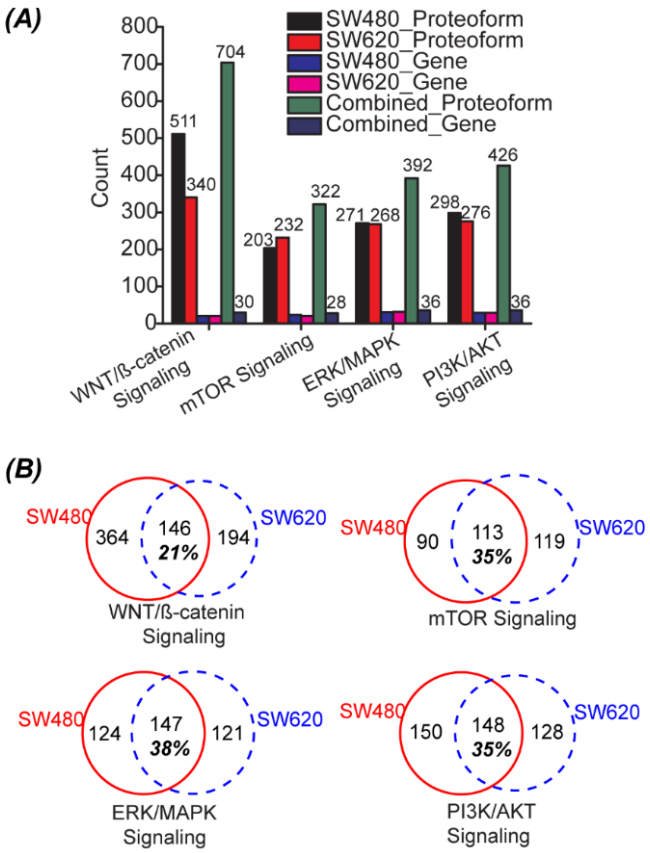

Figure 7.3 Summary of proteoforms from genes involved in well-known CRC-related pathways. (A) The number of proteoforms and genes in four CRC-related pathways identified from SW480 and SW620 cells. (B) Overlaps of identified and pathway-related proteoforms between SW480 and SW620 cells.

We identified some phosphorylated proteoforms of those genes, which are unique to either SW480 or SW620 cells, Table 7-1. For example, two phosphorylated proteoforms of MARK2 and Sox9 in the WNT/β-catenin Signaling were exclusively identified in the SW480 cells; two phosphorylated proteoforms of EIF4B in the mTOR Signaling pathway were identified solely in the SW620 cells. SW480 and SW620 cells have different phosphorylated proteoforms of EIF4EBP1 in the PI3K/AKT Signaling pathway. We further manually checked the intensities of those proteoforms in the SW480 and SW620 raw files by matching the m/z, charge state, and migration time information from the database search. The proteoform intensity data agree well with the database search results, Table 7-1. For example, the three phosphorylated proteoforms identified solely in SW620 cells have roughly 6-60-fold higher intensity in SW620 cells compared to SW480 cells. The extracted ion electropherograms (EIEs) of the two EIF4B phosphorylated proteoforms from triplicate CZE-MS/MS analyses suggested good reproducibility of proteoform

measurements in terms of base peak proteoform intensity from technical triplicates (RSDs (relative standard deviations) ≤20%). Protein phosphorylation is well known for modulating cancer progression, including CRC. Although the roles of those four genes in regulating cancer progression have been well studied, the specific functions of those phosphorylated proteoforms of the genes have not been investigated. Here, for the first time, we documented the significant differences in protein phosphorylation of those genes between a non-metastatic and a metastatic CRC cell lines in a proteoform-specific manner. Those phosphorylated proteoforms could be central to the progression of CRC metastasis.

Table 7-1 Selected proteoforms of important genes related to WNT/β-catenin Signaling, mTOR Signaling, and PI3K/AKT Signaling pathways. "X" suggests that the proteoform is identified in the sample. "ND" indicates that the proteoform is not identified in the sample.

Gene	Pathway	Proteoform	SW480 cells	SW620 cells
MARK2	WNT/β-catenin Signaling	M.(S)[Acetyl]SARTPLPTLNERDTEQPTLGHLDSK(PSSKSNMIRGRNSAT)[mass shift: 96 Da, phospho and oxidation]SADEQPHIGNY.R	×	ND
SOX9	WNT/β-catenin Signaling	R.SQYDYTDHQNSSSYYSHAAGQGTGLYSTFTYMNPAQRPMYTPIADTSGV(PSIPQTHS)[mass shift: 78 Da, phospho]PQHWEQPVYTQLTRP.	×	ND
EIF4B	mTOR Signaling	M.AASAKKKNK(KGKTISLTDFL)[mass shift: 122 Da, phospho and acetylation/trimethylation]AEDGGTGGGSTYVSKPVSWADETDDLEGDVSTTWHSNDDDVYRAPPIDRSILPTAPR.A	ND	×
EIF4B	mTOR Signaling	M.(A)[Acetyl]ASAKKKNKKGKTISLTDFLAEDGG(T)[mass shift: 80 Da,phospho]GGGSTYVSKPVSWADETDDLEGDVSTTWHSNDDDVYRAPPIDR.S	ND	×
EIF4EBP1	PI3K/AKT Signaling	.MSGGSS(C)[Carbamidomethylation]SQTPSRAIPAT(RRVVLGDGVQLPPGDYSTT)[mass shift:81Da,phospho]PGGTLFSTTPGGTRIIYDRKFLME(C)[Carbamidomethylation]RNSPVTKTPPRDLPTIPGVTSPSSDEPPMEASQSHLRNSPEDKRAGGEESQFEMDI.	ND	×
EIF4EBP1	PI3K/AKT Signaling	R.NSPVTK(T)[mass shift: 80 Da, phospho]PPRDLPTIPGVTSPSSDEPPMEASQSHLRNSPEDKRAGGEESQFEMDI.	ND	×
EIF4EBP1	PI3K/AKT Signaling	K.TPPRDLPTIPGVTS(PSSDEPPMEASQSHLRNS)[mass shift: 81Da, phospho]PEDKRAGGEESQFEMDI.	×	ND

7.3.3 Proteoforms with PTMs and single amino acid variants

Protein PTMs modulate their biological function. For example, protein N-terminal acetylation influences the stability, folding, binding, and subcellular targeting of proteins.[37] Protein phosphorylation is well known for regulating cell signaling, gene expression, and differentiation.[38] Protein methylation plays important roles in modulating transcription.[39] All the data analyses in the following parts of the manuscript are based on the combined data from SEC-CZE-MS/MS, RPLC-CZE-MS/MS, and SEC-RPLC-CZE-MS/MS corresponding to 23,319 proteoforms unless specified otherwise.

This large-scale TDP study identified 4,872 proteoforms with N-terminal acetylation (+42 Da mass shift), 319 proteoforms with phosphorylation [+80 Da (single phosphorylation) or +160 Da (double phosphorylation) mass shift], 321 proteoforms with methylation (+14 Da mass shift), and 241 proteoforms with oxidation (+16 Da mass shift), Figure 7.4A. TDP is powerful for the characterization of combinations of various PTMs on proteoforms. Here we identified 54 proteoforms with two phosphorylation sites and 90 proteoforms with both acetylation and phosphorylation PTMs. Figure 7.4B shows the sequences and fragmentation patterns of 28 kDa heat- and acid-stable phosphoprotein (PDAP1) and Calmodulin-1 (CALM1) proteoforms with either two phosphorylation sites or the combination of N-terminal acetylation and one lysine trimethylation. Those PTMs of the two proteins agree with the literature data.[40, 41] Those two proteoforms were identified with high confidence and were well characterized in terms of PTMs. PDAP1 and CALM1 are both prognostic markers of cancer according to the Human Protein Atlas (https://www.proteinatlas.org/). However, the potential roles of those specific proteoforms of PDAP1 and CALM1 in cancer are still not clear. The capability of TDP for delineating those proteoforms opens the door of further investigating their potential functions in CRC.

One important value of TDP is its capability for delineation of various proteoforms from the same gene (proteoform family).[42] Figure 7.4C shows one example of CALM1 proteoform family. CALM1 modulates many enzymes (kinases and phosphatases), ion channels, and many other proteins by calcium-binding. We identified 75 proteoforms of CALM1. Nearly 70% of those proteoforms start at the position 2 with the N-terminal methionine removal. Various truncated proteoforms, for example, with the starting positions around 40, 60, 80 and 120, were identified in a much lower frequency. The number of proteoform spectrum matches (PrSMs) can be used to roughly estimate the relative abundance of proteoforms.[21] For the CALM1 proteoforms starting from position 2, about 90% of the corresponding PrSMs match to proteoforms covering the whole protein sequence (2-149), called intact proteoforms. The PrSMs corresponding to other C-terminally truncated proteoforms only account for 3% or lower. The intact proteoforms have

various PTMs, including acetylation/trimethylation, oxidation, and phosphorylation. The intact proteoforms of CALM1 with a 42-Da mass shift (acetylation/trimethylation) are the most abundant forms; intact proteoforms with additional oxidation (a 58-Da mass shift) or phosphorylation (a 122-Da mass shift) have much lower abundance according to the number of PrSMs of those proteoforms.

Cancers result from gene mutations, which produce proteoforms containing amino acid variants (AAVs). Although transcriptomic analysis can provide ample information about gene mutations and possible AAVs on proteins, it is valuable to detect proteoforms containing AAVs directly because gene expression can be regulated post-transcriptionally. BUP has been used for the identification of peptides containing single AAVs (SAAVs) from cancer cells.[43] The Kelleher group reported the identification of 10 proteoforms containing SAAVs from breast tumor xenografts in one TDP study.[44] Here we identified 111 proteoforms containing SAAVs of 82 genes from the SW480 and SW620 cell lines with a proteogenomic approach with a 5% proteoform-level FDR, representing one order of magnitude improvement in the number of identified proteoforms containing SAAVs compared to previous studies of cancer cells, Figure 7.4D. The SEC-CZE-MS/MS and RPLC-CZE-MS/MS (RPLC 6 fractions) data were used for the analysis. The transcriptomic variants based on the available RNA-Seq data were incorporated into the protein database for the identification of proteoforms containing SAAVs using TopPG, a recently developed bioinformatics tool.[45] We also manually inspected the MS/MS spectra of proteoforms containing the SAAV sites to ensure high-confidence IDs. Only 20% of the 111 proteoforms were identified from both cell lines, indicating potentially different SAAV profiles between the two cell lines, Figure 7.4D. To confirm the conclusion about SAAV proteoform profile differences, we further analyzed the SAAV-containing proteoforms from 1-D CZE-MS/MS. Although the number of SAAV proteoforms from SW620 cells is about twice as many as that from SW480 cells, only half of the SW480 SAAV proteoforms are covered by the SW620 ones. Manual evaluation of some SAAV proteoforms exclusively identified from SW480 and SW620 cells in raw MS data supported the conclusion.

Figure 7.4E shows the sequences and fragmentation patterns of two examples of proteoforms containing SAAVs. TP53 is an important tumor suppressor closely related to CRC development, and it is an essential member in WNT/β-catenin Signaling and PI3K/AKT Signaling pathways. We identified one TP53 proteoform containing an AAV at position 72 (P$\rightarrow$R) due to the codon 72 polymorphism. Studies have shown the functional differences of the P72 and R72 proteoforms of TP53.[46,47] For example, the R72 proteoform does a markedly better job of inducing apoptosis compared to the P72 proteoform.[46] Another study indicated that the expression of P72

proteoform increased CRC metastasis, and that the R72 proteoform does not exist in the non-metastatic CRC cell line (SW480) based on the nucleic-acid data.[47] Interestingly, we only identified the R72 proteoform of TP53 in the SW620 cell line, not in the SW480 cell line, from the top-down MS data. MSH6 is one of the DNA mismatch repair genes and its mutations play a crucial role in Lynch syndrome, which is an inherited form of CRC. We identified one MSH6 proteoform containing a SAAV due to polymorphism at position 39 (G$\rightarrow$E). The G39E SAAV has been associated with an increased risk of CRC according to the nucleic-acid data.[48] We identified G39 proteoforms of MSH6 in both SW480 and SW620 cells, but identified the E39 proteoform only in the SW480 cells, not in the SW620 cells.

For the proteoforms containing SAAVs, we further performed QIAGEN Ingenuity Pathway Analysis (IPA) of the corresponding 82 genes. We revealed that 75 of those genes are associated with tumorigenesis of tissue (p-value: 0.0001), and three genes (MSH6, PITX1 and TP53) relate to the development of colon tumor (p-value: 0.002). Five of the genes related to tumorigenesis of tissue (AURKA, EIF5A, PFKFB3, POLE4, and TP53) are targets of cancer drugs. We further performed IPA network analysis and revealed that 17 out of the 82 genes are involved in a cancer-related network (network score 36), Figure 7.4F, suggesting their crucial roles in cancer and development. The 17 genes are highlighted in purple and those proteins belong to several different families, including enzyme (diamond shape, LARS1, PARS1, ALDOA, MSH6, and PPIF), phosphatase/kinase (triangle shape, PGAM1, SET, and PFKFB3), transcription regulator (oval shape, TP53 and PITX1), and others (circle shape, PSG1, SRP14, MAGEB2, MT1G, MT1H, MT1M, and ISG15). Nine of those highlighted proteins have direct (solid line) or indirect (dotted line) interactions with TP53.

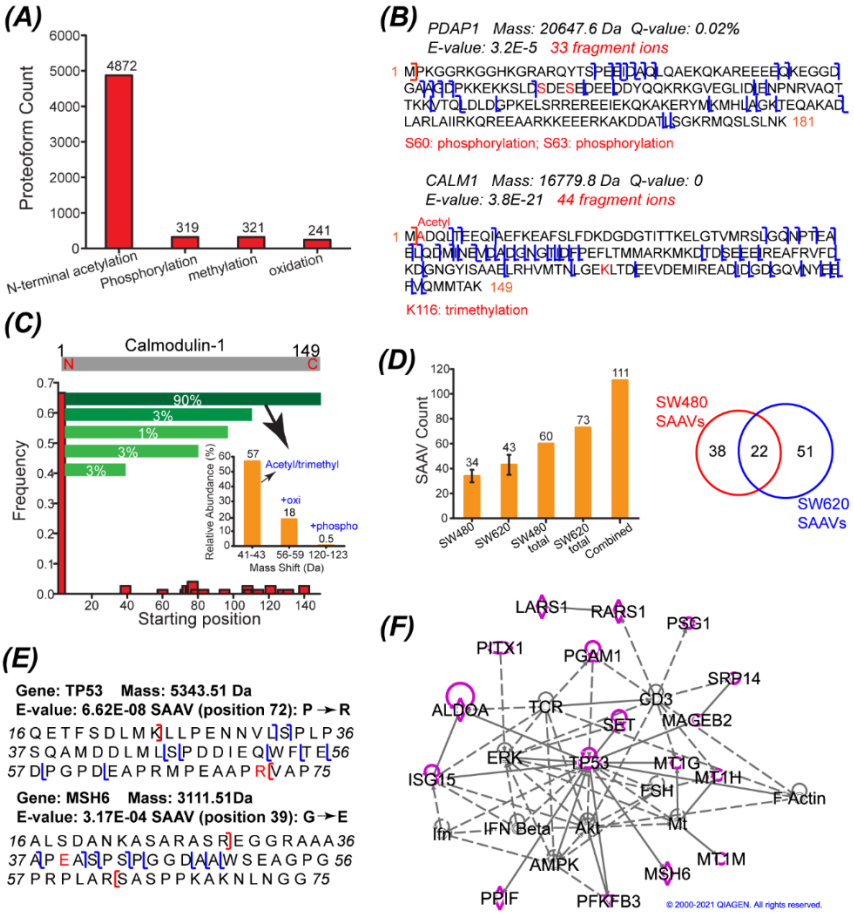

Figure 7.4 Analyses of the identified proteoforms from CRC cells with PTMs and single amino acid variants (SAAVs). (A) Proteoforms with various PTMs, including N-terminal acetylation, phosphorylation, methylation, and oxidation. (B) Sequences and fragmentation patterns of two proteoforms, one proteoform of PDAP1 with two phosphorylation sites and one proteoform of CALM1 with N-terminal acetylation and one lysine trimethylation. (C) Summary of all the identified proteoforms of calmodulin-1 (CALM1) regarding starting positions, relative abundance based on the number of PrSMs, and PTMs. (D) The number of proteoforms containing SAAVs identified from the SW480 and SW620 cells and the overlap of those proteoforms. The SEC-CZE-MS/MS and RPLC-CZE-MS/MS (RPLC 6 fractions) data were used for the analysis. The error bars in the figure represent the standard deviations of proteoforms from triplicate measurements. (E) Sequences and fragmentation patterns of two proteoforms containing SAAVs. (F) SAAVs containing proteoforms correspond to many genes (highlighted in purple) that are involved in a cancer related network according to the IPA analysis.

7.3.4 Quantitative TDP of metastatic and non-metastatic human CRC cell lines

We further carried out the first quantitative TDP study of a pair of metastatic (SW620) and non-metastatic (SW480) human CRC cell lines. The cell lysates of SW480 and SW620 cells were fractionated by SEC and each fraction was analyzed by CZE-MS/MS in technical triplicate. After database search with TopPIC, we identified roughly 4,000 proteoforms per replicate per cell line with a 1% proteoform-level FDR. The intensity distributions of identified proteoforms across technical triplicates and the two cell lines are consistent. We performed label-free quantification (LFQ) analysis using TopDiff (version 1.3.4), a tool in the TopPIC suite, which reported about 1,500 proteoforms with measured intensities in all the six samples (three replicates per cell line and two cell lines). The SEC-CZE-MS/MS system shows reasonably good reproducibility regarding the intensities of shared proteoforms, as evidenced by the strong linear correlations of proteoform intensities between technical replicates of SW480 or SW620 cells (Pearson correlation coefficients: 0.86-0.93). The Pearson correlation coefficients of proteoform intensity between SW480 and SW620 cells are statistically significantly lower than that between technical replicates of one cell line (0.71±0.01 vs. 0.90±0.03, $p<10^{-10}$, two-tailed student's t-test), indicating significant differences between the two cell lines in terms of proteoform intensity. We used the Perseus software for further data analysis.[49] The two cell lines can be easily distinguished using the proteoform quantification profiles, Figure 7.5A. Two clusters of differentially expressed proteoforms across the six samples were revealed.

According to the volcano plot in Figure 7.5B, 460 proteoforms of 248 proteins showed statistically significant differences in abundance between the two cell lines (FDR<0.05). Specifically, 244 proteoforms of 152 proteins had higher abundance in the SW480 cell line and 216 proteoforms of 132 proteins had higher expression in the SW620 cell line. Figure 7.5B shows that one HMGN1 proteoform and one RBM8A proteoform have the most significant abundance changes between SW480 and SW620 cells. HMGN1 regulates gene expression and PTMs of core histones, affecting DNA repair and tumor progression.[50] It has been reported that RBM8A promotes tumor cell migration and invasion in the most common type of primary liver cancer.[51]

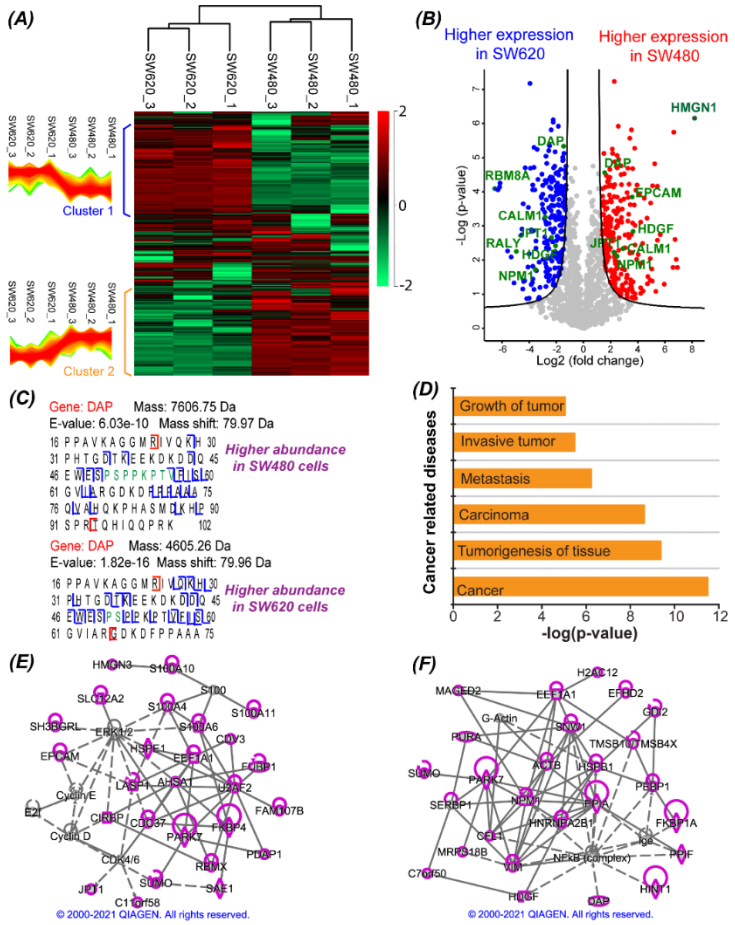

Figure 7.5 Summary of the LFQ data of SW480 and SW620 cells. (A) Heat map and cluster analysis of the quantified proteoforms regarding LFQ intensities. A Z-score normalization was employed. The red color represents high intensity and the green color indicates low intensity. (B) Volcano plot showing differentially expressed proteoforms between the two cell lines. Red dots and blue dots represent proteoforms having statistically significantly higher abundance in SW480 and in SW620, respectively. Gene names of some differentially expressed proteoforms are labeled. The Perseus software was used for generating the heat map in (A) and Volcano plot in (B) with the following settings (S0=1 and FDR = 0.05).[47] (C) Sequences and fragmentation patterns of two phosphorylated proteoforms of the gene DAP. One has higher abundance in SW480 cells and the other has higher expression in SW620 cells. (D) An IPA analysis reported some cancer related diseases that are related to the differentially expressed genes in the two cell lines. Proteoforms with higher abundance in SW480 cells (E) or higher abundance in SW620 cells (F) correspond to genes that are involved in cancer-related networks with high scores.

Comparing the overexpressed and underexpressed proteoforms in the two cell lines revealed that 36 genes (e.g., DAP, CALM1, HDGF, JPT1, and NPM1) have both overexpressed and underexpressed proteoforms in one cell line, suggesting that different proteoforms of the same gene had completely different expression patterns in the two cell lines. Figure 7.5C shows two differentially expressed proteoforms of DAP (Death-associated protein 1), one of those 36 genes. It has been reported that DAP modulates cell death and correlates with the clinical outcome of CRC patients.[52] Interestingly, we revealed that one phosphorylated proteoform of DAP (~7,607 Da, phosphorylation site S51 or T56) had a higher abundance in SW480 cells and another phosphorylated proteoform (~4,605 Da, phosphorylation site S51) showed higher expression in SW620 cells. Both the S51 and T56 are known to be phosphorylated according to PhosphoSitePlus, with S51 being the most common phosphorylation site of DAP. We noted that the differentially expressed proteoforms in this study include phosphorylated proteoforms of several important genes related to CRC, i.e., RALY,[53] NPM1,[54] DAP,[52] and HDGF.[55] The functions of phosphorylated forms of those four proteins in modulating CRC development are still unclear. However, the differential expressions of those phosphorylated proteoforms in the metastatic and non-metastatic CRC cells suggest their potential roles in regulating CRC metastasis.

We highlight several differentially expressed proteoforms of CALM1, JPT1 (HN1), and EPCAM. CALM-dependent systems play important roles in cancer metastasis.[56] JPT1 (HN1) promotes cancer metastasis via activating the NF-ƙB signaling pathway.[57] EPCAM is a human cell surface glycoprotein and plays crucial roles in tumor biology, especially CRC.[58] EPCAM has been recognized as an important therapeutic target for cancer. We discovered two CALM1 proteoforms having significantly higher abundance in SW620 cells compared to SW480 cells; one of them contains K116 trimethylation. We revealed one CALM1 proteoform showing higher abundance in SW480 cells and the proteoform carries N-terminal acetylation and a 58-Da mass shift between amino acid residues 73 and 89. The 58-Da mass shift can be explained as a trimethylation/acetylation plus oxidation. Three of JPT1 proteoforms have higher abundance in SW480 cells and one of them contains a 167-Da mass shift between the amino acid residues 66 and 89, where seven serine residues can be phosphorylated according to the PhosphoSitePlus database (https://www.phosphosite.org/). The 167-Da mass shift most likely represents a combination of phosphorylation and other PTMs. Interestingly, one JPT1 proteoform shows higher abundance in SW620 cells. We also observed two EPCAM proteoforms having higher abundance in SW480 cells.

We then performed IPA analyses of the genes of those differentially expressed proteoforms between SW480 and SW620 cells. Those genes are heavily involved in cancer-related diseases, for example, tumorigenesis of tissue and metastasis, Figure 7.5D. Five of those proteins (EIF4E, EPCAM, FKBP1A, GAA, and HSP90AB1) are drug targets. IPA network analyses revealed that 26 proteins (highlighted in purple) whose proteoforms showed higher abundance in SW480 compared to SW620 were involved in a cancer-related network (score 51), Figure 7.5E. Those proteins belong to several families, including enzyme (diamond shape, e.g., PARK7 and FKBP4), transcription regulator (oval shape, e.g., FUBP1), translation regulator (hexagon shape, e.g., CIRBP and EEF1A1), transporter (trapezium shape, e.g., SLC12A2 and LASP1), and others (circle shape, e.g., EPCAM and JPT1). Most of those proteins have direct (solid line) and indirect (dotted line) interactions with one another. We also carried out network analysis for the proteins whose proteoforms had higher expression in SW620 cells, and observed high-scores for cancer-related networks. Figure 7.5F shows one cancer-related network (score 54), and 26 of those proteins are involved in the network (highlighted in purple). Those proteins include several CRC-related important proteins, NPM1 (oval shape, transcription regulator, located in nucleus), DAP (transcription regulator, located in cytoplasm), and HDGF (square shape, growth factor, located in extracellular space). NPM1 is a crucial protein in the network and many of the highlighted proteins have direct interactions (solid line) with NPM1, for example, PARK7, VIM, and PPIA. NPM1 also has indirect interaction (dotted line) with the NFkB complex, which plays crucial roles in modulating DNA transcription and cell survival. Human NPM1 boosts the activation of NFkB according to Ingenuity relationships from the IPA analysis. Besides NPM1, several other highlighted proteins (e.g., HDGF and DAP) also have indirect interactions with the NFkB complex. For example, NFkB regulates the transcription of HDGF, and DAP deactivates the NFkB according to the IPA network analysis results. The IPA analysis also revealed that 13 proteoforms of three genes (EIF4B, EIF4E, EIF4EBP1) in the mTOR Signaling pathway had statistically significant differences in abundance between the SW480 and SW620 cells.

7.4 Conclusions

In this study, we advanced TDP of human cells drastically in terms of the number of proteoform IDs per human cell line compared to previous LC-MS/MS-based studies (~16,000 vs. ~3,000) via coupling LC fractionations to CZE-MS/MS. This work represents an important progress in TDP, which aims to characterize the human proteome in a proteoform-specific manner (Human Proteoform Project).[16]

TDP of metastatic and non-metastatic cells is crucial for discovering new protein biomarkers and providing a more accurate understanding of molecular mechanisms of cancer metastasis. According to the results from our qualitative and quantitative TDP of SW480 and SW620 cells, we had several conclusions about CRC metastasis. First, CRC cells have a significant transformation in proteoforms and SAAVs after metastasis, evidenced by obvious differences of proteoform and SAAV profiles between SW480 and SW620 cells. Second, different proteoforms from the same cancer-related gene (e.g., DAP, CALM1, HDGF, JPT1, RALY, and NPM1) may have potentially varied biological functions in modulating CRC metastasis, because they show opposite expression profiles between the SW480 and SW620 cells, Figure 7.5B. Some proteoforms of those genes have higher abundance in SW480 cells; some of their proteoforms show higher expression in SW620 cells. Third, PTMs (i.e., phosphorylation) of important cancer-related genes (i.e., DAP, HDGF, JPT1, RALY, NPM1, MARK2, SOX9, EIF4B, and EIF4EBP1) could play important roles in regulating CRC metastasis, evidenced by the significant abundance differences of phosphorylated proteoforms from those genes between the SW480 and SW620 cells. The differentially expressed proteoforms, especially those with PTMs, of important cancer-related genes could be novel proteoform biomarkers of CRC metastasis. Fourth, proteoforms of genes in well-known CRC-related pathways (WNT/β-catenin Signaling, PI3K/AKT Signaling, mTOR Signaling, and ERK/MAPK Signaling) are different between SW480 and SW620 cells, and those proteoforms could play vital roles in modulating CRC metastasis.

Our TDP strategies still have some technical limitations. One relates to the identification of large proteoforms. In this work, we focused on the characterization of proteoforms smaller than 30 kDa. CZE-MS/MS has much lower sample loading capacity compared to RPLC-MS/MS (nL vs. µL), resulting in a limited mass of protein materials that can be injected for measurements with CZE-MS/MS. This issue is particularly severe for the characterization of large proteoforms in a complex proteome sample because large proteoforms tend to have drastically lower signal-to-noise ratios than small proteoforms due to the much wider charge state distributions. Highly efficient size-based fractionation techniques must be employed to enrich large proteoforms before CZE-MS/MS. Additionally, more effort needs to be made to improve the sample loading capacity of CZE-MS/MS via investigating online sample stacking techniques or solid phase microextraction (SPME) methods. Another limitation relates to the extensive fragmentation of proteoforms for accurate localization of PTMs. The backbone cleavage coverage of proteoforms from commonly used collision-based fragmentation techniques (i.e., collision-induced dissociation (CID) and higher energy collision dissociation (HCD)) is limited. We expect that coupling our LC-CZE-MS/MS technique to a mass spectrometer with electron- or photon-based gas-phase

fragmentation techniques (i.e., electron-capture dissociation (ECD),[59] electron-transfer dissociation (ETD),[60] and ultraviolet photodissociation (UVPD)[61]) will revolutionize TDP for the Human Proteoform Project.[16]

7.5 Acknowledgments

7.5.1 Funding

The work was funded by National Cancer Institute (NCI) through the grant R01CA247863 (Sun, Hummon, and Liu). We also thank the support from National Institute of General Medical Sciences (NIGMS) through grants R01GM125991 (Sun and Liu) and R01GM118470 (Liu and Sun). Sun also thanks the support from the National Science Foundation (CAREER Award, Grant DBI1846913). We thank MSU AgBioResearch and Michigan State University for the access to QIAGEN Ingenuity Pathway Analysis (IPA) platform.

7.5.2 Contributions

Elijah N. McCool (E.N.M.) performed the experiments for proteoform identifications using RPLC-CZE-MS/MS and SEC-RPLC-CZE-MS/MS. Tian Xu (T.X.) performed the experiment for proteoform identification and/or quantification using SEC-CZE-MS/MS and 1D-CZE-MS/MS. Wenrong Chen (W.C.) carried out all the database search using TopPIC for proteoform ID and quantification. E.N.M., T.X., and W.C. worked together for data analysis and made the first draft of the manuscript. Nicole C. Beller (N.C.B.) did all the cell culture and initial sample preparation of SW480 and SW620 cells. Scott M. Nolan (S.M.N.) performed the LC fractionations. Amanda B. Hummon (A.B.H.), Xiaowen Liu (X.L.), and Liangliang Sun (L.S.) conceived the original idea. X.L. supervised the database search part of the project. L.S. supervised the project.

REFERENCES

(1) Schmitt, M.; Greten, F. R.The inflammatory pathogenesis of colorectal cancer. *Nat. Rev. Immunol.* **2021**, *21*, 653-667.

(2) Rehman, S. K.; Haynes, J.; Collignon, E.; Brown, K. R.; Wang, Y.; Nixon, A. M.; Bruce, J. P.; Wintersinger, J. A.; Mer, A. S.; Lo, E. B.Colorectal cancer cells enter a diapause-like DTP state to survive chemotherapy. *Cell* **2021**, *184*, 226-242. e221.

(3) Markowitz, S. D.; Bertagnolli, M. M.Molecular basis of colorectal cancer. *New Engl. J. Med.* **2009**, *361*, 2449-2460.

(4) Schunter, A. J.; Yue, X.; Hummon, A. B.Phosphoproteomics of colon cancer metastasis: comparative mass spectrometric analysis of the isogenic primary and metastatic cell lines SW480 and SW620. *Anal. Bioanal. Chem.* **2017**, *409*, 1749-1763.

(5) Zhang, B.; Whiteaker, J. R.; Hoofnagle, A. N.; Baird, G. S.; Rodland, K. D.; Paulovich, A. G.Clinical potential of mass spectrometry-based proteogenomics. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 256-268.

(6) Xu, L.; Wang, R.; Ziegelbauer, J.; Wu, W. W.; Shen, R.-F.; Juhl, H.; Zhang, Y.; Pelosof, L.; Rosenberg, A. S.Transcriptome analysis of human colorectal cancer biopsies reveals extensive expression correlations among genes related to cell proliferation, lipid metabolism, immune response and collagen catabolism. *Oncotarget* **2017**, *8*, 74703.

(7) Huo, T.; Canepa, R.; Sura, A.; Modave, F.; Gong, Y.Colorectal cancer stages transcriptome analysis. *PLoS One* **2017**, *12*, e0188697.

(8) Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513*, 382-387.

(9) Besson, D.; Pavageau, A.-H.; Valo, I.; Bourreau, A.; Bélanger, A.; Eymerit-Morin, C.; Moulière, A.; Chassevent, A.; Boisdron-Celle, M.; Morel, A.A quantitative proteomic approach of the different stages of colorectal cancer establishes OLFM4 as a new nonmetastatic tumor marker. *Mol. Cell Proteomics* **2011**, *10*.

(10) Ghosh, D.; Yu, H.; Tan, X. F.; Lim, T. K.; Zubaidah, R. M.; Tan, H. T.; Chung, M. C.; Lin, Q.Identification of key players for colorectal cancer metastasis by iTRAQ quantitative proteomics profiling of isogenic SW480 and SW620 cell lines. *J. Proteome Res.* **2011**, *10*, 4373-4387.

(11) Smith, L. M.; Kelleher, N. L.Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10*, 186-187.

(12) Smith, L. M.; Kelleher, N. L.Proteoforms as the next proteomics currency. *Science* **2018**, *359*, 1106-1107.

(13) Toby, T. K.; Fornelli, L.; Kelleher, N. L.Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Ana.l Chem.* **2016**, *9*, 499.

(14) Ntai, I.; Fornelli, L.; DeHart, C. J.; Hutton, J. E.; Doubleday, P. F.; LeDuc, R. D.; van Nispen, A. J.; Fellers, R. T.; Whiteley, G.; Boja, E. S.Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 4140-4145.

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

(16) Smith, L. M.; Agar, J. N.; Chamot-Rooke, J.; Danis, P. O.; Ge, Y.; Loo, J. A.; Paša-Tolić, L.; Tsybin, Y. O.; Kelleher, N. L.; Proteomics, C. f. T.-D.The human proteoform project: defining the human proteome. *Sci. Adv.* **2021**, *7*, eabk0734.

(17) 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.How many human proteoforms are there? *Nat. Chem. Biol.* **2018**, *14*, 206-214.

(18) 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.Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011**, *480*, 254-258.

(19) 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*, 3465-3473.

(20) 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.Identification and characterization of human proteoforms by top-down LC-21 tesla FT-ICR mass spectrometry. *J. Proteome Res.* **2017**, *16*, 1087-1096.

(21) Melani, R. D.; Gerbasi, V. R.; Anderson, L. C.; Sikora, J. W.; Toby, T. K.; Hutton, J. E.; Butcher, D. S.; Negrão, F.; Seckler, H. S.; Srzentić, K.The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* **2022**, *375*, 411-418.

(22) 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*, 5467-5475.

(23) 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*, 5529-5533.

(24) Yang, Z.; Shen, X.; Chen, D.; Sun, L.Toward a universal sample preparation method for denaturing top-down proteomics of complex proteomes. *J. Proteome Res.* **2020**, *19*, 3315-3325.

(25) 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*, 1435-1445.

(26) McCool, E. N.; Liangliang, S.Comparing nanoflow reversed-phase liquid chromatography-tandem mass spectrometry and capillary zone electrophoresis-tandem mass spectrometry for top-down proteomics. *Se pu= Chinese journal of chromatography* **2019**, *37*, 878.

(27) Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates III, 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.

(28) Yang, Z.; Shen, X.; Chen, D.; Sun, L.Improved nanoflow RPLC-CZE-MS/MS system with high peak capacity and sensitivity for nanogram bottom-up proteomics. *J. Proteome Res.* **2019**, *18*, 4046-4054.

(29) Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L.Recent advances (2019–2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. *Mass Spectrom. Rev.* **2021**.

(30) Gomes, F. P.; Yates III, J. R.Recent trends of capillary electrophoresis-mass spectrometry in proteomics research. *Mass Spectrom. Rev.* **2019**, *38*, 445-460.

(31) Koveitypour, Z.; Panahi, F.; Vakilian, M.; Peymani, M.; Seyed Forootan, F.; Nasr Esfahani, M. H.; Ghaedi, K.Signaling pathways involved in colorectal cancer progression. *Cell Biosci.* **2019**, *9*, 1-14.

(32) Francipane, M. G.; Lagasse, E.mTOR pathway in colorectal cancer: an update. *Oncotarget* **2014**, *5*, 49.

(33) Pasapera, A. M.; Heissler, S. M.; Eto, M.; Nishimura, Y.; Fischer, R. S.; Thiam, H. R.; Waterman, C. M.MARK2 regulates directed cell migration through modulation of myosin II contractility and focal adhesion organization. *Curr. Biol.* **2022**.

(34) Lü, B.; Fang, Y.; Xu, J.; Wang, L.; Xu, F.; Xu, E.; Huang, Q.; Lai, M.Analysis of SOX9 expression in colorectal cancer. *Am. J. Clin. Pathol.* **2008**, *130*, 897-904.

(35) Shahbazian, D.; Parsyan, A.; Petroulakis, E.; Hershey, J. W.; Sonenberg, N.eIF4B controls survival and proliferation and is regulated by proto-oncogenic signaling pathways. *Cell cycle* **2010**, *9*, 4106-4109.

(36) Chen, Y.; Wang, J.; Fan, H.; Xie, J.; Xu, L.; Zhou, B.Phosphorylated 4E-BP1 is associated with tumor progression and adverse prognosis in colorectal cancer. *Neoplasma* **2017**, *64*, 787-794.

(37) Ree, R.; Varland, S.; Arnesen, T.Spotlight on protein N-terminal acetylation. *Exp. Mol. Med.* **2018**, *50*, 1-13.

(38) Kalume, D. E.; Molina, H.; Pandey, A.Tackling the phosphoproteome: tools and strategies. *Curr. Opin. Chem. Biol.* **2003**, *7*, 64-69.

(39) Lee, D. Y.; Teyssier, C.; Strahl, B. D.; Stallcup, M. R.Role of protein methylation in regulation of transcription. *Endocr. Rev.* **2005**, *26*, 147-170.

(40) Zhou, H.; Di Palma, S.; Preisinger, C.; Peng, M.; Polat, A. N.; Heck, A. J.; Mohammed, S.Toward a comprehensive characterization of a human cancer cell phosphoproteome. *J. Proteome Res.* **2013**, *12*, 260-271.

(41) Sasagawa, T.; Ericsson, L. H.; Walsh, K. A.; Schreiber, W. E.; Fischer, E. H.; Titani, K.Complete amino acid sequence of human brain calmodulin. *Biochemistry* **1982**, *21*, 2565-2569.

(42) 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*, 3671-3680.

(43) Tan, Z.; Zhu, J.; Stemmer, P. M.; Sun, L.; Yang, Z.; Schultz, K.; Gaffrey, M. J.; Cesnik, A. J.; Yi, X.; Hao, X.Comprehensive detection of single amino acid variants and evaluation of their deleterious potential in a PANC-1 cell line. *J. Proteome Res.* **2020**, *19*, 1635-1646.

(44) Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D.Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Mol. Cell Proteomics* **2016**, *15*, 45-56.

(45) Chen, W.; Liu, X.Proteoform identification by combining RNA-Seq and top-down mass spectrometry. *J. Proteome Res.* **2020**, *20*, 261-269.

(46) Jeong, B. S.; Hu, W.; Belyi, V.; Rabadan, R.; Levine, A. J.Differential levels of transcription of p53-regulated genes by the arginine/proline polymorphism: p53 with arginine at codon 72 favors apoptosis. *FASEB J.* **2010**, *24*, 1347-1353.

(47) Katkoori, V. R.; Manne, U.; Chaturvedi, L. S.; Basson, M. D.; Haan, P.; Coffey, D.; Bumpers, H. L.Functional consequence of the p53 codon 72 polymorphism in colorectal cancer. *Oncotarget* **2017**, *8*, 76574-76586.

(48) Zelga, P.; Przybyłowska-Sygut, K.; Zelga, M.; Dziki, A.; Majsterek, I.Polymorphism of Gly39Glu (c.116G>A) hMSH6 is associated with sporadic colorectal cancer development in the Polish population: Preliminary results. *Adv Clin Exp Med* **2017**, *26*, 1425-1429.

(49) 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*, 731-740.

(50) Postnikov, Y.; Bustin, M.Regulation of chromatin structure and function by HMGN proteins. *Biochim. Biophys. Acta* **2010**, *1799*, 62-68.

(51) Liang, R.; Lin, Y.; Ye, J. Z.; Yan, X. X.; Liu, Z. H.; Li, Y. Q.; Luo, X. L.; Ye, H. H.High expression of RBM8A predicts poor patient prognosis and promotes tumor progression in hepatocellular carcinoma. *Oncol. Rep.* **2017**, *37*, 2167-2176.

(52) Jia, Y.; Ye, L.; Ji, K.; Toms, A. M.; Davies, M. L.; Ruge, F.; Ji, J.; Hargest, R.; Jiang, W. G.Death associated protein 1 is correlated with the clinical outcome of patients with colorectal cancer and has a role in the regulation of cell death. *Oncol. Rep.* **2014**, *31*, 175-182.

(53) Sun, L.; Wan, A.; Zhou, Z.; Chen, D.; Liang, H.; Liu, C.; Yan, S.; Niu, Y.; Lin, Z.; Zhan, S.; Wang, S.; Bu, X.; He, W.; Lu, X.; Xu, A.; Wan, G.RNA-binding protein RALY reprogrammes mitochondrial metabolism via mediating miRNA processing in colorectal cancer. *Gut* **2021**, *70*, 1698-1712.

(54) Grisendi, S.; Mecucci, C.; Falini, B.; Pandolfi, P. P.Nucleophosmin and cancer. *Nat. Rev. Cancer* **2006**, *6*, 493-505.

(55) Sun, B.; Gu, X.; Chen, Z.; Xiang, J.MiR-610 inhibits cell proliferation and invasion in colorectal cancer by repressing hepatoma-derived growth factor. *Am J Cancer Res* **2015**, *5*, 3635-3644.

(56) Villalobo, A.; Berchtold, M. W.The Role of Calmodulin in Tumor Cell Migration, Invasiveness, and Metastasis. *Int J Mol Sci* **2020**, *21*.

(57) Chen, J.; Qiu, J.; Li, F.; Jiang, X.; Sun, X.; Zheng, L.; Zhang, W.; Li, H.; Wu, H.; Ouyang, Y.; Chen, X.; Lin, C.; Song, L.; Zhang, Y.HN1 promotes tumor associated lymphangiogenesis and lymph node metastasis via NF-κB signaling activation in cervical carcinoma. *Biochem. Biophys. Res. Commun.* **2020**, *530*, 87-94.

(58) Armstrong, A.; Eck, S. L.EpCAM: A new therapeutic target for an old cancer antigen. *Cancer Biol Ther* **2003**, *2*, 320-326.

(59) Ge, Y.; Lawhorn, B. G.; ElNaggar, M.; Strauss, E.; Park, J. H.; Begley, T. P.; McLafferty, F. W.Top down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry. *J. Am. Chem. Soc.* **2002**, *124*, 672-678.

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

(61) Shaw, J. B.; Li, W.; Holden, D. D.; Zhang, Y.; Griep-Raming, J.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S.Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. *J. Am. Chem. Soc.* **2013**, *135*, 12646-12651.

CHAPTER 8. Conclusions and future directions

This dissertation is dedicated to advance cIEF-MS/MS and CZE-MS/MS for TDP. First, high-capacity and high-throughput cIEF-MS/MS approaches were developed and optimized for qualitative and quantitative TDP. Second, we greatly improved robustness and resolution of cIEF-MS for characterization of charge variants of mAbs. The method can be easily adapted to perform targeted TDP for investigating proteoform heterogeneity of important proteoform families (e.g KRAS4b). Third, non-denaturing cIEF-MS was developed for studying the microheterogeneity of protein complexes. The approach is highly promising for implementing native TDP to study the protein complex heterogeneity in complex proteome samples. Fourth, FAIMS was coupled to CZE-MS/MS and greatly boosted the sensitivity of the system and number of proteoform IDs, offering a new multidimensional separation tool for deep TDP. Finally, cIEF-MS/MS and CZE-MS/MS-based platforms were employed for two biological applications, in which we delineated sex-dependent proteoform profiles in brains and uncovered proteoform-level differences between metastatic and non-metastatic CRC cells, respectively.

For CZE-MS/MS, improvement of sample loading capacity remains desired. Incorporating solid phase microextraction (SPME) with CZE can potentially enhance the sample loading amount. Alternatively, cIEF has a high loading capacity and can be an ideal option for TDP. cIEF-MS/MS typically provides ultrahigh resolving power for the proteoforms/protein complexes with microheterogeneity on PTMs, however, it is restricted in sensitivity due to ESI interference by ampholyte. As we learned from our CZE-FAIMS-MS/MS study, FAIMS has the capability to fractionate ions based on size by varying CVs. Considering ampholytes generally have small molecular weights, future application of FAIMS in cIEF-MS/MS is promising to address the sensitivity issue by removing the ampholytes before MS detection.

Furthermore, current TDP studies mostly focus on proteoforms below 30 kDa. Identification of larger proteoforms faces various challenges from protein loss during sample preparation, signal suppression from coelution, limitation of MS instrumentations, inefficient fragmentations, and lack of effective bioinformatics tools to interpret data. In Chapter 5, we attempted to combine an improved sample preparation workflow with CZE-FAIMS-MS/MS, which greatly expanded the number of large proteoform IDs (20 kDa-40 kDa). More effort is required to apply advanced fragmentation tools such as ECD and UVPD to achieve better sequence coverage and PTM localization of larger proteoforms (>30 kDa). The improvement of bioinformatic tools is also necessary for interpreting MS spectra with low resolution and MS/MS spectra with high complexity.

TDP has served as important strategy for characterizing the proteoforms related to cancers and discovery of biomarkers for disease. In our study, we achieved large datasets of proteoform IDs from metastatic and non-metastatic CRC cell lines and disclosed a variety of proteoforms with discrepancy on SAAVs or level of expression. However, the functions of the individual proteoforms and crosstalk regulation of proteoforms remain unknown. Future study to explore the proteoform associated function and biological processes can deepen our understanding of the mechanism of cancer occurrence and metastasis. To achieve this goal, the targeted TDP will be combined with molecular biology techniques for studying proteoforms of specific cancer related genes.