METHOD DEVELOPMENT FOR CAPILLARY ELECTROPHORESIS MASS SPECTROMETRY (CE-MS)-BASED PROTEOMICS AND APPLICATION TO UNCOVERING PROTEOME DYNAMICS OF ZEBRAFISH EMBRYOS DURING EARLY EMBRYOGENSIS

By

Daoyang Chen

A DISSERTATION

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

Chemistry-Doctor of Philosophy

2021

ABSTRACT

METHOD DEVELOPMENT FOR CAPILLARY ELECTROPHORESIS MASS SPECTROMETRY (CE-MS)-BASED PROTEOMICS AND APPLICATION TO UNCOVERING PROTEOME DYNAMICS OF ZEBRAFISH EMBRYOS DURING EARLY EMBRYOGENSIS

By

Daoyang Chen

Reversed-phase liquid chromatography (RPLC) coupling with tandem MS (MS/MS) is often the method of choice in both peptide-centric bottom-up proteomics (BUP) and proteoform-centric top-down proteomics (TDP) studies. In recent years, capillary zone electrophoresis (CZE)-MS has attracted attention as another platform in proteomics due to high separation efficiency, high sensitivity, and complementarity to LC-MS. This work is dedicated to developing novel CE-MS-based methods for large-scale proteomics and applies them to study the proteome dynamics of zebrafish embryos during early embryogenesis.

In Chapter 2, a sample stacking method, dynamic pH junction, was systematically investigated and employed to improve CZE's sample loading capacity for large-scale BUP. The results of the optimized system represent the highest loading capacity, the highest peak capacity, and the widest separation window of CZE for peptide separation to date. The automated CZE-MS system opened the door to using CZE-MS for large-scale BUP.

In Chapter 3, for the first time, a strong cation exchange (SCX)-RPLC-CZE-MS/MS platform was established for deep BUP and phosphoproteomics. The platform approached comparable performance to the modern 2D-LC-MS/MS for deep proteomic sequencing evident by identifying 8200 protein groups and 65,000 unique peptides from a mouse brain proteome digest, 11,555 phosphopeptides from

the HCT116 cell line. SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS showed good complementarity in protein, peptide, and phosphopeptide IDs.

In Chapter 4, a quantitative BUP study was performed on zebrafish embryos across four developmental stages during the maternal-to-zygotic transition (MZT) via coupling isobaric tag for relative and absolute quantitation (iTRAQ) chemistry with both RPLC-MS/MS and CZE-MS/MS. Expression kinetics of nearly 5000 proteins including over 100 transcription factors (TFs) across four early embryonic stages were determined. The protein expression profiles fall into several different clusters and accurately reflect the important events during early embryogenesis. Further studies of the expression profiles of TFs revealed that the differentially expressed TFs during the MZT show wave-like expression patterns.

Top-down proteomics (TDP) aims to directly characterize proteoforms in cells. CZE-MS/MS has been demonstrated as a useful tool for TDP. In Chapter 5, for the first time, we evaluated various semiempirical models for predicting proteoforms' electrophoretic mobility using large-scale TDP data sets from earlier CZE–MS/MS studies. Linear correlations were achieved between the experimental and predicted μ_{ef} of *E. coli* proteoforms and histone proteoforms (R^2 = 0.98), demonstrating that the μ_{ef} of proteoforms in CZE-MS can be predicted accurately, which could be useful for validating the confidence of proteoform IDs from a database search.

In Chapter 6, we concluded the results of this dissertation and provided our expectations for future studies.

This thesis is dedicated to my parents who have been supporting me unconditionally.
This thesis is also dedicated to my family and friends, you always cheer me up.
To my grandpa, I will miss you and love you always.

ACKNOWLEDGEMENTS

Without the help and support of many people, I can never complete this project. I would like to share my sincere appreciation with the following people:

Dr. Liangliang Sun, my mentor, has given a massive amount of support since day one we have met. I first joined the Sun group as a research assistant, not a student. By then, Dr. Sun just started his career as an Assistant Professor at Michigan State University. Therefore, we had a lot of time working together in the lab during my first year in the Sun group. Lucky for me, I have got his direct mentorship. I started my Ph.D. application after working at MSU for half a year, he offered continuous help throughout the process. He motivates me with his passion, patience, and diligence in my daily research for the last four and a half years. I could not have finished the project without him. I sincerely wish him and his family all the best. And I hope the Sun group “Live long and prosper”.

I want to thank Dr. Dana Spence, who has been always friendly to me, for his guidance in and out of the classroom. I would like to thank Dr. Gary Blanchard for his various help in literally every aspect even since before I became a graduate student. I want to thank Dr. Greg Swain for his kindness and guidance with humor. I thank all my committees for pushing me to a point where I can realize my weak spot so I can further improve myself academically.

I sincerely appreciate the help from our collaborators: Dr. Heedeok Hong, Dr. Xuefei Huang, Dr. Chao Wei, Dr. Chen Chen, and Dr. Jose Cibelli from Michigan State University, Dr. Katelyn Ludwig and Dr. Amanda Hummon from Ohio State University, Dr. James Xia from CMP Scientific, Dr. Vic Spicer and Dr. Oleg V.

Krokhin from the University of Manitoba, Dr. Xiaowen Liu from IUPUI, Dr. Fan Zhang and Dr. Daniela Tomazela from Merck, Dr. Yansheng Liu and Dr. Wenxue Li from Yale University, Jake Melby, Dr. Yanlong Zhu and Dr. Ying Ge from University of Wisconsin.

I want to thank my group members. Xiaojing has helped me a lot including finding a place to live on my first day in Michigan. Eli and Rachele took me to recreations like sports, clubs, NBA games which helped me relax a lot outside of the lab. Zhichang shared lots of valuable experiences in lab work and life. I often asked Tian, Qianyi, and Qianjie to help me with some lab works during the pandemics. They all helped me a lot in or out of the lab. I truly appreciate their kindness and friendship.

I would like to thank Jiaqi Yao, Zhilin Hou, Fangchun Liang, Mengxia Sun, Yijin Zhang, Li Zheng, Zhen Li, Kunli Liu, Ke Ma, Linqing Mo, Nathan Landes, and all other friends. A special thanks go to Dr. Shujia Dai, my former supervisor in Sanofi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

2D-PAGE	Two-dimensional polyacrylamide gel electrophoresis
AA	Acetic acid
ACN	Acetonitrile
AGC	Automatic gain control
APS	Ammonium persulfate
BCA	Bicinchoninic acid
BGE	Background electrolyte
BSA	Bovine serum albumin
BUP	Bottom-up proteomics
CDC	Centers for Disease Control and Prevention
CE	Capillary electrophoresis
CGE	Capillary gel electrophoresis
CID	Collision induced dissociation
CIEF	Capillary isoelectric focusing
C-Score	Characterization Score
CZE	Capillary zone electrophoresis
DDA	Data dependent acquisition
DMA	Dimethylacetamide
DTT	Dithiothreitol
E. coli	*Escherichia coli*
ECD	Electron capture dissociation
EMRD	Electrophoretic mobility relative difference
EOF	Electroosmotic flow

ERLIC	Electrostatic repulsion hydrophilic interaction chromatography
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FA	Formic acid
FASP	Filter-aided sample preparation
FASS	Field-amplified sample stacking
FBS	Fetal bovine serum
FDR	False discovery rate
FWHM	Full width at half maximum
HCD	Higher-energy collisional dissociation
HF	Hydrofluoric acid
HILIC	Hydrophilic interaction chromatography
HPC	Hydroxypropyl cellulose
hpf	Hours post fertilization
i.d.	Inner diameter
IAA	Iodoacetamide
ID	Identification
IEC	Ion-exchange chromatography
IPA	Isopropanol
iTRAQ	Isobaric tags for relative and absolute quantification
IVF	In vitro fertilization
LB	Luria-Bertani
LC	Liquid chromatography
LE	Leading electrolyte
LIT	Linear ion trap
LPA	Linear polyacrylamide

m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser induced ionization
MBT	Mid-blastula transition
MDLC	Multi-dimensional liquid chromatography
MEKC	Micellar electrokinetic capillary chromatography
MO	Morpholino
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MudPIT	Multidimensional protein identification technology
MW	Molecular weight
MZT	Maternal-zygotic-transition
nanoRPLC	RPLC with reduced-inner-diameter columns
NH_4HCO_3	Ammonium bicarbonate
o.d.	Outter diameter
PBS	Phosphate-buffered saline
pI	Isoelectric point
PrSM	Proteoform-spectrum match
PSM	Peptide-spectrum match
PTM	Post-translational modifications
Q	Quadrupole
RF	Radio frequency
RPLC	Reversed phase liquid chromatography
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
STR	Short tandem repeat
TE	Terminating electrolyte

timsTOF	Trapped ion mobility-spectrometry-TOF
tITP	Transient isotachophoresis
TMT	Tandem mass tags
WAX	Weak anion exchange chromatography
ZGA	Zygotic genome activation
μ_{ef}	Electrophoretic mobility

CHAPTER 1. Introduction

1.1. Proteomics

Proteins serve as the functional basics in a vast number of biological processes. Comprehensive characterization of protein cellular localization, post-translational modifications (PTMs), interactions, and expression levels under perturbation can advance our understanding of molecular mechanisms in many biological activities. However, the extremely high complexity underlined the technical challenge of complete investigation of a proteome. After various biological events such as single amino acid polymorphisms, alternative splicing, truncations, and PTMs, one gene can produce multiple forms of protein molecules, termed as “proteoforms” [1,2]. Based on estimation, the approximate 20000 protein-coding genes in the human genome can produce well over 1 million proteoforms [2,3]. Although comprehensive delineation of complex proteomes is still an analytical challenge, mass spectrometry (MS)-based proteomics has made tremendous progress for the characterization of proteins and proteoforms in complex samples in the last decades.

Proteomics is a study of proteomes with the dedication of understanding all protein sequences, PTMs, interactions, and functions. In the 1980s, the milestone development of two soft ionization techniques, matrix-assisted laser-induced ionization (MALDI) and electrospray ionization (ESI), has fueled the MS-based analysis of biomolecules [4,5]. Nowadays, proteomics is widely used in fields of forensics, cancer biology, hematology, development biology, and so forth [6-15]. A typical MS-based proteomics workflow comprises 5 general steps: protein exaction and preparation, liquid phase separation, biomolecule ionization, parent and

fragment ions measurement, and database searching for protein identification. Depending on whether enzymatic digestion of proteins is involved in sample preparation or not, proteomics is branched into two strategies, i.e., bottom-up and top-down, **Figure 1.1** [16].

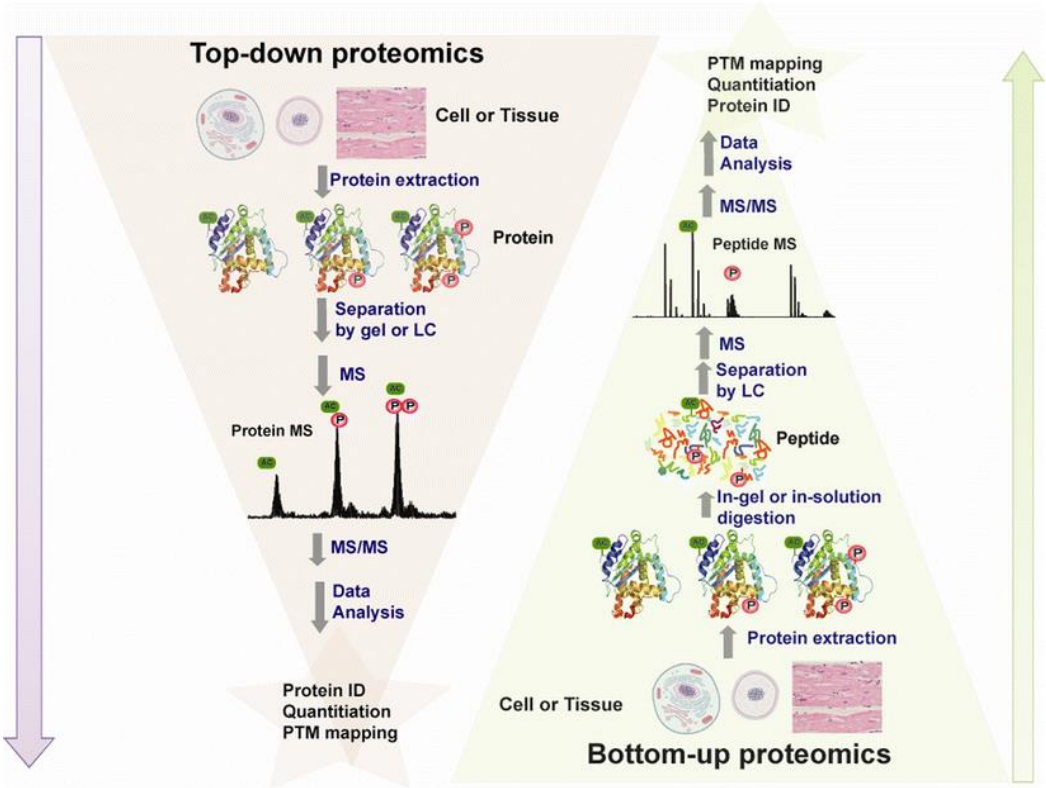

Figure 1.1. Schematic illustration of the difference between top-down and bottom-up proteomics. The figure is reprinted with permission from reference [16].

1.1.1. Bottom-up proteomics

Bottom-up proteomics (BUP) measures the product peptides of enzymatic digestion of proteins. In a classic BUP workflow, **Figure 1.1(Right)**, proteins are extracted from cells or tissues followed by either in-gel or in-solution enzymatic digestion. One or multiple dimensional liquid-phase separations of peptides are then performed followed by soft ionization of enzymatic digests. The tandem MS (MS/MS)-based

mass spectra collection of peptides is performed in two steps. First, a survey mass spectrum where the signal of all parent ions at a certain time is acquired. Second, ions having the top N (usually 10 to 12) intense signals in the survey mass spectrum are isolated for gas-phase fragmentation in sequence. Through *in-silico* digestion of the corresponding protein database that is established by the known genome sequence, a theoretical tandem mass spectra database is generated. In the database searching step, tandem mass spectra containing the fragment ions information are compared to the theoretical mass spectra database and scored. Peptides assigned with high matching scores are considered as confident identifications (IDs). The protein ID is achieved by the alignment between identified peptide sequence and protein sequence. A peptide can be annotated to one unique protein or several proteins that are typically referred to as a protein group. Noteworthily, each mass spectrum scan only consumes milliseconds. Therefore, BUP has high-throughput and compatibility with most liquid chromatography (LC) methods where the average peak width is on a sub-minute scale.

1.1.2. Peptide enrichment, separation, and fractionation methods in BUP

Reduction of proteome complexity is essential to approach large-scale peptide IDs in BUP. The protein abundance dynamic range of a complex human serum proteome is up to 12 orders of magnitude [17]. The *in-silico* digestion of a yeast proteome, which theoretically contains far fewer gene products than the human serum, generated about 300,000 peptides [18]. Although modern mass spectrometers can acquire tens of mass spectra within 1 second, the extremely high complexity of the peptide pool makes large-scale profiling of a proteome impossible by using MS alone. Therefore, methods including enrichment of peptides of interest,

chromatographic separation, and pre-fractionation are frequently employed to decrease the complexity and diversity of the peptide pool.

PTMs on proteins like methylation, glycosylation, phosphorylation, etc., play crucial roles in various cellular processes such as cell differentiation, cellular signaling, and metabolism [19-24]. However, post-translationally modified proteins are usually dynamic and low abundant. Enrichment of post-translationally modified peptides is essential for the global analysis of PTMs.

In BUP studies on phosphorylated proteins, namely phospho-proteomics, a variety of strategies has been established for the enrichment of phosphorylated peptides based on their inherent characteristics such as immunoaffinity, charge, hydrophobicity, and Lewis basicity [25-33]. Antibodies have shown good specificity for the immunoaffinity-based enrichment of low-abundant phosphorylated peptides [25, 26]. However, they are expensive and not always commercially available, thus not suited for the global study of phosphorylation. LC methods based on analyte charge and hydrophobicity, e.g., ion-exchange chromatography (IEC) [27, 28], and hydrophilic interaction chromatography (HILIC) are relatively cost-effective and high-throughput but also suffered from limited specificity for phosphorylated peptide enrichment [28-30]. As well known, a phosphoryl group shows strong Lewis base properties so that it can interact with metal cations, which are typical Lewis acids, via electrostatic interaction and/or chelation. A great number of immobilized metal affinity materials, especially titanium-based materials, have shown high specificity and high throughput for phosphorylated peptide enrichment [31-33].

Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) had been widely used for protein separation in the early era of BUP. In the two dimensions of

2D-PAGE, proteins are separated based on molecular weight (MW) and isoelectric point (pI), respectively. Stained gel pieces where proteins underlie are cut out and subjected to in-gel enzymatic digestion. 2D-PAGE has a laborious and time-consuming workflow with the risk of losing proteins that have extreme MWs and pIs [34]. In the contrast, LC-based methods provide unbiased separation with high separation efficiency, wide separation window, and high peak capacity. That is why LC, especially reversed-phase (RP)LC has been substantially developed and became the dominant separation method in BUP. As an alternative separation method to LC, capillary electrophoresis (CE) has also been applied in BUP studies and has shown some unique features [32, 35, 36].

Although LC is a robust separation method, single-dimensional separation still cannot offer large-scale protein IDs with high protein sequence coverages. Additional separation obtained by offline peptide prefractionation or online multi-dimensional separations is frequently employed in large-scale BUP analysis. Fractionation methods prior to RPLC-MS/MS analysis are frequently employed off-line due to the use of MS-incompatible buffers. Methods like strong cation exchange (SCX) [37, 38, 45], weak anion exchange chromatography (WAX) [39], HILIC [40, 45], high-pH RPLC [41, 42, 45], electrostatic repulsion hydrophilic interaction chromatography (ERLIC) [43] and free-flow electrophoresis [44] have shown their good orthogonality to RPLC. Nevertheless, the protein ID numbers in studies might have compromised from the extensive sample handling in offline fractionation. The development of online multidimensional separation embarked on in the early twenty-one century. By then, the Yates group established the multidimensional protein identification technology (MudPIT) online coupling SCX and RPLC for peptide separation first by charge and then by hydrophobicity [46, 47]. The pioneering study inspired many

developments of two- and three-dimensional LC platforms that are routinely used in proteomics. Further discussions of separation and fractionation methods are available in the **Peptide separation methods** section.

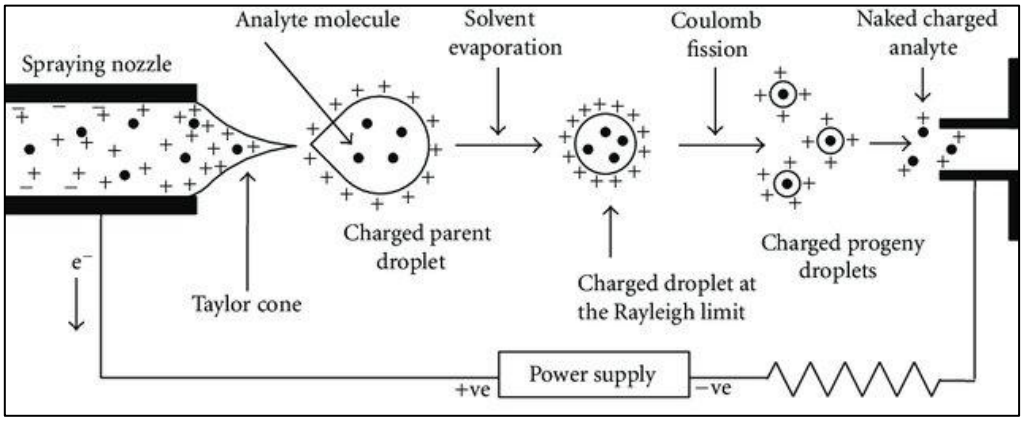

Figure 1.2. Schematic representation of the electrospray ionization process. The figure is reprinted with permission from reference [48].

1.1.3. Mass spectrometers

Analysis of large biomolecules had been crippled by the instability of ionization. This gap was bridged by the Nobel-prize-winning developments of ESI and MALDI [4, 5]. ESI is a popular ionization method due to its feasibility coupling to different liquid phase separations. As depicted in **Figure 1.2**, a liquid Taylor cone is first formed at the end of the separation after the spray voltage application [48]. Followed by the formation of a charged parent droplet. Then solvent evaporation and coulomb fission take place and split the parent droplet into smaller charged droplets that carry analytes. Eventually, the desolvated analytes fly into a mass spectrometer for detection. The implementation of soft ionization methods has enabled extensive studies of proteins on large scales.

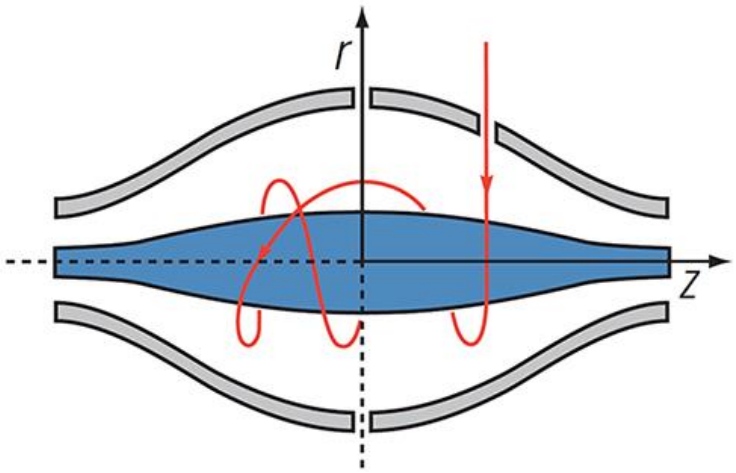

Figure 1.3. Orbitrap mass analyzer. This picture is reprinted from (https://www.creative-proteomics.com/support/q-exactive-hybrid-quadrupole-orbitrap-mass-spectrometer.htm)

In addition to ionization, crucial optimizations in mass analyzers ensured detections with high sensitivity and high mass accuracy. Mass analyzers including linear ion trap (LIT), Orbitrap, quadrupole (Q), TOF are commonly used in BUP. In a LIT mass analyzer, peptide ions are trapped radically by 2D radio frequency (RF) and axially by the stopping potential on both end electrodes. Trapped peptide ions are ejected by the increasing RF voltages. The mass resolution of LIT is about 2 k with a mass accuracy of around 1000 ppm [34]. Limited mass resolution and accuracy burdened LIT from separating peptide ions with similar m/z. The demand for high-resolution m/z separation was met by high-resolution mass analyzers such as Orbitrap. In 2000, upon the Kingdon trap model, Makarov developed the Orbitrap mass analyzer in which peptide ions are trapped between a spindle-like inner electrode and an outer electrode [49]. As illustrated in **Figure 1.3**, ions are circulating the inner electrode meanwhile oscillating back and forth axially in a harmonic fashion. The frequency of ion oscillation is proportional to $(m/z)^{½}$. Ion frequency is detected by the image current detector located on the outer electrode and then transferred to ion signals via Fourier transformation. The Orbitrap can provide mass

resolution up to half a million with approximate 1 ppm mass accuracy ensuring accurate peptide mass assignment and sufficient separation power for peptide ions.

On top of the development of high-resolution mass analyzers, hybrid mass spectrometers nowadays such as Q-Orbitrap [50], and Q-Orbitrap-LIT [51], showed their outstanding performances in BUP due to their fast scan rates and multi-principal capabilities. Furthermore, the interfacing of external ion mobility spectrometry to mass spectrometers greatly improved BUP in noise reduction, sensitivity, dynamic range, and coverage of proteome [52-56]. More recently, the trapped ion mobility-spectrometry-TOF (timsTOF) mass spectrometer, integrating the gas separation power of the trapped ion mobility spectrometry, the fast scan rate, high mass accuracy, and high resolution of the Q-TOF, demonstrated its robustness in BUP with over 2500 protein IDs from sorely 10 ng of Hela proteome digest [57].

Sufficient peptide fragmentation is another prerequisite for confident peptide/protein ID. Peptide fragmentation methods with different principles have been integrated with mass spectrometers to produce informative spectra of complementary fragment ions. A widely used peptide fragmentation method is collision-induced dissociation (CID). In CID, accelerated peptides that have high ion kinetic energy collide with neutral gas. The collision transfers the kinetic energy to internal energy resulting in the breakage of peptide bonds. In the Orbitrap mass spectrometer, the ions were introduced into the higher-energy collision cell where the higher-energy C-trap dissociation (HCD) takes place. The resultant fragment ions are then accumulated in the C-trap followed by the measurement by the Orbitrap mass analyzer [34, 58, 59].

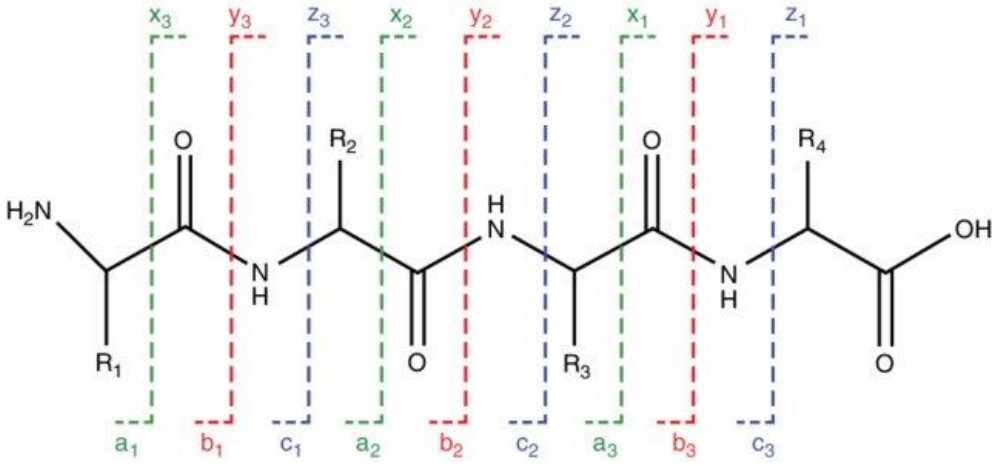

Figure 1.4. The nomenclature for peptide fragmentation. The figure is reprinted with permission from reference [59].

Another well-known strategy is electron-based fragmentation including electron transfer dissociation (ETD) [60], and electron capture dissociation (ECD) [61]. Both fragmentation processes are initiated by interaction between an ion with multiply charges and a free electron either from a low energy electron pool in ECD, or an electron from an anion radical in ETD. The interactions form unstable odd-electron ions which will spontaneously be fragmented. Based on their principles, ETD and ECD provide peptide fragmentation with higher sequence coverage than CID or HCD. ETD and ECD can better conserve labile PTMs leading to better PTM localization of PTMs [62]. Both HCD and CID predominantly produce b- and y-ions when ETD and ECD form c-, z- ions [34], **Figure 1.4**. Noteworthily, HCD forms low-mass ions such as a2, b2, y1, and y2 ions which are not routinely observed in CID mass spectra [63]. Therefore, HCD outperforms CID in the identification of low-mass reporter ions in isobaric labeling-based quantitative proteomics.

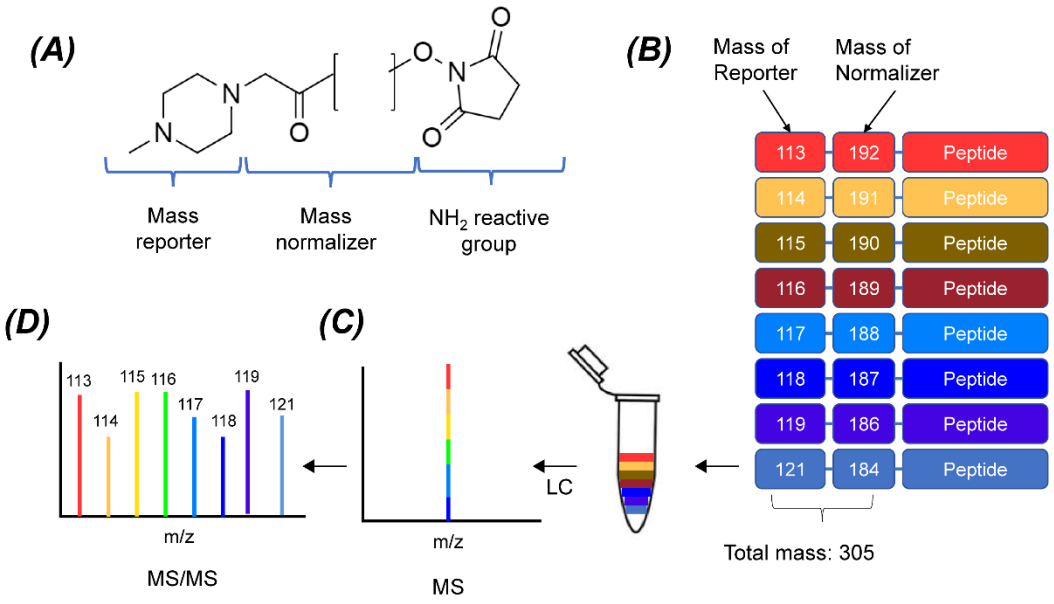

Figure 1.5. Schematic representation of the Isobaric labeling-based quantitation. (A) Structure backbone of isobaric labeling reagent. (B) Example labeling scheme of 8-plex labeling reagent. (C) Illustration of identical peptides labeled with different channels; each color represents a channel of labeling reagent. (D) An example tandem mass spectrum of isotopic mass reporters.

1.1.4. Isobaric labeling for quantitative proteomics

Isobaric labeling is a quantitative proteomics strategy that allows simultaneous analysis of multiple samples, **Figure 1.5**. Two most used multi-channel labeling reagents, isobaric tags for relative and absolute quantification (iTRAQ) [64] and tandem mass tags (TMT) [65], both consist of three major parts: the primary amine-reactive group, the mass reporter group, and the mass balance group, **Figure 1.5A**. All channels in a set of multi-channel reagents have the same total mass, **Figure 1.5B**. Samples are labeled with reagents of different channels via NHS-ester reaction and then pooled together followed by LC separation, where identical peptides labeled with different channels are co-eluted and inseparable by MS due to their same molecular masses, **Figure 1.5C**. During the fragmentation, the isotopic

mass reporters are released, and their intensities represent the relative abundance of the corresponding peptide between different biological conditions, **Figure 1.5D**.

1.2. Peptide separation methods

Ion suppression first reported by Buhrman *et al*. is a form of matrix effect [66]. The researchers found that the ion intensity of the spike-in compound had an inverse relationship to the diversity of the matrix in ESI. In theory, matrix compounds that coelute with the analyte can suppress the ionization of the analyte which leads to inaccuracy and low sensitivity of the MS-based quantification. Therefore, robust peptide separation methods that are compatible with ESI-MS are much-needed for reducing the ion suppression effect in BUP.

1.2.1. One dimensional LC-based methods

RPLC is the most used separation scheme in BUP, attributed to its buffer compatibility with MS. In the RP separation mode, peptides are separated based on their partitioning between the stationary phase and mobile phase that typically has an increasing concentration of organic solvent (i.e., acetonitrile). Peptides are eluted from an RPLC column by the order of increasing hydrophobicity. Reducing the inner diameter (i.d.) can boost the ionization efficiency of ESI, the consequently the sensitivity of a RPLC-MS/MS platform [67-69]. In a recent study, Xiang *et al.* reduced the column i.d. to 2 μm and obtained a flow rate at the pico-liter per minute range [70]. As a result, they identified near 1000 proteins from sorely 75 pg of sample proving the high sample concentration sensitivity of nanoRPLC columns. As is well-known, longer LC columns provide better separation efficiency. In nanoRPLC-

MS/MS BUP experiments, columns shorter than 20 cm are frequently used [37,40,43,45]. Using a 1-meter-long nanoRPLC column, Zhou *et al.* identified over 4000 proteins with a high separation peak capacity of 700 in 10 hours [71]. In a milestone study, the Coon lab demonstrated the separation power of RPLC. A 35-cm long nanoRPLC column (i.d. 75 μm) was used for separation, the RPLC-MS/MS platform to identify around 4000 yeast proteins with an only 70-min gradient when the yeast proteome temporally contains a total of 4500 proteins based on estimation [42].

As an alternative mode for peptide separation, HILIC has been used for peptide fractionation and rarely coupled with MS for applications in BUP [72, 73]. In the HILIC mode, peptides are loaded onto the HILIC column by a high concentration of organic solvent and interact with the ionic stationary phase within the column. Elution in HILIC was done by the increasing concentration of the hydrophilic mobile phase. HILIC and RPLC have the inverted peptide elution orders between each other, but the same buffer compatibility to MS. ERLIC, a variation of HILIC, has been directly coupled with MS for BUP or peptide mapping of an antibody [74,75].

1.2.2. Peptide fractionation and multi-dimensional LC

Although one-dimensional RPLC has proven its robustness and reproducibility in peptide separation, the separation was merely based on hydrophobicity resulting in failures to resolve peptides with similar hydrophobicity. Theoretically, the human proteome contains over twenty thousand gene products. One-dimensional LC-MS/MS platform can routinely identify thousands of proteins covering less than 50% of human proteome because of its limited peak capacity and insufficient separation

power. Therefore, peptide fractionation or multi-dimensional LC (MDLC) is required for extensive applications of proteomics in studies of variety.

The principle in peptide fractionation or MDLC is the orthogonality between separation mechanisms. Since RPLC usually serves as the last dimension of separation, other dimensions are better to be providing different peptide retention mechanisms other than hydrophobicity. For example, in the pioneer study reported by the Yates group, peptides were separated first by charge in the SCX column and then by hydrophobicity in the RPLC column [47]. SCX-RPLC is a commonly used online separation platform in BUP [46, 47, 76-78]. ERLIC was also involved for iTRAQ-labelled peptide fractionation and demonstrated comparable or even better performance than SCX [79]. More recently, a study has revealed that SCX and HILIC have lower orthogonality to low-pH RPLC as compared to high-pH RPLC [80]. The increase in buffer pH alters peptide hydrophobicity in the first dimensional RPLC separation which brings good orthogonality to the second dimension of the RPLC separation. This pH mediated 2D-RP separation strategy facilitated the completion of a draft map of the human proteome [81]. The draft map annotated over 17000 gene products covering 84% of the annotated genes. Three-dimensional LC configuration was also explored towards peptide ID comprehensiveness and received great improvement [82, 83]. However, high proteome coverage requires better peptide separation to boost the coverage of the proteome. An alternative method that offers orthogonal peptide separation to the existing LC methods is desired for further improvement.

1.3. Capillary electrophoresis

1.3.1. Capillary zone electrophoresis for peptide separation

CE has attracted a lot of attention as an alternative separation approach for BUP in the past decade. CE has different modes including capillary isoelectric focusing (CIEF), capillary zone electrophoresis (CZE), capillary gel electrophoresis (CGE), micellar electrokinetic capillary chromatography (MEKC), etc. Due to the buffer compatibility, CZE is the most popular mode for peptide separation.

In CZE, analytes are separated based on their electrophoretic mobility (μ_{ef}) that is determined by charge-to-size ratios [84]. In a conventional setting, a fused silica open-tubular capillary with 10 to 75 μm i.d. and 20 to 100 cm length is employed for the CZE separation. The total mobility of an analyte is the sum of electrophoretic mobility (μ_{ef}) and electroosmotic mobility (μ_{eof}).

$$\mu_{total} = \mu_{ef} + \mu_{eof} \quad (1)$$

Based on the Debye-Hückel-Henry theory, the μ_{ef} is determined by

$$\mu_{ef} = \frac{ze}{6\pi\eta r} \quad (2)$$

Where z is the analyte's net charge, e is the elementary charge, η is the viscosity of the background electrolyte (BGE), and r is the analyte's radius. During the separation, the μ_{ef} of positive ions direct towards the cathode, the μ_{ef} of negative ions direct towards anode while the μ_{ef} of neutral ions is equal to 0, **Figure 1.6**. μ_{eof} of all analytes equals the speed of electroosmotic flow (EOF). The inner wall of the fused silica capillary is full of silanol groups that are negatively charged when the pH of BGE is higher than 3. The negative charges attract cations that consequently form

the electrical double layer near the inner wall. After separation voltage application, the outer cation layer moves and forms EOF that carries the BGE along with all analytes towards the cathode. The EOF gives all analytes the same μ_{eof}.

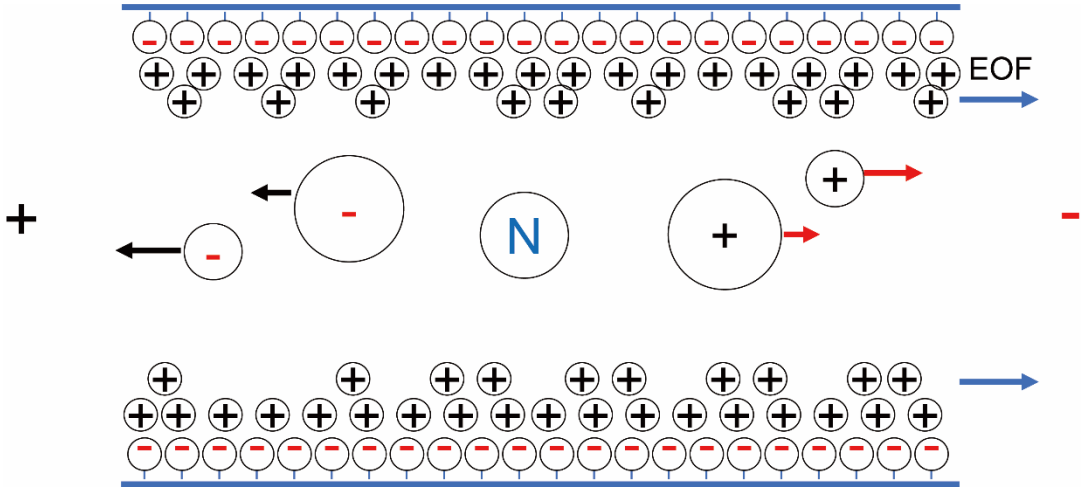

Figure 1.6. CZE separation mechanism.

CZE and RPLC are orthogonal for peptide separation. Li *et al.* found a small overlap in peptide IDs between CZE-MS/MS and nanoRPLC-MS/MS when they were analyzing the tryptic digest of *M. marinum* [85]. A similar small overlap was also found in phospho-proteomics [86]. Besides the orthogonality in peptide separation, CZE also allows accurate prediction of peptides' μ_{ef}. In 2017, Krokhin *et al*. reported a semi-empirical model that achieved linear correlation (R^2~0.995) between predicted μ_{ef} and experimental μ_{ef} [87]. The accurate prediction of peptide μ_{ef} can facilitate or even guide the confident peptide ID. Although CZE has shown great potentials, it requires a mature, stable CE-MS interface. Besides, inherent features of CZE, including narrow separation window and limited loading capacity, have been hampering it from practical applications for large-scale BUP.

1.3.2. CE-MS Interface

Configuration of CE and ESI-MS appears to be more straightforward than CE and MALDI [88]. The hyphenation of CE-ESI-MS demands a close electric circuit and high voltage supply for forming the electrospray. In 1988, the Smith group developed the first coaxial sheath-liquid CE-MS interface [89]. In their design, the end of the capillary is introduced into a stainless-steel spray needle that is filled with sheath liquid. The electric circuit is closed by the contact of sheath liquid and CE eluate. The mixture of sheath liquid and CE eluate is electro-sprayed by applying a voltage to the metal needle. Formation of electrospray is also assisted by neutral sheath gas. This design was pioneering, stable and the foundation of many modern CE-MS interface designs. However, the flow rate of sheath liquid in Smith's model is at least 100-fold higher than that in CZE (μL/min v.s. nL/min), which results in significant sample dilution and compromising of sensitivity.

Another strategy for CE-ESI-MS interfaces, the sheathless strategy, theoretically provides higher sensitivity for sample concentration than the sheath liquid-based strategy. A classic sheathless model was reported by Moini in 2007 [90]. In Moini's design, the outlet end of the capillary is etched by hydrofluoric acid (HF) and becomes porous. The porous end allows the ions to transfer between conductive liquid and BGE. Droplets are formed via electrospray ionization at the exit end of the capillary when a voltage is applied to the outer metal ESI needle that is in contact with the conductive liquid. Although the sheathless interface provides high sensitivity, the fabrication of the capillary with a porous tip and the setup of the interface are not simple and easy to use. The requirement of separation current lower than 10 μA also limited its application for sophisticated BUP analysis.

EOF-driven sheath liquid-based interfaces provided an interesting solution to compensate for the loss in sensitivity of the sheath liquid interfaces [91-93]. As depicted in **Figure 1.7A**, one end of the capillary is threaded through a cross and introduced into a borosilicate glass emitter filled with sheath liquid. The sheath liquid is electrokinetically pumped by the applied voltage in the sheath electrolyte reservoir. Since the flow rate of the sheath liquid is EOF-driven, it is at the same scale with the CE eluate (nL/min), much lower than that in the conventional sheath liquid-based design. By adjusting the diameter of the exit end of the capillary and the electrospray emitter, Sun *et al*. made the capillary end closer to the orifice of the emitter, **Figure 1.7B**. This effort resulted in less sample dilution and further boosted the sample concentration sensitivity. This design has been commercialized by CMP Scientific (https://www.cmpscientific.com/).

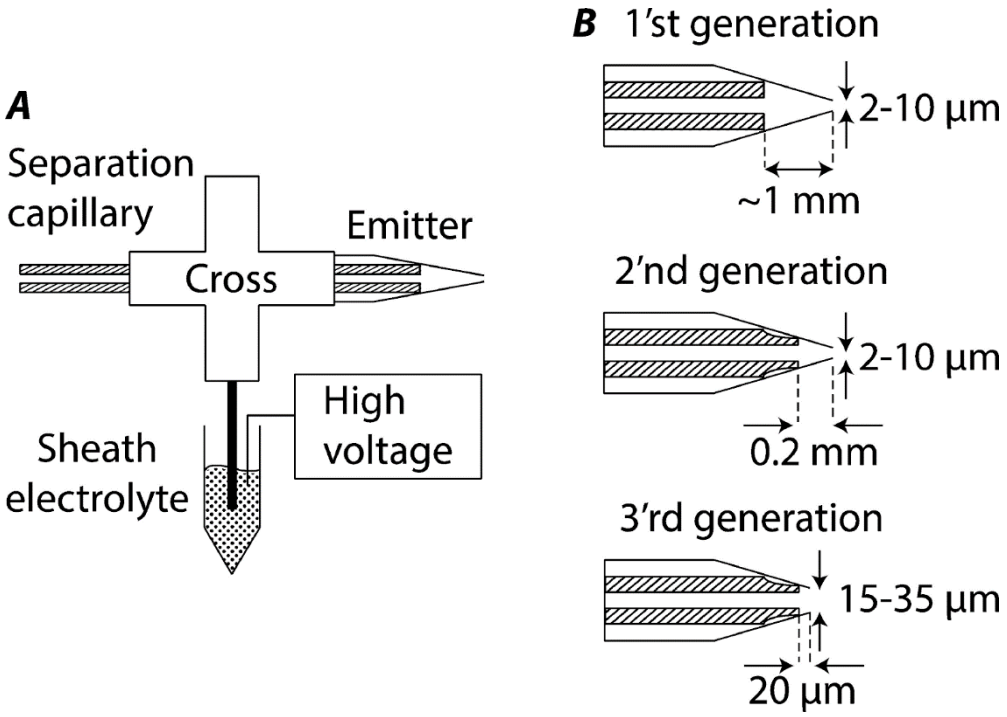

Figure 1.7. Diagrams of the basic design of the electrokinetically pumped sheath flow CE-MS interface (A) and its three different generations (B). The figure is reprinted with permission from reference [93].

1.3.3. Narrow separation window of conventional CZE

In the conventional CZE mode using a fused silica capillary, EOF is fast and drives all analytes out of the capillary quickly. So that, the separation windows using the conventional CZE are usually narrower than 30 min [93-95]. The narrow separation window limits the number of acquirable MS/MS spectra, and consequently the number of protein IDs in large-scale BUP.

Since EOF is the determining factor of the fast migration of analytes, suppression of EOF via neutral coating can widen the separation window for CZE. Hydrophilic and neutral materials such as linear polyacrylamide (LPA) [96, 97] and hydroxypropyl cellulose (HPC) [98, 99] have been employed for capillary coating. The Dovichi group has demonstrated that CZE-MS/MS analysis of bovine serum albumin (BSA) digest using an LPA-coated capillary generated a separation window

exceeding one hour [100], significantly wider than using a fused silica capillary [93-95].

1.3.4. Low loading capacity of CZE

Typically, sample volume cannot exceed 1% of total capillary volume to avoid significant peak broadening in CZE. For instance, the total volume of a capillary with 1-m long, 50-µm i.d. is ~2 µL, meaning that the sample volume is limited to 20 nL. The upper limit of CZE loading capacity is approximately 100-fold lower than nanoRPLC. The extremely low sample amount prohibits the detection of low abundant peptides in a complex mixture. Online sample stacking methods including field-amplified sample stacking (FASS), transient isotachophoresis (tITP), and dynamic pH junction can improve the loading capacity of CZE.

FASS is based on the analyte ion stacking at the interface between low-conductivity sample buffer and high conductivity BGE [101, 102]. To perform FASS, the sample is dissolved in a low-conductivity solution and injected into a capillary filled with high conductivity BGE. When the separation voltage is applied, the high electric field strength over the low-conductivity sample matrix provides high analyte ions velocity. The ions eventually reach the interface between the BGE and sample buffer where they drastically slow down and be stacked in a narrow zone. FASS stacking can significantly improve both the loading capability of CZE and the sensitivity. A sample buffer containing 0.04% (v/v) formic acid (FA) and 30% (v/v) acetonitrile was used to create a low-conductivity sample plug in a study by the Dovichi group [103]. Sample volume was successfully boosted to about 100 nL, equivalent to 5% of the capillary. More importantly, the FASS-based CZE separation

produced 2100 protein IDs from Hela proteome digest with high peak capacity (~300).

In tITP, the sample and the leading electrolyte (LE) are first introduced into a capillary that is filled with BGE. Then a plug of terminating electrolyte (TE) is injected into the capillary. After voltage application, analytes, based on their original concentration and electrophoretic mobilities, are pre-concentrated into narrow zones that have the same migrating velocity between LE and TE. After tITP preconcentration, the boundaries built by LE and TE are disrupted by the surrounding BGE. Analytes are then separated by CZE. Guo *et al.* systematically evaluated the loading capacity, sensitivity, and separation window of a tITP/CZE-MS/MS system [104]. Although they enlarged the sample loading capacity from less than 1% to 32%, they also found a trade-off between the tITP/CZE separation window and loading capacity [104].

Dynamic pH junction developed by the Chen group is based on pH-dependent μ_{ef} changes [105]. Typically, the sample is dissolved in a basic solution (i.e., 50 mM NH_4HCO_3, pH 8) when the BGE is acidic (i.e., 5% (v/v) acetic acid, pH 2.4). After sample injection, two pH boundaries are formed at both ends of the sample plug interfacing with BGE. Within the basic sample solution, the analytes are negatively charged. When the positive separation voltage is applied on the injection end, hydrogen protons start to migrate towards the detection and titrate the basic sample plug. Therefore, the pH boundary near the injection end begins to move towards the other pH boundary when negative analytes move towards the injection end due to the electrostatic force. Once the negative analytes meet the moving pH boundary, their mobilities drastically decrease because of the pH-dependent charge turn-over. Eventually, analytes are concentrated at the moving pH boundary. Until the two pH

boundaries emerge, analytes are separated by CZE. Dynamic pH junction has been used for pre-concentration of samples in CZE and achieved up to 400-nL sample injection volume without losing separation efficiency [106].

1.4. Summary

This chapter introduced MS-based BUP protein studies. Advance in ESI, liquid phase separation, and mass spectrometer has made BUP a powerful tool in the delineation of protein dynamics, PTMs on proteins, protein function, and protein interactions. Current LC-based one- or multi-dimensional separation methods allowed large-scale protein IDs with limited sequence and proteome coverages. To approach the comprehensive characterization of proteomes, CZE as an alternative liquid phase separation method to RPLC has shown its potential for BUP. Improvement in CZE regarding the CE-MS interfaces, capillary coating, online pre-concentration method boosted its performance for BUP with high separation efficiency, good reproducibility, and high sensitivity. The following chapters in this dissertation will discuss further improvement and application of CZE for BUP.

REFERENCES

REFERENCES

[1] Schlüter, H.; Apweiler, R.; Holzhütter, H, G.; Jungblut, P, R. Finding one's way in proteomics: a protein species nomenclature. *Chem. Cent. J.* **2009,** 3, 11.

[2] Smith, LM.; Kelleher, N, L.; Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013,** 10(3), 186-7.

[3] Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek, Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van, Eyk, J, E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How many human proteoforms are there? *Nat. Chem. Biol.* ***2018***, 14(3), 206-214.

[4] Yamashita, M.; Fenn, J. B. Electrospray ion-source—another variation on the free-jet theme. *J. Phys. Chem.* **1984**, 88, 4451–4459.

[5] Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T.; Matsuo, T. Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of-flight Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **1988**, 2 (20), 151–3.

[6] Chu, F.; Mason, K.E.; Anex, D. S.; Jones, A, D.; Hart, B. R. Proteomic Characterization of Damaged Single Hairs Recovered after an Explosion for Protein-Based Human Identification. *J. Proteome Res.* **2020,** 19(8), 3088-3099.

[7] Merkley, E. D.; Wunschel, D. S.; Wahl, K. L.; Jarman, K. H. Applications and challenges of forensic proteomics. *Forensic Sci Int.* **2019**, 297, 350-363.

[8] Tan, H. T.; Lee, Y. H.; Chung, M. C. Cancer proteomics. *Mass Spectrom. Rev.* **2012**, 31(5), 583-605.

[9] Cheung, C. H. Y.; Juan, H. F. Quantitative proteomics in lung cancer. *J Biomed Sci.* **2017**, 24(1), 37.

[10] Doll, S.; Gnad, F.; Mann, M. The Case for Proteomics and Phospho-Proteomics in Personalized Cancer Medicine. *Proteomics Clin. Appl.* **2019**, 13(2), e1800113.

[11] Huang, Z.; Ma, L.; Huang, C.; Li, Q.; Nice, E. C. Proteomic profiling of human plasma for cancer biomarker discovery. *Proteomics* **2017**, 17(6).

[12] Rolland, D. C. M.; Lim, M. S.; Elenitoba-Johnson, K. S. J. Mass spectrometry and proteomics in hematology. *Semin. Hematol.* **2019**, 56(1), 52-57.

[13] Peuchen, E. H.; Sun, L.; Dovichi, N. J. Optimization and comparison of bottom-up proteomic sample preparation for early-stage Xenopus laevis embryos. *Anal. Bioanal. Chem.* **2016**, 408(17), 4743-9.

[14] Lombard-Banek, C.; Portero, E. P.; Onjiko, R. M.; Nemes, P. New-generation mass spectrometry expands the toolbox of cell and developmental biology. *Genesis* **2017**, 55(1-2), 10.

[15] Hashimoto, Y.; Greco, T. M.; Cristea, I. M. Contribution of Mass Spectrometry-Based Proteomics to Discoveries in Developmental Biology. *Adv. Exp. Med. Biol.* **2019**, 1140, 143-154.

[16] Gregorich, Z. R.; Chang, Y. H.; Ge, Y. Proteomics in heart failure: top-down or bottom-up? *Pflugers Arch.* **2014**, 466(6),1199-209.

[17] Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **2002**, 1, 845.

[18] Cagney, G.; Amiri, S.; Premawaradena, T.; Lindo, M.; Emili, A. In silico proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome Sci.* **2003**, 1, 5.

[19] Bedford, M. T.; Clarke, S.G. Protein arginine methylation in mammals: who, what, and why. *Mol. Cell.* **2009**, 33(1), 1-13.

[20] Clarke, S. Protein methylation. *Curr. Opin. Cell Biol.* **1993**, 5(6), 977-83.

[21] Gong, F. Miller, K. M. Histone methylation and the DNA damage response. *Mutat. Res.* **2019**, 780:37-47.

[22] Stowell, S.R.; Ju, T.; Cummings, R. D. Protein glycosylation in cancer. *Annu. Rev. Pathol.* **2015**,10, 473-510.

[23] Clerc, F.; Reiding, K. R.; Jansen, B. C.; Kammeijer, G. S.; Bondt, A.; Wuhrer, M. Human plasma protein N-glycosylation. *Glycoconj. J.* **2016**, 33(3), 309-43.

[24] Humphrey, S. J.; James, D. E.; Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends. Endocrinol. Metab.* **2015,** 26(12), 676-687.

[25] Whiteaker, J. R.; Zhao, L.; Yan, P.; Ivey, R. G.; Voytovich, U. J.; Moore, H. D.; Lin, C.; Paulovich, A. G. Peptide Immunoaffinity Enrichment and Targeted Mass Spectrometry Enables Multiplex, Quantitative Pharmacodynamic Studies of Phospho-Signaling. *Mol. Cell. Proteomics* **2015**, 14(8), 2261-73.

[26] Whiteaker, J. R.; Zhao, L.; Schoenherr, R. M.; Kennedy, J. J.; Ivey, R. G.; Paulovich, A. G. Peptide Immunoaffinity Enrichment with Targeted Mass Spectrometry: Application to Quantification of ATM Kinase Phospho-Signaling. *Methods Mol. Biol.* **2017**, 1599, 197-213.

[27] Villén, J.; Gygi, SP. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* **2008**, 3(10), 1630-8.

[28] Hennrich, ML.; van den Toorn, H. W.; Groenewold, V.; Heck, A. J.; Mohammed, S. Ultra acidic strong cation exchange enabling the efficient enrichment of basic phosphopeptides. *Anal. Chem.* **2012**, 84(4), 1804-8.

[29] Engholm-Keller, K.; Birck, P.; Størling, J.; Pociot, F.; Mandrup-Poulsen, T.; Larsen, M. R. TiSH--a robust and sensitive global phosphoproteomics strategy

employing a combination of TiO2, SIMAC, and HILIC. *J. Proteomics* **2012**, 75(18), 5749-61.

[30] Cui, Y. Yang, K. Tabang, D. N.; Huang, J.; Tang, W.; Li, L. Finding the Sweet Spot in ERLIC Mobile Phase for Simultaneous Enrichment of N-Glyco and Phosphopeptides. *J. Am. Soc. Mass. Spectrom.* **2019**, 30(12), 2491-2501.

[31] Lundby, A.; Secher, A.; Lage, K.; Nordsborg, N. B.; Dmytriyev, A.; Lundby, C.; Olsen, JV. Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nat. Commun.* **2012**, 3, 876.

[32] Ludwig, K. R.; Sun, L.; Zhu, G.; Dovichi, N. J.; Hummon, A. B. Over 2300 phosphorylated peptide identifications with single-shot capillary zone electrophoresis-tandem mass spectrometry in a 100 min separation. *Anal. Chem.* **2015,** 87(19), 9532-7.

[33] Leitner, A. Enrichment Strategies in Phosphoproteomics. *Methods Mol. Biol.* **2016,** 1355, 105-21.

[34] Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R, 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013,** 113(4), 2343-94.

[35] Sun, L.; Zhu, G.; Yan, X.; Zhang, Z.; Wojcik, R.; Champion, M. M.; Dovichi N. J. Capillary zone electrophoresis for bottom-up analysis of complex proteomes. *Proteomics* **2016,** 16(2), 188-96.

[36] Wojcik, R.; Zhu, G.; Zhang, Z.; Yan, X.; Zhao, Y.; Sun, L.; Champion, M. M.; Dovichi, N. J. Capillary zone electrophoresis as a tool for bottom-up protein analysis. *Bioanalysis* **2016**, 8(2), 89-92.

[37] Lau, E.; Lam, M. P.; Siu, S. O.; Kong, R. P.; Chan, W. L.; Zhou, Z.; Huang, J.; Lo, C.; Chu, I. K. Combinatorial use of offline SCX and online RP-RP liquid chromatography for iTRAQ-based quantitative proteomics applications. *Mol. Biosyst.* **2011**, 7(5), 1399-408.

[38] Shen, Y.; Jacobs, J. M.; Camp, D. G 2nd.; Fang, R.; Moore, R. J.; Smith, R. D.; Xiao, W.; Davis, R. W.; Tompkins, R. G. Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. Chem.* **2004,** 76(4), 1134-44.

[39] Motoyama, A.; Xu, T.; Ruse, C. I.; Wohlschlegel, J. A.; Yates, J. R, 3rd. Anion and cation mixed-bed ion exchange for enhanced multidimensional separations of peptides and phosphopeptides. *Anal. Chem.* **2007**, 79(10), 3623-34.

[40] Di Palma, S.; Boersema, P. J.; Heck, A. J.; Mohammed, S. Zwitterionic hydrophilic interaction liquid chromatography (ZIC-HILIC and ZIC-cHILIC) provide high resolution separation and increase sensitivity in proteome analysis. *Anal. Chem.* **2011**, 83(9), 3440-7.

[41] Zhou, F.; Cardoza, J. D.; Ficarro, S. B.; Adelmant, G. O.; Lazaro, J. B.; Marto, J. A. Online nanoflow RP-RP-MS reveals dynamics of multicomponent Ku complex in response to DNA damage. *J. Proteome Res.* **2010**, 9(12), 6242-55.

[42] Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The one hour yeast proteome. *Mol. Cell. Proteomics* **2014**, 13(1), 339-47.

[43] Hao, P.; Qian, J.; Ren, Y.; Sze, S. K. Electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) versus strong cation exchange (SCX) for fractionation of iTRAQ-labeled peptides. *J. Proteome Res.* **2011**, 10(12), 5568-74.

[44] Malmström, J.; Lee, H.; Nesvizhskii, AI.; Shteynberg, D.; Mohanty, S.; Brunner, E.; Ye, M.; Weber, G.; Eckerskorn, C.; Aebersold, R. Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **2006**, 5(9), 2241-9.

[45] Yeung, D.; Mizero, B.; Gussakovsky, D.; Klaassen, N.; Lao, Y.; Spicer, V.; Krokhin, OV. Separation Orthogonality in Liquid Chromatography-Mass Spectrometry for Proteomic Applications: Comparison of 16 Different Two-Dimensional Combinations. *Anal. Chem.* **2020**, 92(5), 3904-3912.

[46] Wolters, D. A.; Washburn, M. P.; Yates, J.R, 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **2001**, 73(23), 5683-90.

[47] Washburn, M. P.; Wolters, D.; Yates, J. R, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, 19(3), 42-7.

[48] Banerjee, S.; Mazumdar, S. Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. *Int. J. Anal. Chem.* **2012**, 282574.

[49] Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **2000**, 72(6), 1156-62.

[50] Kelstrup, C. D.; Bekker-Jensen, D. B.; Arrey, T. N.; Hogrebe, A.; Harder, A.; Olsen, J. V. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **2018**, 17(1), 727-738.

[51] Senko, M. W.; Remes, P. M.; Canterbury, J. D.; Mathur, R.; Song, Q.; Eliuk, S. M.; Mullen, C.; Earley, L.; Hardman, M.; Blethrow, J. D.; Bui, H.; Specht, A.; Lange, O.; Denisov, E.; Makarov, A.; Horning, S.; Zabrouskov, V. Novel parallelized quadrupole/linear ion trap/Orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. *Anal. Chem.* **2013**, 85(24), 11710-4.

[52] 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.

[53] Bekker-Jensen, D. B.; Martínez-Val, A.; Steigerwald, S.; Rüther, P.; Fort, K. L.; Arrey, T. N.; Harder, A.; Makarov, A.; Olsen, J. V. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol. Cell. Proteomics* **2020**, 19(4), 716-729.

[54] Saba, J.; Bonneil, E.; Pomiès, C.; Eng, K.; Thibault, P. Enhanced sensitivity in proteomics experiments using FAIMS coupled with a hybrid linear ion trap/Orbitrap mass spectrometer. *J. Proteome Res.* **2009**, 8(7), 3355-66.

[55] Canterbury, J. D.; Yi, X.; Hoopmann, M. R.; MacCoss, M. Assessing the dynamic range and peak capacity of nanoflow LC-FAIMS-MS on an ion trap mass spectrometer for proteomics. *Anal. Chem.* **2008**, 80(18), 6888-97.

[56] 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), M111.014985.

[57] Meier, F.; Brunner, A. D.; Frank, M.; Ha, A.; Bludau, I.; Voytik, E.; Kaspar-Schoenefeld, S.; Lubeck, M.; Raether, O.; Bache, N.; Aebersold, R.; Collins, BC.; Röst, H. L.; Mann, M. diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. *Nat. Methods* **2020**, 17(12), 1229-1236.

[58] Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, 4(9):709-12.

[59] Hao, Z.; Hong, Q.; Zhang, F.; Wu, S. L. Bennett, P. Current Methods for the Characterization of Posttranslational Modifications in Therapeutic Proteins Using Orbitrap Mass Spectrometry, John Wiley & Sons, Inc., **2017**.

[60] Kim, M. S.; Pandey, A. Electron transfer dissociation mass spectrometry in proteomics. *Proteomics* **2012,** 12(4-5), 530-42.

[61] Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *J. Am. Chem. Soc.* **1998**, 120, 3265-3266.

[62] Sarbu, M.; Ghiulai, R. M.; Zamfir, A. D. Recent developments and applications of electron transfer dissociation mass spectrometry in proteomics. *Amino Acids* **2014**, 46(7):1625-34.

[63] McAlister, G. C.; Phanstiel, D. H.; Brumbaugh, J.; Westphall, M. S.; Coon, J. J. Higher-energy collision-activated dissociation without a dedicated collision cell. *Mol. Cell. Proteomics* **2011**, 10(5), O111.009456.

[64] Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, 3(12), 1154-69.

[65] Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem*. **2003**, 75(8), 1895-904.

[66] Buhrman, D. L.; Price, P. I.; Rudewiczcor, P. J. Quantitation of SR 27417 in human plasma using electrospray liquid chromatography-tandem mass spectrometry: A study of ion suppression. *J. Am. Soc. Mass Spectrom.* **1996**, 7(11), 1099-105.

[67] Zhou, F.; Lu, Y.; Ficarro, S. B.; Webber, J. T.; Marto, J. A. Nanoflow low pressure high peak capacity single dimension LC-MS/MS platform for high-throughput, in-depth analysis of mammalian proteomes. *Anal. Chem.* **2012**, 84(11), 5133-9.

[68] Vissers, J. P. Recent developments in microcolumn liquid chromatography. *J. Chromatogr. A* **1999**, 856(1-2), 117-43.

[69] Masuda, T.; Sugiyama, N.; Tomita, M.; Ishihama, Y. Microscale phosphoproteome analysis of 10,000 cells from human cancer cell lines. *Anal. Chem.* **2011**, 83, 7698.

[70] Xiang, P.; Zhu, Y.; Yang, Y.; Zhao, Z.; Williams, S. M.; Moore, R. J.; Kelly, R. T.; Smith, R. D.; Liu, S. Picoflow Liquid Chromatography-Mass Spectrometry for Ultrasensitive Bottom-Up Proteomics Using 2-μm-i.d. Open Tubular Columns. *Anal. Chem.* **2020**, 92(7), 4711-4715.

[71] Zhou, F.; Lu, Y.; Ficarro, S. B.; Webber, J. T.; Marto, J. A. Nanoflow low pressure high peak capacity single dimension LC-MS/MS platform for high-throughput, in-depth analysis of mammalian proteomes. *Anal. Chem.* **2012**, 84(11):5133-9.

[72] Boersema, P. J.; Mohammed, S.; Heck, A. J. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Anal. Bioanal. Chem.* **2008**, 391(1), 151-9.

[73] Bensaddek, D.; Nicolas, A.; Lamond, A. I. Evaluating the use of HILIC in large-scale, multi dimensional proteomics: Horses for courses? *Int. J. Mass Spectrom.* **2015**, 391, 105-114.

[74] de Jong, E. P.; Griffin, T. J. Online nanoscale ERLIC-MS outperforms RPLC-MS for shotgun proteomics in complex mixtures. *J. Proteome Res.* **2012**, 11(10):5059-64.

[75] Zhen, J.; Kim, J.; Zhou, Y.; Gaidamauskas, E.; Subramanian, S.; Feng, P. Antibody characterization using novel ERLIC-MS/MS-based peptide mapping. *MAbs.* **2018,** 10(7), 951-959.

[76] Shen, Y.; Jacobs, J. M.; Camp, D. G., 2nd; Fangqq, R.; Moore, R. J.; Smith, R. D.; Xiao, W.; Davis, R. W.; Tompkins, R. G. Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. Chem.* **2004**, 76, 1134.

[77] Peng, J.; Schwartz, D.; Elias, J. E.; Thoreen, C. C.; Cheng, D.; Marsischky, G.; Roelofs, J.; Finley, D.; Gygi S. P. A proteomics approach to understanding protein ubiquitination. *Nat. Biotechnol.* **2003**, 21(8):921-6.

[78] Zhu, M. Z.; Li, N.; Wang, Y. T.; Liu, N.; Guo, M. Q.; Sun, B. Q.; Zhou, H.; Liu, L.; Wu, J. L. Acid/Salt/pH Gradient Improved Resolution and Sensitivity in Proteomics Study Using 2D SCX-RP LC-MS. *J. Proteome Res.* **2017**, 16(9):3470-3475.

[79] Hao, P.; Qian, J.; Ren, Y.; Sze, S. K. J. Electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) versus strong cation exchange (SCX) for fractionation of iTRAQ-labeled peptides. *Proteome Res.* **2011**, 10, 5568

[80] Yeung, D.; Mizero, B.; Gussakovsky, D.; Klaassen, N.; Lao, Y.; Spicer, V.; Krokhin, O. V. Separation Orthogonality in Liquid Chromatography-Mass Spectrometry for Proteomic Applications: Comparison of 16 Different Two-Dimensional Combinations. *Anal. Chem.* **2020**, 92(5), 3904-3912.

[81] Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A., A draft map of the human proteome. *Nature* **2014**, 509(7502), 575-81.

[82] Spicer, V.; Ezzati, P.; Neustaeter, H.; Beavis, R. C.; Wilkins, J. A.; Krokhin, O. V. 3D HPLC-MS with Reversed-Phase Separation Functionality in All Three Dimensions for Large-Scale Bottom-Up Proteomics and Peptide Retention Data Collection. *Anal. Chem.* 2016, 88(5):2847-55.

[83] Zhou, F.; Lu, Y.; Ficarro, S. B.; Adelmant, G.; Jiang, W.; Luckey, C. J.; Marto, J. A. Genome-scale proteome quantification by DEEP SEQ mass spectrometry. *Nat. Commun.* **2013**, 4, 2171.

[84] Jorgenson, J. W.; Lukacs, K. D. Capillary zone electrophoresis. *Science* **1983**, 222(4621), 266-72.

[85] Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J. Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry as an alternative proteomics platform to ultraperformance liquid chromatography-electrospray ionization-tandem mass spectrometry for samples of intermediate complexity. *Anal. Chem.* **2012**, 84(3), 1617-22.

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

[87] Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem.* **2017**, 89(3), 2000-2008.

[88] 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(13), 3814-28.

[89] Smith, R. D.; Udseth, H. R. Capillary zone electrophoresis-MS. *Nature* **1988**, 331(6157), 639-40.

[90] Moini, M. Simplifying CE-MS operation. 2. Interfacing low-flow separation techniques to mass spectrometry using a porous tip. *Anal. Chem.* **2007**, 79(11), 4241-6.

[91] 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(17), 2554-60.

[92] Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J. Ultrasensitive and fast bottom-up analysis of femtogram amounts of complex proteome digests. *Angew Chem. Int. Ed. Engl.* **2013**, 52(51), 13661-4.

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

[94] Sanz-Nebot, V.; Balaguer, E.; Benavente, F.; Barbosa, J. Comparison of sheathless and sheath-flow electrospray interfaces for the capillary electrophoresis-electrospray ionization-mass spectrometry analysis of peptides. *Electrophoresis* **2005**, 26(7-8), 1457-65.

[95] Lombard-Banek, C.; Reddy, S.; Moody, S. A.; Nemes, P. Label-free Quantification of Proteins in Single Embryonic Cells with Neural Fate in the Cleavage-Stage Frog (Xenopus laevis) Embryo using Capillary Electrophoresis Electrospray Ionization High-Resolution Mass Spectrometry (CE-ESI-HRMS). *Mol. Cell. Proteomics* **2016**,15(8), 2756-68.

[96] Hjerten, S.; High-performance electrophoresis: elimination of electroendosmosis and solute adsorption. *J. Chromatogr. A* **1985**, 347, pp. 191-198

[97] Gao, L.; Liu, S. Cross-linked polyacrylamide coating for capillary isoelectric focusing. *Anal. Chem.* **2004**, 76(24), 7179-86.

[98] Danel, C.; Melnyk, P.; Azaroual, N.; Larchanché, P. E.; Goossens, J. F.; Vaccher, C. Evaluation of three neutral capillary coatings for the determination of analyte-cyclodextrin binding constants by affinity capillary electrophoresis. Application to N,N'-disubstituted piperazine derivatives. *J. Chromatogr. A* **2016**, 1455, 163-171.

[99] Shen, Y.; Smith, R.D. High-resolution capillary isoelectric focusing of proteins using highly hydrophilic-substituted cellulose-coated capillaries. *J. Microcolumn Sep.* **2000**, 12, pp. 135-141

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

[101] Stroink, T.; Paarlberg, E.; Waterval, J. C.; Bult, A.; Underberg, W. J. On-line sample preconcentration in capillary electrophoresis, focused on the determination of proteins and peptides. *Electrophoresis* **2001**, 22(12), 2375-83.

[102] Zhang, Z.; Qu, Y.; Dovichi, N.J. Capillary zone electrophoresis-mass spectrometry for bottom-up proteomics. *TrAC Trends in Anal. Chem.* 2018, 108, 23-37.

[103] Sun, L.; Hebert, A. S.; Yan, X.; Zhao, Y.; Westphall, M. S.; Rush, M. J.; Zhu, G.; Champion, M. M.; Coon, J. J.; Dovichi, N. J. Over 10,000 peptide identifications from the HeLa proteome by using single-shot capillary zone electrophoresis combined with tandem mass spectrometry. *Angew Chem. Int. Ed. Engl.* **2014,** 8, 53(50), 13931-3.

[104] Guo, X.; Fillmore, T. L.; Gao, Y.; Tang, K. Capillary Electrophoresis-Nanoelectrospray Ionization-Selected Reaction Monitoring Mass Spectrometry via a True Sheathless Metal-Coated Emitter Interface for Robust and High-Sensitivity Sample Quantification. *Anal. Chem.* **2016**, 88(8), 4418-25.

[105] 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), 1242-52.

[106] Zhu. G.; Sun, L.; Yan, X.; Dovichi, N. J. Bottom-up proteomics of Escherichia coli using dynamic pH junction preconcentration and capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry. *Anal. Chem.* **2014**, 86(13), 6331-6.

CHAPTER 2. Capillary zone electrophoresis-mass spectrometry with microliter-scale loading capacity, 140 min separation window and high peak capacity for bottom-up proteomics

Part of this chapter was adapted from Analyst 2017, 142(12), 2118-2127 with permission

2.1. Introduction

2D-LC-ESI-MS/MS has been routinely used for large-scale BUP for a decade starting from the invention of MudPIT [1]. Recently, the Coon group reported the complete yeast proteome using only one-hour RPLC-MS/MS analysis [2]. However, for deep proteome profiling of mammalian cells, 2D-LC-MS/MS is still required. Two kinds of 2D-LC systems are widely used. One system typically employs SCX/SAX and RPLC to separate peptides based on their charge and hydrophobicity [1,3–6]. Another system usually employs high-pH RPLC (i.e., pH 10) and low-pH RPLC (i.e., pH 3) to separate peptides based on their hydrophobicity [7–10]. The 2D-LC-MS/MS approaches have achieved great success in the last decade, thus leading to the identification of 10,000 protein groups from a mammalian cell line [8,11], and the generation of the draft human proteome in 2014 [12,13]. However, as is well known, the median protein sequence coverage from large-scale BUP is well below 30%, thus leading to challenges for the identification of protein isoforms, which typically have similar amino acid sequences. Although thousands of LC-MS/MS runs were performed for the draft human proteome work, the median sequence coverage of identified proteins is only about 28% [12]. Accordingly, proteome coverage from bottom-up proteomics is still limited in terms of protein isoform identification. Better

peptide separation is required to improve the protein sequence coverage, thus leading to better characterization of protein isoforms and deeper proteome coverage.

On-line 3D-LC–MS/MS has been established to improve the proteome coverage from bottom-up proteomics [14,15]. The system employed high pH RPLC-SAX–low pH RPLC for peptide separation, and indeed improved the total peak capacity of peptide separation to 1.3×10^4, thus leading to the quantification of 11,352 gene products from mammalian cell lines [15]. Spintip-based 2D-LC fractionation (SCX/SAX-high pH RPLC)–low pH RPLC-MS/MS systems have also demonstrated their power for large-scale and highly sensitive bottom-up proteomics [16,17] However, the 3D-LC systems still only explore the differences among peptides in their hydrophobicity and charge.

We need to consider alternative separation techniques with different separation mechanisms in order to enhance the peptide separation significantly. CZE–MS has been suggested as an attractive alternative platform for large-scale and highly sensitive BUP [18–30]. First, CZE separates analytes based on their size-to-charge ratios, which is complementary with LC [31]. Second, CZE has high efficiency for separation of biomolecules (i.e., peptides and proteins) [24,31,32]. Third, the commercialization of several new CE–MS interfaces recently demonstrates their maturity for robust and highly sensitive CZE–MS experiments [24,33–36]. Fourth, using the improved CE–MS interfaces, CZE–MS has better sensitivity for peptide and protein detection, and it can produce more protein IDs from mass-limited complex proteome samples than typical RPLC–MS [18,19,33]. RPLC–MS has significant sample loss during sample loading and on the column due to the existence of very hydrophilic peptides/proteins, the valve and the large surface area of beads [32,37]. CZE–MS can dramatically attenuate that sample loss due to its

simple design, and can be routinely operated at a low nL/min flow rate for ESI, thus leading to very high ESI efficiency [38]. The Dovichi group recently reported a low zmole peptide detection limit using CZE–MS and 10000 peptide IDs using single-shot CZE–MS/MS from a human cell line proteome digest, which clearly demonstrates the great potential of CZE–MS/MS for large-scale and highly sensitive BUP [24,25]. However, the typical loading capacity of CZE is only on the low nL level and the typical separation window is around 1 hour or narrower. The low loading capacity and narrow separation window limit the further improvement of the number of peptide/protein IDs using CZE–MS/MS from complex proteome samples.

Several methods have been used to improve the loading capacity of CZE–MS for proteomics without significant loss of separation efficiency, e.g., FASS [23,25]. tITP [29,30,39] solid phase micro-extraction [18,40] and dynamic pH junction [41,42]. Dynamic pH junction is a very simple method for improving the loading capacity of CZE, and it was invented by Aebersold's group and Chen's group [43,44]. At least 95% of target molecules injected into the capillary could be captured and concentrated easily with the dynamic pH junction method [45].

Dynamic pH junction based CZE–MS/MS has been used for bottom-up proteomics recently and around 500 nL of the peptide sample was loaded for CZE–MS analysis [42]. In that work, the background electrolyte (BGE) of CZE was 0.1% (v/v) formic acid (pH 3) and the sample was simply dissolved in a buffer with a much higher pH than BGE (ammonium acetate, pH 7) for CZE–MS analysis. A commercialized LPA-coated fused silica capillary from Polymicro Technologies was used for CZE–MS. The results showed that the dynamic pH junction based CZE–MS could approach reasonably good peptide separation when about 500 nL of the sample was loaded for analysis [42]. However, the results also indicated that when

the volume of the loaded sample increased from 20 nL to around 500 nL, the separation window of the system became significantly narrower [42]. The widest separation window of dynamic pH junction based CZE–MS for peptide samples is around 80 min [46]. The widest separation window of CZE–MS for peptide samples is around 90 min with 100 nL sample loading volume based on a field enhanced sample stacking method [25]. Unfortunately, the separation efficiency of CZE–MS attenuated when larger than 100 nL of the sample was loaded based on the FASS method [42]. A wide separation window and a large loading capacity are both imperative for CZE–MS based large-scale proteomics. A high peak capacity is also crucial. The highest peak capacity for peptide separation using CZE was around 300 [25,39]. CZE–MS systems with a large loading capacity, a wider separation window and a higher peak capacity are required for large-scale proteomics.

In this work, we reported one automated CZE–MS system with a microliter-scale loading capacity, a 140 min separation window and a high peak capacity (~380) for complex proteome digest analysis. It is the first time that CZE–MS approaches both the microliter-scale loading capacity and over 2-hour separation window for analysis of complex samples. The CZE–MS system employed the systematically optimized dynamic pH junction sample stacking and in-house made LPA-coated separation capillary. The results represent the widest separation window and the highest peak capacity of CZE–MS for peptide separation. The CZE–MS system truly opens the door of CZE–MS based large-scale BUP.

2.2. Experimental

2.2.1. Materials and reagents

All reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. LC/MS grade water, formic acid (FA), methanol, acetonitrile (ACN), HPLC grade acetic acid (AA) and hydrofluoric acid (HF) were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide was purchased from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ).

Mammalian Cell-PE LB™ buffer for cell lysis was purchased from G-Biosciences (St. Louis, MO). Complete, mini protease inhibitor cocktail (provided in EASYpacks) was purchased from Roche (Indianapolis, IN).

2.2.2. LPA coating for separation capillary of CZE

Bare fused silica capillary (50 μm i.d./360 μm o.d.) was flushed successively with 1 M hydrochloric acid, water, 1 M sodium hydroxide, water, and methanol. Then the capillary was flushed with nitrogen overnight at room temperature. After that, the capillary was filled with 50% (v/v) 3-(trimethoxysilyl) propyl methacrylate in methanol and was kept at room temperature for at least 24 hours with both ends sealed by silica rubber. (Tips. Based on our experience, the capillary with 3-(trimethoxysilyl) propyl methacrylate inside can be kept at room temperature for up to one week before next step. The silica rubber will block the ends of the capillary, so before next step at least 5 mm length of capillary should be removed from both ends of the capillary.) The capillary was then flushed with methanol to remove the unreacted reagents and dried under nitrogen. The capillary now can be stored at room temperature with both ends sealed with silica rubber before further steps.

40 mg of acrylamide was dissolved in 1 mL of water. Then 500 µL of the acrylamide solution was used for following steps. 3.5 µL of 5% (w/v) ammonium persulfate (APS) in water was added to the 500-µL acrylamide solution. After mixed via vortex, the solution was degassed using nitrogen to remove the oxygen in the solution for 5 minutes. The solution finally was introduced into the pretreated capillary by vacuum. The filled capillary was sealed with silica rubber at both ends and incubated in the water bath at 50 ˚C for 35 min. The capillary was then flushed with water to remove excess reagents and was stored at room temperature before use. (Tips. The reaction in the capillary is initiated by APS at high temperatures. The reaction time can be varied depending on the temperature in the lab. Therefore, the volume of APS and reaction time need to be slightly adjusted. The degassing step is important and this step can be longer (i.e., 10 min). When you flush the capillary with water after the reaction at 50 ˚C, you should be able to see “agarose gel” like substance being pushed out of the capillary at the beginning. The quantity of the “agarose gel” like substance can be small depending on the total volume of the capillary. If you see the “agarose gel” like substance, it means the quality of the coating is good. If the reaction time is too short or APS concentration is too low or the degassing step is not sufficient, you will not see the “agarose gel” like substance. We suggest using a short capillary to test the conditions at the beginning.)

2.2.3. Sample Preparation

Four standard proteins including bovine serum albumin, cytochrome c (from bovine), myoglobin (from equine) and beta casein (from bovine) were dissolved in 8 M urea and 100 mM ammonium bicarbonate (NH_4HCO_3), pH 8.0, and the solutions

were kept at 37 ˚C for 30 min for protein denaturation, followed by protein reduction with dithiothreitol (DTT) at 37 ˚C for 30 min and alkylation with iodoacetamide (IAA) at room temperature for 20 min. After diluting the samples with 100 mM NH_4HCO_3 (pH 8.0) to make the urea concentration lower than 2 M, tryptic digestion was performed at 37 ˚C for overnight with trypsin/protein ratio as 1/30 (w/w). After digestion, the samples were acidified with formic acid to terminate the reactions and desalted with C18 SPE columns (Waters, Milford, MA), followed by lyophilization with a vacuum concentrator (Thermo Fisher Scientific). Finally, we prepared a standard-protein digest sample via mixing the peptides from the four standard proteins. The standard-protein digest sample contains 1 mg/mL of BSA, 0.07 mg/mL of myoglobin, 0.03 mg/mL of beta-casein and 0.006 mg/mL of cytochrome c. This sample was further diluted 10-times with different buffers to get total peptide concentration around 0.1 mg/mL for CZE-MS experiments.

Two 21-week-old female OT-1 mice were sacrificed for collection of brain and liver. The mouse brain and liver samples were kindly provided by Professor Xuefei Huang's group at Department of Chemistry, Michigan State University. The whole protocol related to the mouse samples was performed following guidelines defined by the Institutional Animal Care and Use Committee of Michigan State University. The mouse brain and liver samples taken from the sacrificed mice were stored at -20 ˚C before use. The mouse brain and liver samples were further prepared with the same protocol described below. The tissue was first cut into small pieces and washed with phosphate-buffered saline (PBS) to remove the blood. Then, the sample was suspended in 9 mL of lysis buffer containing mammalian cell-PE LBTM buffer and complete protease inhibitor, followed by homogenization with a Homogenizer 150 (Fisher Scientific, Pittsburgh, PA) on ice and sonication with a

Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 10 minutes. The lysates were then aliquoted equally into 1.7 mL Eppendorf tubes, followed by centrifugation at 10,000 g for 5 min. The supernatants were collected and a small portion of the sample was used to measure the protein concentration with bicinchoninic acid (BCA) assay. The supernatants were then subjected to acetone precipitation to purify the proteins. Briefly, 1 volume of sample was mixed with 4 volumes of cold acetone. Then the mixture was kept at -20 °C overnight, followed by centrifugation at 10,000 g for 5 min. After removed the supernatant, we added cold acetone to the Eppendorf tube again to simply wash the protein pellet, followed by centrifugation. After centrifugation, the supernatant was discarded and the protein pellet was air dried in the chemical hood for several minutes. Finally, the protein pellets were stored at -20 ˚C before use.

The protein pellet from the mouse brain or liver was dissolved in 8 M urea and 100 mM NH_4HCO_3 (pH 8.0) buffer via vortex and sonication to a final protein concentration of approximately 4 mg/mL. The proteins were denatured at 37 °C for 30 min, followed by protein reduction with DTT at 37 °C for 30 min, and protein alkylation with IAA in dark at room temperature for 20 min. The DTT solution was added into the sample again to react with the left IAA for 10 min at room temperature. The sample was then diluted with 100 mM NH_4HCO_3 (pH 8.0) to reduce the urea concentration to 2 M. After that, proteins were digested into peptides with TPCK-treated trypsin at 37 °C overnight with a trypsin/protein ratio of 1/30 (w/w). Then, the protein digests were acidified with formic acid to terminate the reaction. The protein digests were desalted with C18 SPE columns (Waters, Milford, MA), followed by lyophilization with a vacuum concentrator (Thermo Fisher

Scientific). The protein digests were dissolved in different buffers for CZE-MS experiments.

2.2.4. RPLC fractionation of mouse brain proteome digests

500 µg and 50 µg of mouse brain proteome digests were dissolved in 150 µL of 0.1% (v/v) FA for RPLC fractionation. An Agilent Infinity II HPLC system was used. A C18 RP column (Zorbax 300Extend-C18, 2.1 mm i.d. × 150 mm length，3.5 µm particles, Agilent Technologies) was used for peptide separation. Buffer A (water, 0.1% FA) and buffer B (ACN, 0.1%FA) were used as mobile phases to generate a gradient for separation. The flow rate was 0.3 mL/min. The peptide samples (500 µg and 50 µg of mouse brain digests) were loaded onto the RPLC column for 4 min at 2% B. Then the peptides were separated by gradient elution: 2 min from 2% B to 6% B, 70 min from 6% B to 40% B, and 1 min from 40% B to 80% B. The mobile phase was kept at 80% B for 4 min, followed by column equilibration with 2% B for 10 min.

We collected fractions from 11 min to 78 min. The eluate from 11 min to 14 min were collected as one fraction, from 72 min to 78 min was collected as one fraction. From 14 min to 72 min, fractions were collected one fraction/min. In total 60 fractions were collected from each sample.

For the 500 µg of mouse brain digest sample, fraction number “N” and fraction number “N+30” were combined, thus leading to totally 30 fractions. For the 50 µg of protein digest sample, fraction number “N”, fraction number “N+15”, fraction number “N+30” and fraction number “N+45” were combined, thus leading to totally 15 fractions. The fractions were lyophilized and stored at -20 ˚C before use.

Each RPLC fraction from the 500 µg of protein digest sample was dissolved in 8 µL of 10 mM NH_4HCO_3 (pH 8.0) for CZE-ESI-MS/MS analysis. Each RPLC fraction from the 50 µg of protein digests was dissolved in 4 µL of 10 mM NH_4HCO_3 (pH 8.0) for CZE-ESI-MS/MS analysis.

2.2.5. CZE–ESI–MS/MS

An ECE-001 capillary electrophoresis autosampler (CMP Scientific, Brooklyn, NY) was used for CZE separation. An LTQ-XL mass spectrometer (Thermo Fisher Scientific) was used for detection. A commercialized electro-kinetically pumped sheath flow interface (CMP Scientific, Brooklyn, NY) was employed for coupling CZE to MS [24,33,48].

The LPA-coated capillary (50 µm i.d., 360 µm o.d.) made in house was etched with HF based on ref. 24 to reduce the outer diameter of one end of the capillary to around 70 µm with the total length of the etched part around 3 mm. The total length of the capillary for CZE–MS was 96 cm. (Caution: use appropriate safety procedures while handling HF solutions.) The power supply integrated in the autosampler was used for CZE separation and another power supply from CMP Scientific (Brooklyn, NY) was used for electrospray. The electrospray emitter was a borosilicate glass capillary (1.0 mm o.d., 0.75 mm i.d., and 10 cm length) pulled with a Sutter instrument P-1000 flaming/brown micropipette puller. The opening size of the electrospray emitter is around 30 µm. The BGE of CZE was 5% (v/v) AA and the sheath buffer was 0.2% (v/v) FA containing 10% (v/v) methanol.

26 kV or 30 kV was applied at the injection end for CZE separation and around 2 kV was applied in the sheath buffer vial for electrospray. For all the standard-protein digest data, we applied 26 kV at the injection end for separation. The analysis time

ranged from 125 min to 265 min. For all the mouse brain and liver proteome digests, 30 kV was applied at the injection end for separation. The analysis time ranged from 110 min to 210 min. For the RPLC fractionated samples, 30 kV was applied at the injection end for separation. The separation time ranged from 120 min to 150 min. For all the CZE–MS runs, we flushed the capillary with BGE for 10 to 20 min at 5 psi pressure at the end of the separation. Sample injection was performed via applying 5-psi pressure for different periods in order to inject different volumes of samples into the separation capillary. The sample injection volume was calculated based on Poiseuille's law.

The distance between the spray emitter orifice and the mass spectrometer entrance was ~2 mm. The parameters of the LTQ-XL mass spectrometer are listed below. The ion transfer tube temperature was 200 °C. The top 10 data dependent acquisition (DDA) method was used for data acquisition. The full MS scan was acquired with the automatic gain control (AGC) target value as 30 000 and scan range 300 m/z to 1500 m/z. The ten most intense ions were sequentially isolated in the ion trap with 2 m/z isolation window, followed by fragmentation with normalized collision energy as 35%. The AGC target value for MS/MS is 10 000. The maximum ion injection time for MS and MS/MS scans was 50 ms and 100 ms, respectively. Dynamic exclusion was enabled with the following settings: repeat count as 1, repeat duration as 15 s and exclusion duration as 20 s. The minimum ion signal in MS spectra required for triggering MS/MS is 5000 counts.

2.2.6. Data analysis

The RAW files were analyzed by Proteome Discoverer 1.4 software (Thermo Fisher Scientific) with Sequest HT database search engine. The mouse proteome,

bovine proteome and equine proteome databases (ID: UP000000589, UP000009136, UP000002281) downloaded from UniProt (http://www.uniprot.org/) were used for database search. The database search was also performed for the reversed databases to evaluate the false discovery rates (FDRs). For database search, the MS/MS spectra were filtered firstly to remove the background peaks with the parameters as top 6 peaks in every 100 Da window. Fully specific tryptic digestion was chosen. The mass tolerance of parent ions and fragment ions was 2 Da and 1 Da, respectively. The dynamic modification was oxidation (M), and the static modification was carbamidomethyl (C). The Percolator tool integrated in the Proteome Discoverer 1.4 software was used to validate the database search results based on the q-value. The IDs were filtered with peptide confidence value as high to obtain FDR less than 1% on the peptide level. Protein grouping was enabled, and a strict maximum parsimony principle was applied. Therefore, if multiple proteins were identified from same peptides, these proteins were grouped into one protein group, and each protein group has at least one unique peptide.

2.3. Results and discussion

2.3.1. Optimization of the dynamic pH junction based CZE-MS system

CZE-MS systems with large loading capacity, wider separation window and higher peak capacity are required for CZE-MS based large-scale proteomics. Because the dynamic pH junction method has the great potential to dramatically improve the loading capacity of CZE without significant loss of separation efficiency [42], we should be able to optimize the dynamic pH junction based CZE-MS system

to approach large loading capacity, significantly wider separation window and higher peak capacity.

In typical dynamic pH junction based CZE experiments, the sample buffer and BGE of CZE have significantly different pH values [42-44]. For instance, the sample is dissolved in a buffer with pH as 10 and the pH of BGE is 3. The underlying mechanism of dynamic pH junction-based sample stacking is complicated, and its concentration performance may be provided by more than one mechanism, e.g., pH boundary and isotachophoresis [45,49-53] Here we systematically evaluated the effect of sample buffer pH and salt concentration as well as the sample injection volume on the CZE-MS performance for complex peptide mixtures. We chose 5% (v/v) AA (pH 2.4) as the BGE based on the results from initial evaluations of different BGEs (0.1%-0.5% (v/v) FA and 1%-10% (v/v) AA). 5% (v/v) AA as BGE produced stable CZE-MS runs and the widest peptide separation window when 500 nL of the standard protein digest was loaded for analysis. We used the in-house made LPA-coated capillary for peptide separation to reduce the EOF in the capillary.

We firstly optimized the sample injection volume with the standard-protein digest sample, **Figure 2.1**. When the sample injection volume increases from 40 nL to 500 nL, the peptide intensity reasonably increases, suggesting that peptides can be well concentrated via the dynamic pH junction method with even half-a-microliter sample loading volume. Interestingly, when the sample loading volume increases, the migration speed of analytes in the capillary decrease and the separation window of the system widens. The peptides start to migrate out of the capillary at around 20 min, 35 min and 55 min for 40 nL, 240 nL and 500 nL injection volume, respectively, **Figure 2.1**. Because of the slower migration speed of peptides for 500 nL loading volume, peptides spread out in a wider migration time window. The results clearly

demonstrate that dynamic pH junction based CZE-MS system can approach half-a-microliter loading capacity and wider separation window.

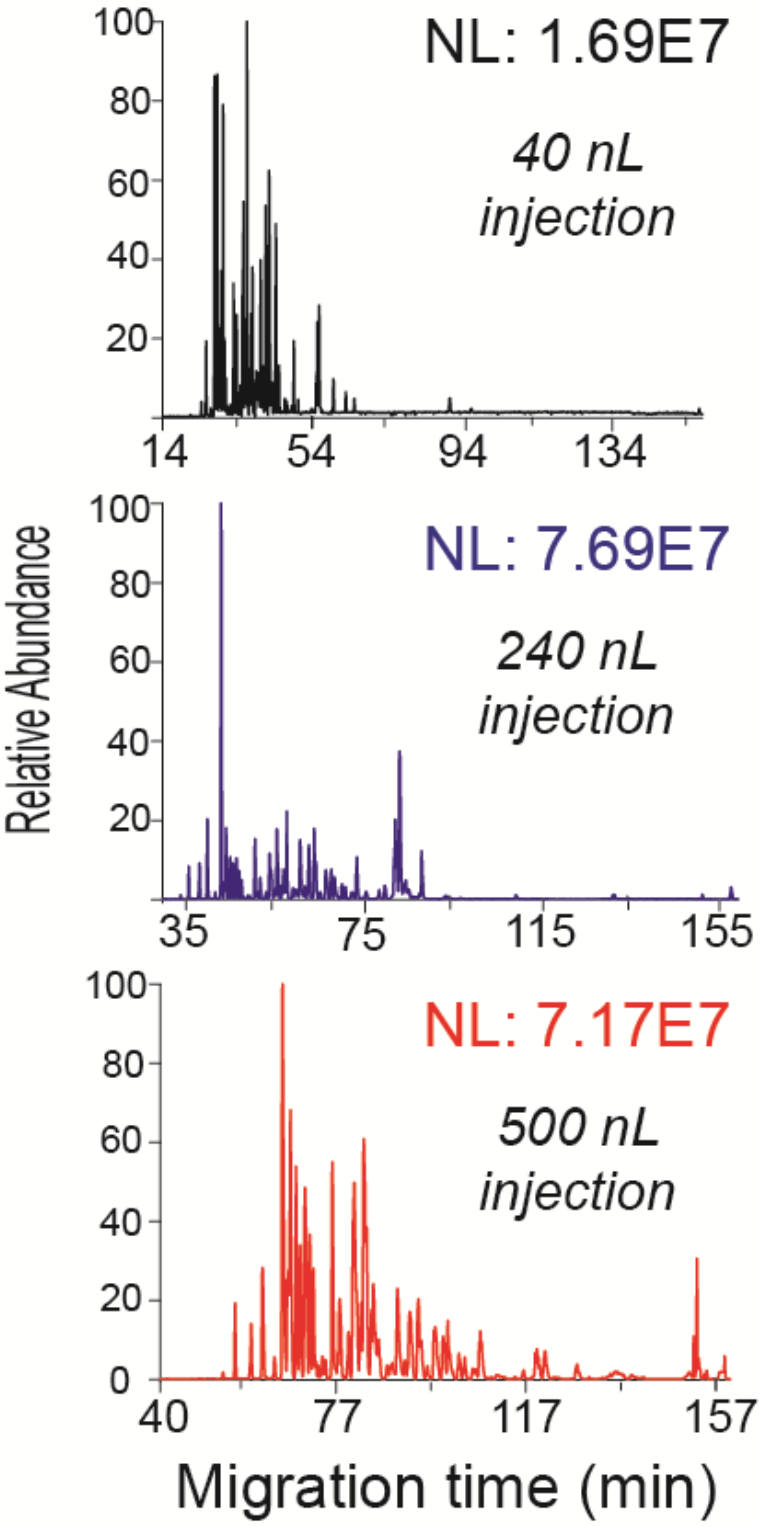

Figure 2.1. Electropherograms of the standard-protein digest sample (0.1 mg/mL in 10 mM NH_4HCO_3, pH 8.0) after CZE-MS analysis with three different sample injection volumes. Top: 40 nL injection; Middle: 240 nL injection; Bottom: 500 nL injection.

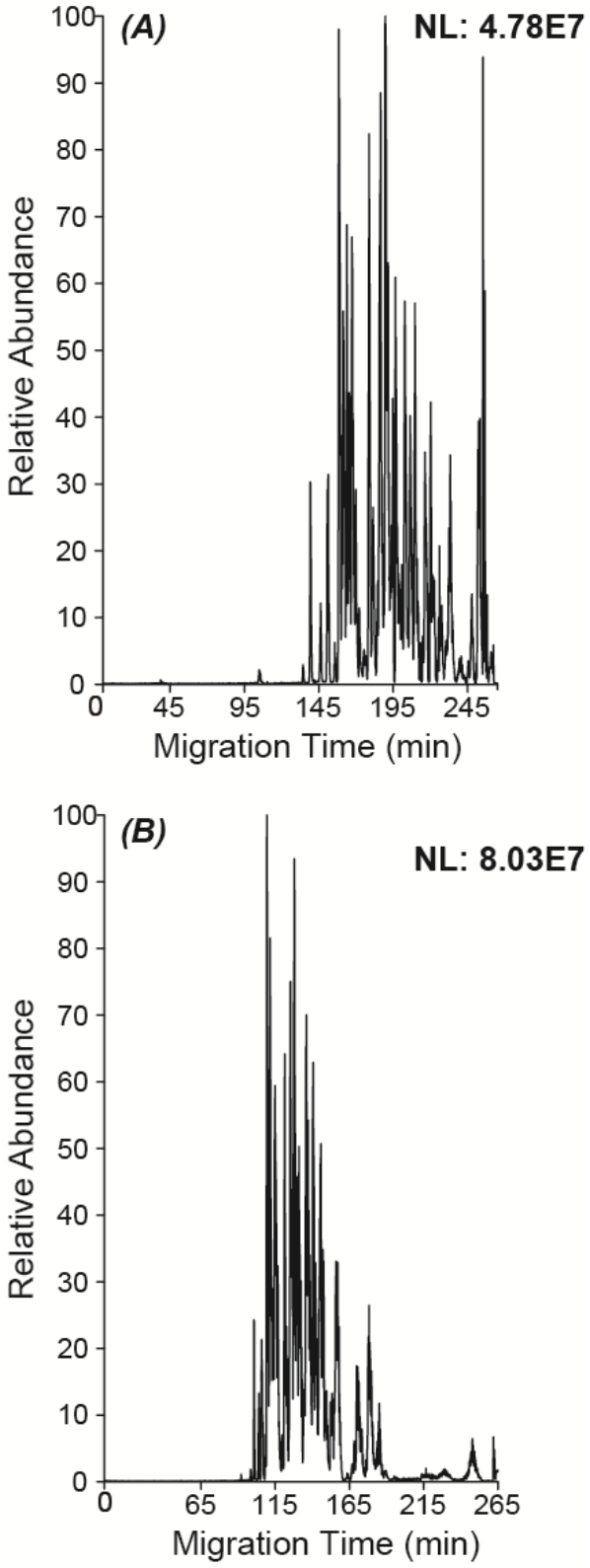

Figure 2.2. Electropherograms of the standard-protein digest sample (0.1 mg/mL in 10 mM NH_4HCO_3, pH 8.0) after CZE-MS analysis with two different sample injection volumes. (A): 1 µL injection; (B) 1.5 µL injection.

We also tried to inject 1 µL (50% of the total capillary volume) and 1.5 µL (75% of the total capillary volume) of the sample for CZE-MS. Surprisingly, we still observed good peptide separation even 50%-75% of the capillary was filled with the sample, **Figure 2.2**. When the sample injection volume increases from 500 nL to 1 µL, the migration speed of peptides in the capillary continue to slow down, and the separation window becomes wider. When the sample injection volume increases from 1 µL to 1.5 µL, the peptides migrate faster in the capillary and the separation window narrows. The results suggest that the CZE-MS system can approach 1 µL of sample loading capacity (50% of the total capillary volume) without effect on the peptide separation. The results also suggest that change of injection volume can be used as one way to modulate the migration speed of peptides in the separation capillary and the separation window. It is worth to note that the numbers of unique peptides matched to the four standard proteins from different sample loading volumes (40 nL to 1.5 µL) are comparable, which is most likely due to the low complexity of the sample.

We chose 500 nL sample injection volume for all the following experiments in order to balance the loading capacity, separation window and the migration speed of peptides in the capillary. Although there is still around 50 min dead time for the 500 nL sample injection data, we can easily reduce the dead time via applying higher separation voltage across the capillary. Interestingly, it has been demonstrated in the literature that when the sample injection volume of CZE increased from 20 nL to around 500 nL, the separation window of dynamic pH junction based CZE-MS narrowed significantly [42]. We observed different phenomenon in this work. One potential reason for this difference is the higher quality of the LPA coating on the inner wall of the separation capillary in this work (in-house made) compared with

reference [42] (from Polymicro Technologies), thus leading to much lower EOF in the capillary. Based on our previous experience, in order to get good intact protein separation, the LPA coated capillary (50 µm i.d.) from Polymicro Technologies need to be flushed with the protein sample first to reduce the dead adsorption of proteins on the inner wall of the separation capillary before protein separation. The in-house made LPA-coated capillary can be directly used for intact protein separation.

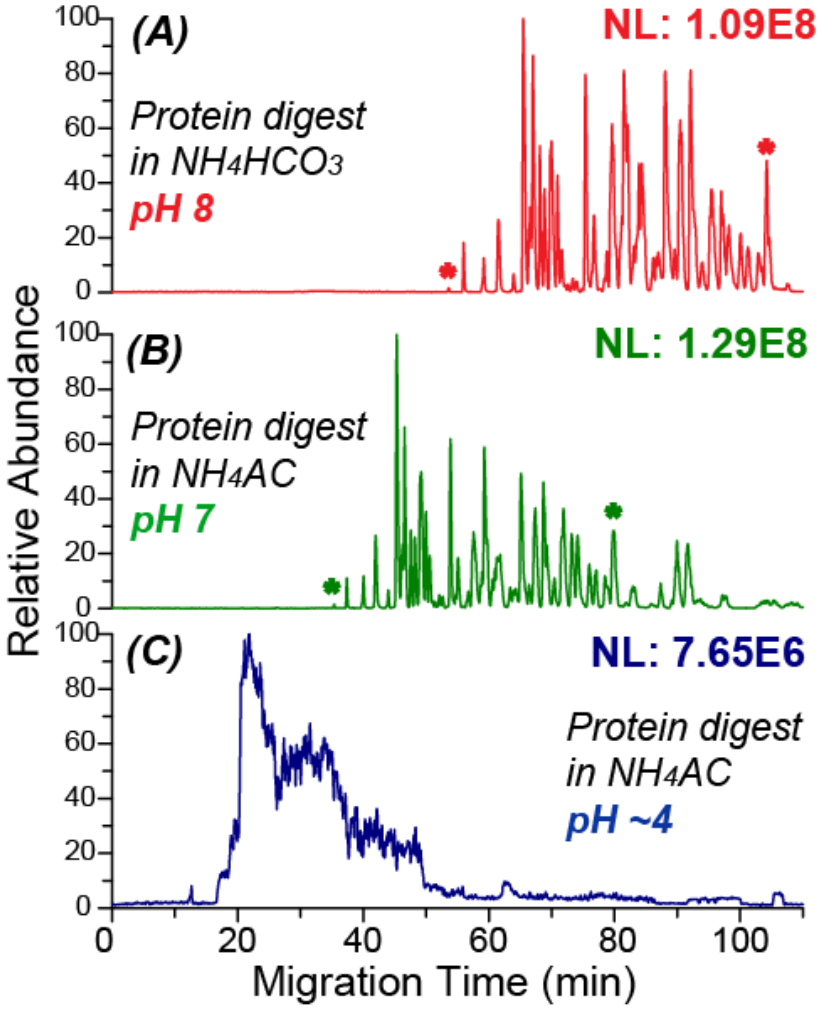

Figure 2.3. Electropherograms of the standard-protein digest sample (0.1 mg/mL) in 10 mM NH_4HCO_3, pH 8.0 (A), in 10 mM ammonium acetate (NH_4AC), pH ~7 (B) and 10 mM NH4AC, pH ~4 (C) after CZE-MS analysis with 500 nL sample injection volume. Two peptides were marked (*) in electropherograms (A) and (B) to show the different distance between those two peptides.

Next, we optimized the pH of the sample buffer with the standard-protein digest sample, **Figure 2.3**. The results clearly show that sample buffer with higher pH (pH 7

and 8) leads to better peptide separation when 500 nL of sample was injected for CZE-MS. The results also show that sample buffer with higher pH leads to lower migration speed of peptides in the capillary and wider separation window, which are clearly demonstrated by the migration time of the two marked peptides and the migration time difference between them for different sample buffers (~50 min for pH 8 sample buffer vs. ~45 min for pH 7 sample buffer). The total numbers of unique peptides matched to the four standard proteins from pH 7 and 8 sample buffers are similar and are about 34% higher than that from pH 4 sample buffer. We did not try sample buffers with pH higher than 8, because the LPA coating of separation capillaries is not stable in high basic conditions. Accordingly, we chose pH 8 (NH_4HCO_3) as the optimum pH of sample buffer.

Finally, we optimized the salt concentration of the sample buffer with both the standard-protein digest sample and the mouse brain proteome digest sample, **Figure 2.4,** and **Figure 2.5.** Higher salt concentration in the sample buffer leads to higher peptide intensity and narrower peptide separation window. The narrower separation window is demonstrated by the shorter distance between the two marked peptides in the electropherograms (38 min, 34 min and 27 min for 5 mM, 10 mM, and 20 mM NH_4HCO_3 in **Figure 2.5**, respectively). The numbers of peptide IDs from the mouse brain proteome digests in 5-20 mM NH_4HCO_3 (pH ~8.0) are comparable. In order to balance the peptide intensity and separation window, we chose the 10 mM NH_4HCO_3 (pH ~8.0) as the optimum sample buffer. We also need to note that 30 kV was applied for CZE separation of the mouse brain proteome digest (**Figure 2.5**.) and 26 kV was applied for the standard-protein digest (**Figure 2.4**), thus leading to faster peptide migration in the electropherograms in **Figure 2.5**.

The results mentioned above suggest that the pH boundary and isotachophoresis contribute to the analyte focusing during dynamic pH junction process. Higher sample buffer pH produces better peptide separation. Higher salt concentration in sample buffer generates higher peptide intensity due to higher concentration of leading electrolyte (NH_4^+) for isotachophoresis. We also observed that at the beginning of the dynamic pH junction based CZE-MS, the current of CZE was quickly and dramatically decreased (i.e., from 8 µA to 2 µA when 500 nL of sample in 10 mM NH_4HCO_3 was loaded for analysis). Interestingly, larger sample injection volume led to more obvious current drop. The results suggest that low conductivity zone is produced in the capillary during dynamic pH junction based sample stacking. Accordingly, the conventional field enhanced sample stacking may also contribute to the analyte stacking during dynamic pH junction process. The results here further indicate that the mechanism of dynamic pH junction based sample stacking is complicated.

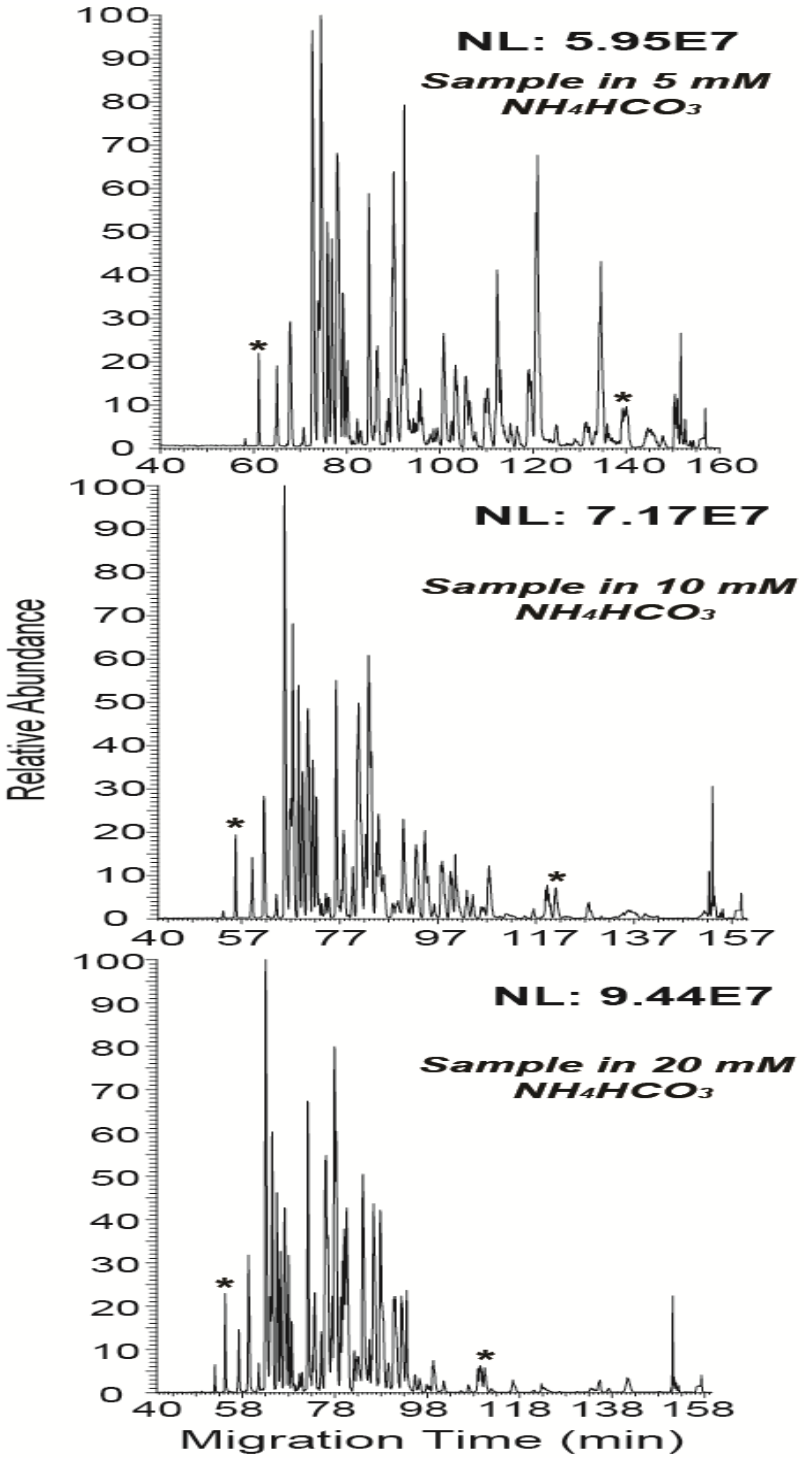

Figure 2.4. Electropherograms of the standard-protein digest sample (0.1 mg/mL) in 5 mM NH_4HCO_3, pH ~8.0 (top), 10 mM NH_4HCO_3, pH ~8.0 (middle), and 20 mM NH_4HCO_3, pH ~8.0 (bottom) after CZE-MS analysis with 500 nL sample injection volume. Two peptides were marked (*) in the electropherograms to show the different distance between those two peptides.

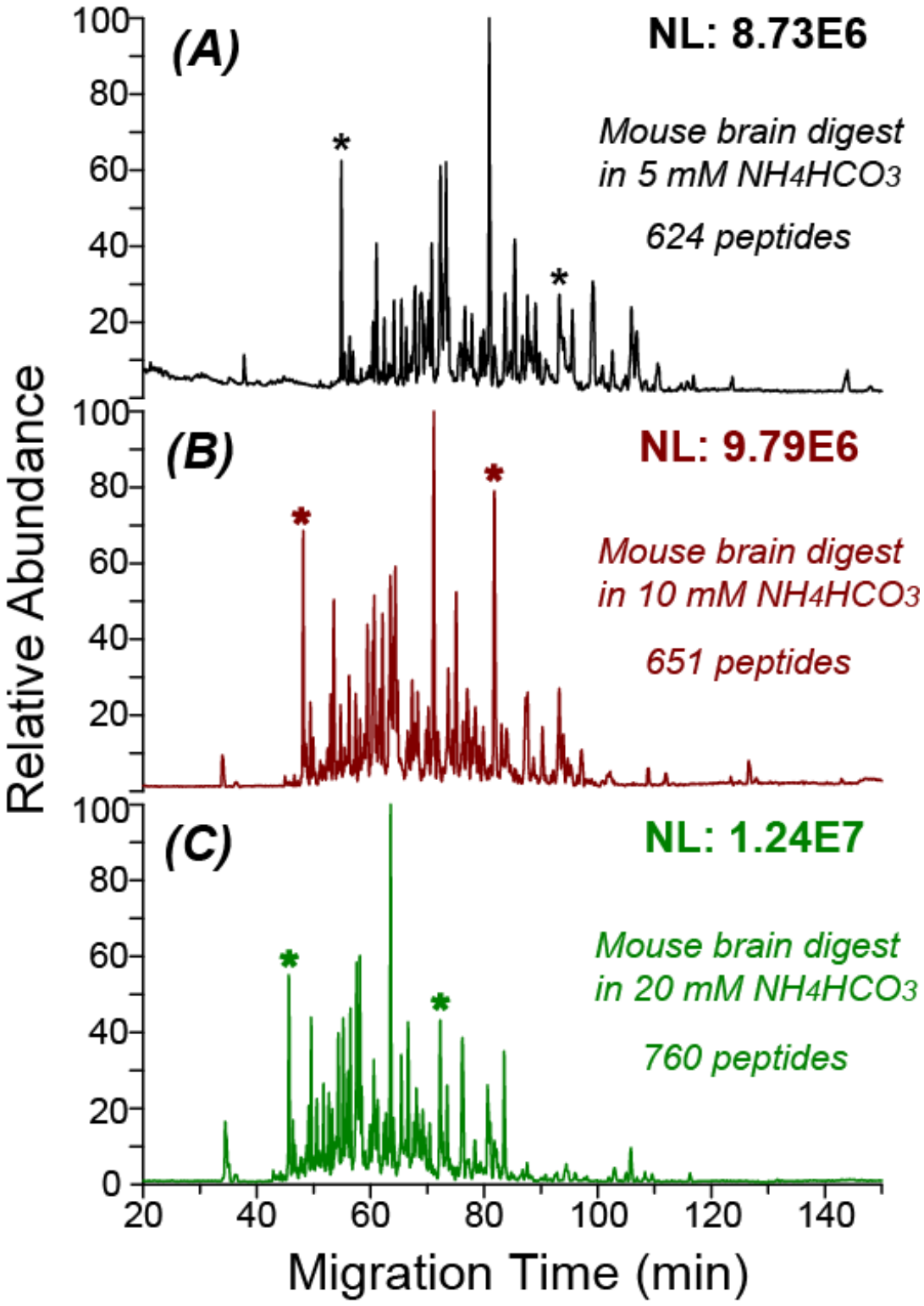

Figure 2.5. Electropherograms of the mouse brain proteome digests (0.4 mg/mL) in 5 mM NH_4HCO_3, pH ~8.0 (A), 10 mM NH_4HCO_3, pH ~8.0 (B), and 20 mM NH_4HCO_3, pH ~8.0 (C) after CZE-MS/MS analysis with 500 nL sample injection volume. Two peptides were marked (*) in the electropherograms to show the different distance between those two peptides.

The results also suggest that the characteristics (e.g., separation window, stacking performance, migration speed of analytes in the capillary) of dynamic pH junction based CZE-MS can be manipulated via simple adjustments of the sample buffer and/or sample injection volume.

Based on the optimized results above, we employed the dynamic pH junction based CZE-MS with 500 nL sample loading volume and 10 mM NH_4HCO_3 (pH 8) sample buffer for the following experiments.

2.3.2. Reproducibility and robustness of the dynamic pH junction based CZE-MS system

We evaluated the reproducibility of the system with both the standard-protein digest sample and the mouse brain proteome digest sample, **Figure 2.6**. and **Figure 2.7**. CZE-MS system with 500 nL injection volume produced reproducible separation and detection of both the standard-protein digest and complex proteome digest samples in terms of the separation profile, peak intensity, migration time and the number of peptide IDs. The relative standard deviation (RSD) of number of peptide IDs from the mouse brain proteome digest sample with the CZE-MS/MS system in quintuplicates is about 7%. We also calculated RSDs of the migration time and intensity of peptides based on five randomly selected peptides from the standard-protein digest data. The RSDs of migration time and intensity of peptides are around 5% and less than 16%, respectively. We continuously used the dynamic pH junction based CZE-MS system for analysis of different peptide samples for about one week, suggesting the good robustness of the system.

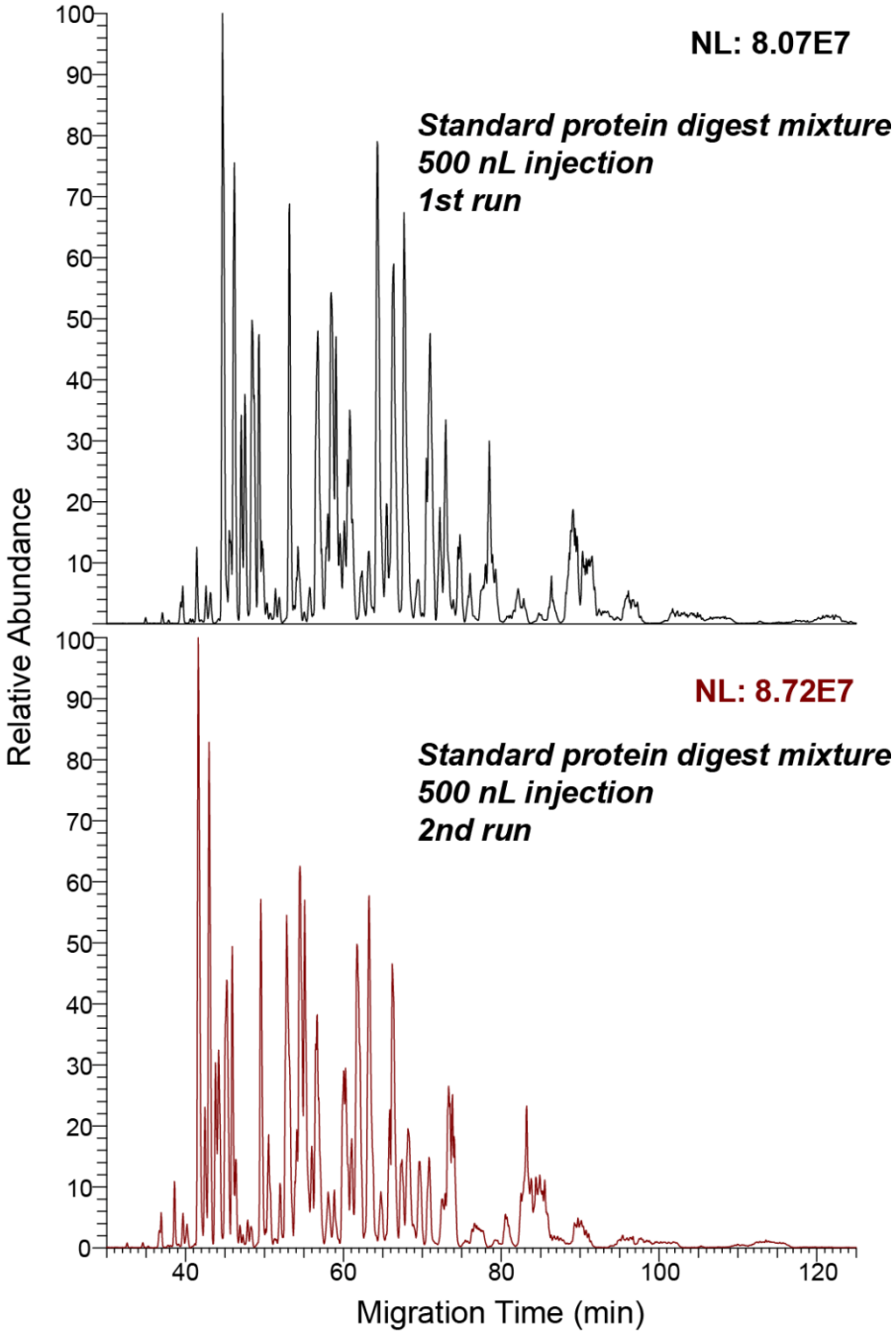

Figure 2.6. Electropherograms of the standard-protein digest sample (0.1 mg/mL) in 10 mM NH_4HCO_3 (pH ~8.0) after CZE-MS analysis in duplicates with 500 nL sample injection volume per run.

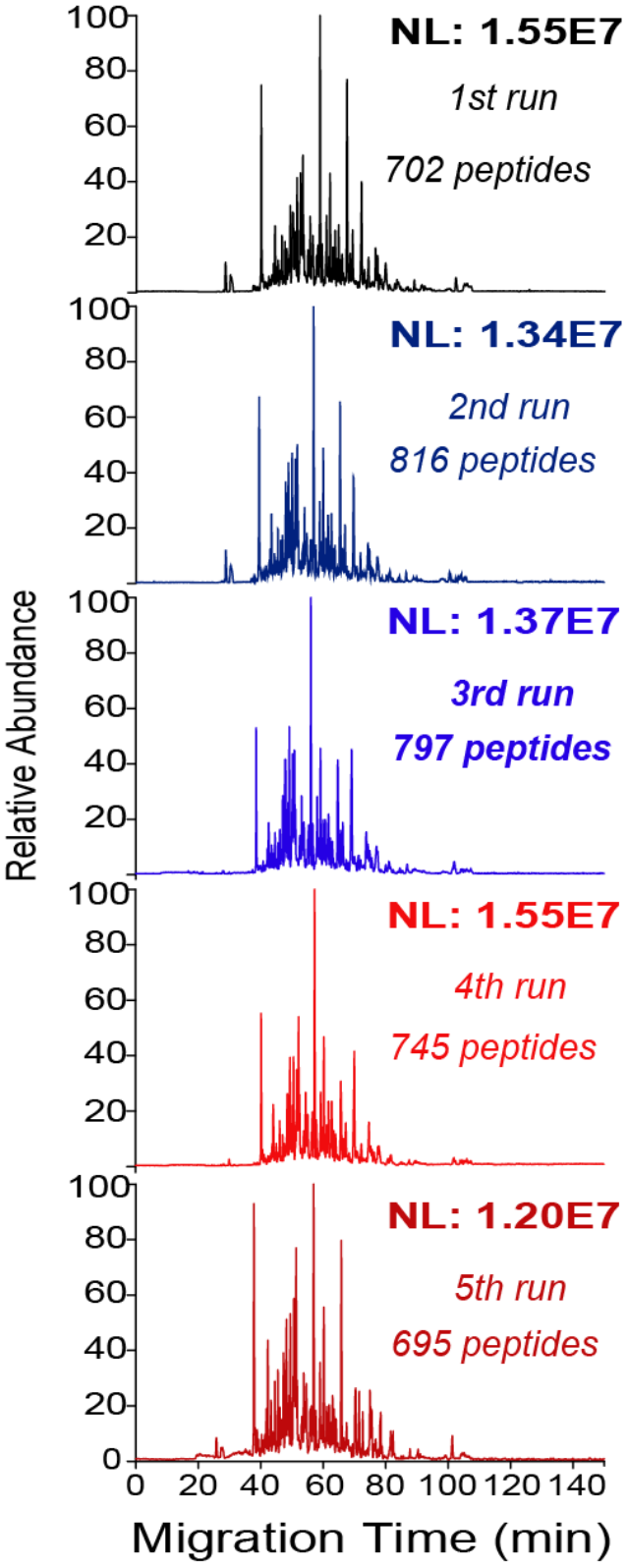

Figure 2.7. Electropherograms of the mouse brain proteome digests (0.4 mg/mL) in 10 mM NH_4HCO_3 (pH 8) after CZE-MS/MS analysis in quintuplicates with 500 nL sample injection volume per run.

2.3.3. 140-min peptide separation window with the dynamic pH junction based CZE-MS system

We further applied the optimized dynamic pH junction based CZE-MS/MS system for analysis of a mouse liver proteome digest sample, **Figure 2.8**. We approached nearly 140-min separation window and the separation window is much wider than that reported in the literature (140 min vs. 90 min) [25,39] We calculated the peak capacity of the system based on the peak width at 50% height, and we observed significantly higher peak capacity than that reported in the literature (380 vs. 300) [25,39]. The results here represent the widest separation window and the highest peak capacity for peptide separation using CZE-MS. It is the first time that CZE-MS approaches both microliter-scale loading capacity and over 2-hour separation window for analysis of complex samples. 497 protein groups and 1,400 peptides were identified with this single-shot CZE-MS/MS analysis on an LTQ-XL mass spectrometer.

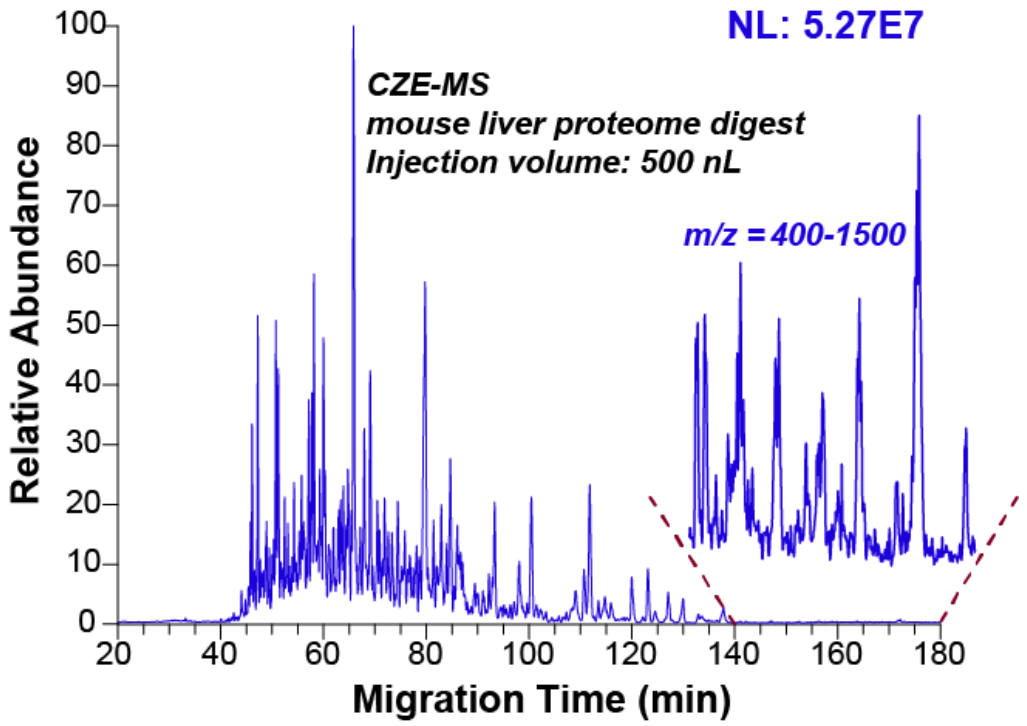

Figure 2.8. Electropherogram of the mouse liver proteome digest (1 mg/mL) in 10 mM NH_4HCO_3 (pH 8) after CZE-MS/MS analysis with 500 nL sample injection volume.

In the literature, single-shot CZE-MS/MS has approached 800-2,000 protein group IDs and over 10,000 peptide IDs from complex proteome samples [25,28,54]. Orbitrap Fusion or Q-Exactive HF mass spectrometer with high mass resolution and high scan speed (up to 20 Hz) was employed in those works. In this work, an LTQ-XL mass spectrometer with dramatically lower mass resolution and scan speed (~3 Hz) was used, thus leading to lower number of peptide and protein IDs compared with those works. Recently, ~500 peptides and 200 protein groups were identified from a complex proteome sample with single-shot CZE-MS/MS on a LTQ-XL mass spectrometer [42]. Our data in this work is about 2-3 times better in terms of the number of peptide and protein group IDs, which is most likely due to the significantly better peptide separation.

2.3.4. Quantitative performance of the dynamic pH junction based CZE-MS system

We analyzed the mouse brain proteome digests with five different concentrations ranging from 0.1 mg/mL to 1.6 mg/mL via the dynamic pH junction based CZE-MS/MS. The sample loading volume of CZE-MS/MS is 500 nL per run. The loaded mass of mouse brain proteome digests ranged from 50 ng to 800 ng. We observed good linear correlations between the mass of loaded mouse brain proteome digests and peptide intensity based on five chosen peptides, **Figure 2.9A**. Those five peptides have different isoelectric points (4.1-9.8), different length (6-26 amino acids) and dramatically different intensity. The result suggests that the dynamic pH junction based CZE-MS system is quantitative and has great potential for label free based quantitative bottom-up proteomics.

We also evaluated the relationship between the number of peptide and protein IDs and the mass of loaded mouse brain proteome digest, **Figure 2.9B**. The number of protein and peptide IDs increase reasonably with the increase of mass of loaded mouse brain proteome digest, further suggesting the system is quantitative. It is worth to note that when the mass of loaded complex proteome digests increases from 400 ng to 800 ng, the number of peptide and protein IDs only increases very slightly, suggesting that 400 ng is very close to the optimum loading mass of complex proteome digests for the dynamic pH junction based CZE-MS/MS system.

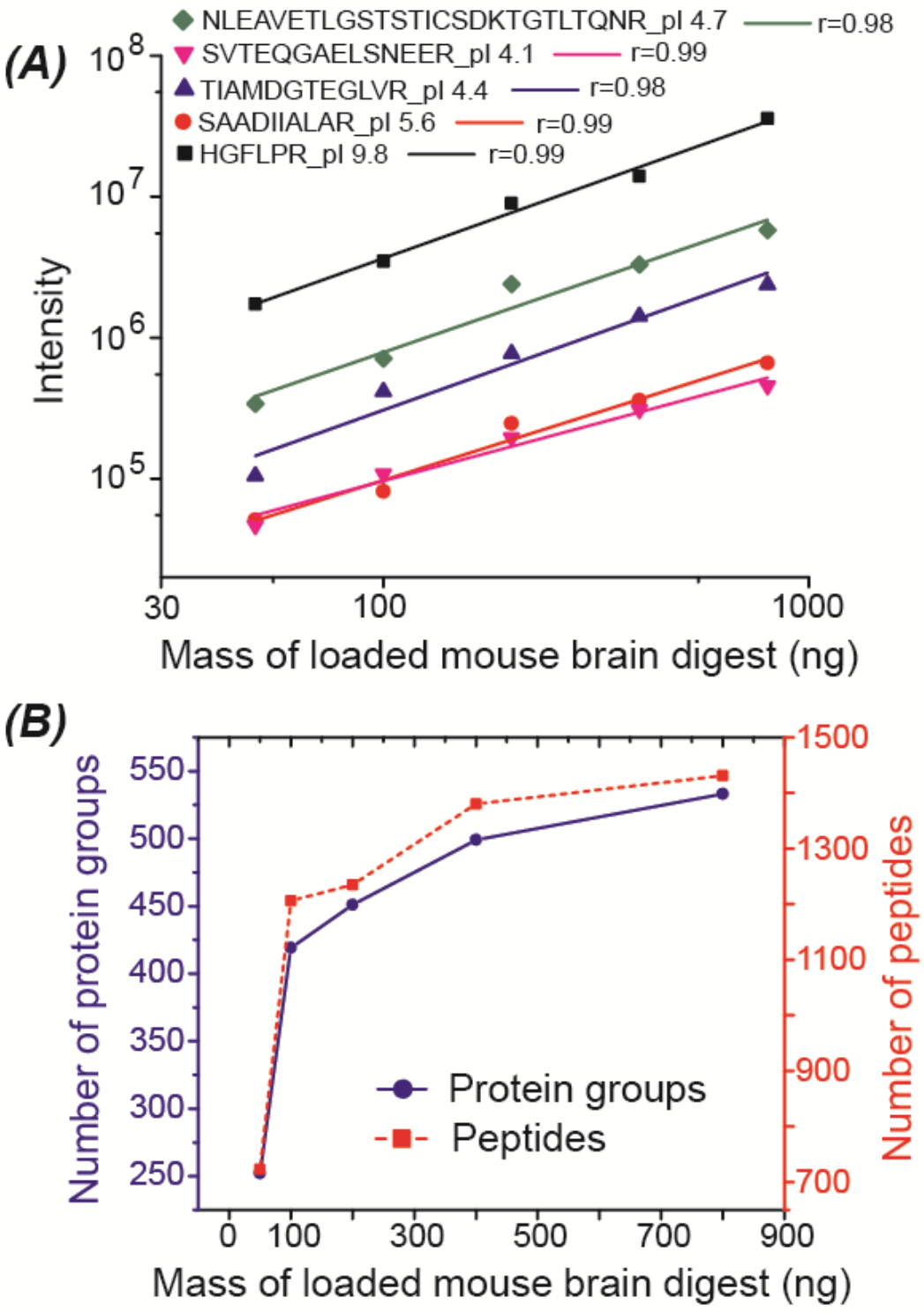

Figure 2.9. (A) Correlations of the mass of loaded mouse brain proteome digests and peptide intensity after the dynamic pH junction based CZE-MS/MS analysis. Five peptides with different length, isoelectric points (pIs) and intensity were chosen for the analysis. (B) Relationship between the mass of loaded mouse brain proteome digest and the number of peptide and protein group IDs after the dynamic pH junction based CZE-MS/MS analysis.

2.3.5. Large-scale BUP with the dynamic pH junction based CZE-MS system

Firstly, we performed two CZE-MS/MS runs with different concentrations (0.4 mg/mL and 4 mg/mL) of mouse brain proteome digests. For the 0.4 mg/mL sample, 500 nL of sample was injected. For the 4 mg/mL sample, 50 nL of sample was injected. The total mass of the loaded mouse brain proteome digests for those two CZE-MS/MS analyses are the same (200 ng). We observed that the CZE-MS/MS analysis of the 0.4 mg/mL sample with 500 nL injection volume generated significantly higher numbers of protein and peptide IDs than the analysis of the 4 mg/mL sample with 50 nL injection volume (423 vs. 294 protein groups and 1,159 vs. 797 peptides). Two factors affect the number of peptide IDs. One is the peptide intensity and the other one is the peptide separation. As shown in **Figure 2.10**, 0.4 mg/mL of digest with 500 nL injection volume generates comparable peptide intensity and much wider separation window compared with 4 mg/mL of digest with 50 nL injection volume, thus leading to much higher numbers of high-quality MS/MS spectra and peptide IDs. The result provides us with two important information. First, the dynamic pH junction based CZE-MS/MS system can significantly benefit analysis of mass-limited proteome samples. Second, the dynamic pH junction based CZE-MS/MS system can enable large-scale proteome analysis with much lower mass of sample material compared with typical CZE-MS/MS systems, thus leading to much higher overall sensitivity.

Secondly, we employed RPLC fractionation and the dynamic pH junction based CZE-MS/MS for large-scale bottom-up proteomics with 50 µg and 500 µg of mouse brain proteome digests. Fifteen LC fractions and 30 LC fractions were analyzed by CZE-MS/MS for the 50 µg and 500 µg sample, respectively. Each LC fraction from the 500-µg sample was dissolved in 8 µL of 10 mM NH4HCO3 (pH 8.0) for CZE-

MS/MS analysis. Each LC fraction from the 50-µg sample was dissolved in 4 µL of 10 mM NH_4HCO_3 (pH 8.0) for CZE-MS/MS analysis. The sample loading volume of CZE-MS/MS was 500 nL per run. An LTQ-XL mass spectrometer was used in these experiments. About 3,000 protein groups and 13,000 peptides were identified from the 500 µg of starting material with peptide-level FDR less than 1%. About 1,600 protein groups and 5,200 peptides were identified from only 50 µg of the complex proteome digest. The identified protein groups are listed in supporting material II.

In the literature, 3,000-4,000 protein group IDs from complex proteome samples have been approached by Dovichi and Lindner groups with offline RPLC-CZE-MS/MS.20,55 LTQ-Orbitrap Velos or LTQ-Orbitrap XL was used in those works, and 600 µg-1 mg of complex proteome digests were used for RPLC fractionation. In this work, we identified slightly lower number of protein group IDs (3,000 vs. 4,000 protein groups) with 500 µg of initial proteome digest and only LTQ-XL mass spectrometer.

We identified 1,600 protein groups from only 50 µg of the complex proteome digest with offline RPLC fractionation and the dynamic pH junction based CZE-MS/MS (LTQ-XL). We compared our data with the LTQ-XL based MudPIT datasets in the literature. Typically, offline multi-dimensional separation-MS/MS has lower sensitivity than online MudPIT due to the sample loss during fraction collection [56]. Our offline RPLC-CZE-MS/MS has similar sensitivity to the online MudPIT technique in terms of the number of protein group IDs from similar mass of initial complex proteome digests [57-61]. The result here is very important. Offline RPLC-CZE-MS/MS has been suggested as a valuable platform for large-scale proteomics [20,55], but hundreds of microgram of proteome digests are typically required due to the low loading capacity of CZE and also limited separation window, thus leading to

much lower overall sensitivity compared with MudPIT (typically 30 µg of peptides/run). This is one bottleneck of CZE-MS/MS for deep proteomics. Our dynamic pH junction based CZE-MS/MS has much larger loading capacity (0.5-1 µL vs. low nL) and significantly wider separation window (140 min vs. 60 min or narrower) than typical CZE-MS/MS systems, thus leading to much higher overall sensitivity. In addition, in this work we simply adjusted the position of the injection end of the separation capillary in the sample vial and guaranteed that the injection end of the separation capillary could be immersed in the sample for injection when only 4 µL of sample was in the sample vial. Injection of 500 nL of sample into the separation capillary from 4 µL of sample in the vial represents the use of 12% of the available sample for CZE-MS analysis, which is at least 6 times higher than that in typical CZE-MS systems (12% vs. 2% or lower) [20,25,28]. Those are the major reasons that our offline RPLC-CZE-MS/MS can approach similar sensitivity to the online MudPIT technique in terms of the number of protein group IDs. Our dynamic pH junction based CZE-MS/MS system truly opens the door of CZE-MS/MS based deep proteomics.

In this work, our RPLC-CZE-MS/MS system (LTQ-XL) need 2-3 days for comprehensive analysis of complex proteome digests. Typical 2D-LC-MS/MS approaches require about 1-day MS time for deep proteomics profiling using modern mass spectrometers [6,8,11]. The MS time required by our RPLC-CZE-MS/MS system can be easily attenuated by using one modern mass spectrometer with high mass accuracy and high scan speed.

We believe that coupling MDLC prefractionation (e.g., SCX-RPLC) with the dynamic pH junction based CZE-MS/MS will further greatly enhance the separation

of complex proteome digests, thus leading to higher protein sequence coverage, better characterization of protein isoforms and deeper proteome coverage.

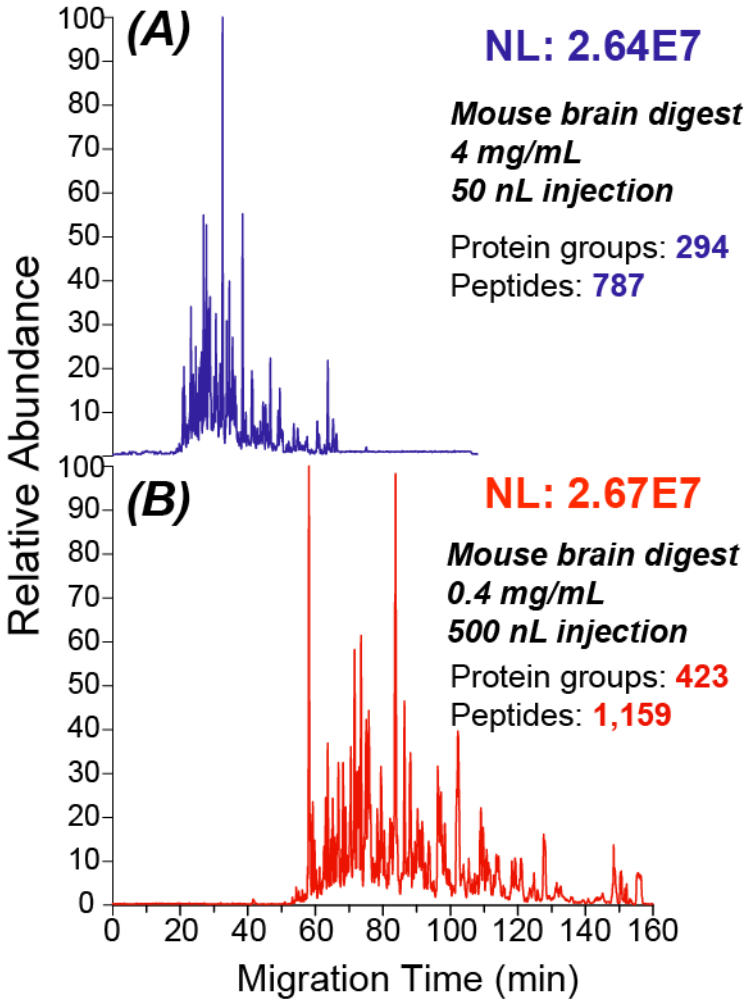

Figure 2.10. Electropherograms of the mouse brain proteome digests in 10 mM NH_4HCO_3 (pH 8) after CZE-MS/MS analysis. (A) 4 mg/mL of the mouse brain proteome digest with 50 nL injection volume; (B) 0.4 mg/mL of the mouse brain proteome digest with 500 nL injection volume.

2.4. Conclusions

In this work, we presented one automated CZE-MS system with microliter-scale loading capacity, 140-min peptide separation window and high peak capacity (~380), which represent the widest peptide separation window and the highest peak capacity for peptide separation using CZE-MS. It is the first time that CZE-MS approaches both microliter-scale loading capacity and over 2-hour separation window for analysis of complex samples. Coupling RPLC fractionation with the CZE-MS/MS (LTQ-XL) yielded 1,600 and 3,000 protein group IDs from 50 µg and 500 µg of mouse brain proteome digests, respectively. The results clearly demonstrate that CZE-MS/MS is ready for large-scale proteomics.

We expect that coupling SCX-RPLC fractionation with CZE-MS/MS will produce an invaluable platform for ultra-deep bottom-up proteomics. The platform will provide orthogonal and high capacity peptide separation. The platform combines the advantages of SCX (large loading capacity), RPLC (high-resolution peptide separation and on-line peptide desalting) and CZE-MS (high efficiency and high sensitivity for peptide separation/detection).

We also expect that the separation window and peak capacity of CZE-MS can be further significantly improved by using a longer separation capillary (e.g., 1.5 meters) and higher separation voltage across the capillary (e.g., 60 kV or higher).

2.5. Acknowledgments

We thank Prof. Xuefei Huang's group at Department of Chemistry, Michigan State University for kindly providing the mouse samples for our experiments. This research was funded by the Michigan State University.

REFERENCES

REFERENCES

[1] Washburn, M. P.; Wolters, D.; Yates, J. R.; 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001,** 19(3), 242-7.

[2] Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The one hour yeast proteome. *Mol. Cell. Proteomics* **2014**, 13(1), 339-347.

[3] Webb, K. J.; Xu, T.; Park, S. K.; Yates, J. R.; 3rd. Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. *J. Proteome Res.* **2013**, 12, 2177-2184.

[4] Wang, F.; Dong, J.; Jiang, X.; Ye, M.; Zou, H., Capillary trap column with strong cation-exchange monolith for automated shotgun proteome analysis., *Anal. Chem.* **2007**, 79, 6599-6606.

[5] Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, 3, 1154-1169.

[6] Low, T. Y.; van, Heesch, S.; van, den, Toorn, H.; Giansanti, P.; Cristobal, A.; Toonen, P.; Schafer, S.; Hübner, N.; van, Breukelen, B.; Mohammed, S.; Cuppen, E.; Heck, A. J.; Guryev, V., Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep.* **2013**, 5, 1469-1478.

[7] Wang, Y.; Yang, F.; Gritsenko, M. A.; Wang, Y.; Clauss, T.; Liu, T.; Shen, Y.; Monroe, M. E.; Lopez-Ferrer, D.; Reno, T.; Moore, R. J.; Klemke, R. L.; Camp, D. G, 2nd.; Smith, R. D. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **2011**, 11, 2019-2026.

[8] Ding, C.; Jiang, J.; Wei, J.; Liu, W.; Zhang, W.; Liu, M.; Fu, T.; Lu, T.; Song, L.; Ying, W.; Chang, C.; Zhang, Y.; Ma, J.; Wei, L.; Malovannaya, A.; Jia, L.; Zhen, B.; Wang, Y.; He, F.; Qian, X.; Qin, J. A fast workflow for identification and quantification of proteomes. *Mol. Cell. Proteomics* **2013**, 12, 2370-2380.

[9] Mertins, P.; Qiao, J. W.; Patel, J.; Udeshi, N. D.; Clauser, K. R.; Mani, D. R.; Burgess, M. W.; Gillette, M. A.; Jaffe, J. D.; Carr, S. A. Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat. Methods* **2013**, 10, 634-637.

[10] Song, C.; Ye, M.; Han, G.; Jiang, X.; Wang, F.; Yu, Z.; Chen, R.; Zou, H. Reversed-phase-reversed-phase liquid chromatography approach with high

orthogonality for multidimensional separation of phosphopeptides. *Anal. Chem.* **2010**, 82, 53-56.

[11], Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **2012**, 11, M111.014050.

[12] Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, 509, 575-581.

[13] Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, 509, 582-587.

[14] Zhou, F.; Sikorski, T. W.; Ficarro, S. B.; Webber, J. T.; Marto, J. A. Online nanoflow reversed phase-strong anion exchange-reversed phase liquid chromatography-tandem mass spectrometry platform for efficient and in-depth proteome sequence analysis of complex organisms. *Anal. Chem.* **2011**, 83, 6996-7005.

[15], Zhou, F.; Lu, Y.; Ficarro, S. B.; Adelmant, G.; Jiang, W.; Luckey, C. J.; Marto, J. A. Genome-scale proteome quantification by DEEP SEQ mass spectrometry. *Nat. Commun.* **2013**, 4, 2171.

[16] Chen, W.; Adhikari, S.; Chen, L.; Lin, L.; Li, H.; Luo, S.; Yang, P.; Tian, R. 3D-SISPROT: A simple and integrated spintip-based protein digestion and three-dimensional peptide fractionation technology for deep proteome profiling. *J. Chromatogr. A* **2017**,1498, 207-214.

[17] Chen, W.; Wang, S.; Adhikari, S.; Deng, Z.; Wang, L.; Chen, L.; Ke, M.; Yang, P.; Tian, R. Simple and Integrated Spintip-Based Technology Applied for Deep Proteome Profiling. *Anal. Chem*. **2016**, 88, 4864-4871.

[18] Wang, Y.; Fonslow, B. R.; Wong, C. C.; Nakorchevsky, A.; Yates, J. R.; 3rd. Improving the comprehensiveness and sensitivity of sheathless capillary electrophoresis-tandem mass spectrometry for proteomic analysis. *Anal. Chem.* **2012**, 84, 8505-8513.

[19] Faserl, K.; Sarg, B.; Kremser, L.; Lindner, H. Optimization and evaluation of a sheathless capillary electrophoresis-electrospray ionization mass spectrometry platform for peptide analysis: comparison to liquid chromatography-electrospray ionization mass spectrometry. *Anal. Chem.* **2011**, 83, 7297-7305.

[20] Faserl, K.; Kremser, L.; Müller, M.; Teis, D.; Lindner, H. H. Quantitative proteomics using ultralow flow capillary electrophoresis-mass spectrometry. *Anal. Chem.* **2015**, 87, 4633-4640.

[21] Heemskerk, A. A.; Deelder, A. M.; Mayboroda, O. A. CE-ESI-MS for bottom-up proteomics: Advances in separation, interfacing and applications. *Mass Spectrom. Rev.* **2016**, 35, 259-271.

[22] Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J. Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry as an alternative proteomics platform to ultraperformance liquid chromatography-electrospray ionization-tandem mass spectrometry for samples of intermediate complexity. *Anal. Chem.* **2012**, 84, 1617-1622.

[23] Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J. Single-shot proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with production of more than 1250 Escherichia coli peptide identifications in a 50 min separation. *Anal. Chem.* **2013**, 85, 2569-2573.

[24] 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.

[25] Sun, L.; Hebert, A. S.; Yan, X.; Zhao, Y.; Westphall, M. S.; Rush, M. J.; Zhu, G.; Champion, M. M.; Coon, J. J.; Dovichi, N. J. Over 10,000 peptide identifications from the HeLa proteome by using single-shot capillary zone electrophoresis combined with tandem mass spectrometry. *Angew. Chem. Int. Ed.* **2014**, 53, 13931-13933.

[26] Ludwig, K. R.; Sun, L.; Zhu, G.; Dovichi, N. J.; Hummon, A. B. Over 2300 phosphorylated peptide identifications with single-shot capillary zone electrophoresis-tandem mass spectrometry in a 100 min separation. *Anal. Chem.* **2015**, 87, 9532-9537.

[27] Choi, S. B.; Zamarbide, M.; Manzini, M. C.; Nemes, P. Tapered-Tip Capillary Electrophoresis Nano-Electrospray Ionization Mass Spectrometry for Ultrasensitive Proteomics: the Mouse Cortex. *J. Am. Soc. Mass Spectrom.* **2017**, 28(4), 597-607.

[28] Lombard-Banek, C.; Moody, S. A.; Nemes, P. Single-Cell Mass Spectrometry for Discovery Proteomics: Quantifying Translational Cell Heterogeneity in the 16-Cell Frog (Xenopus) Embryo. *Angew. Chem. Int. Ed.* **2016**, 55, 2454-2458.

[29] Guo, X.; Fillmore, T. L.; Gao, Y.; Tang, K. Capillary Electrophoresis-Nanoelectrospray Ionization-Selected Reaction Monitoring Mass Spectrometry via a True Sheathless Metal-Coated Emitter Interface for Robust and High-Sensitivity Sample Quantification. *Anal. Chem.* **2016**, 88, 4418-4425.

[30] Wang, C.; Lee, C. S.; Smith, R.D.; Tang, K. Ultrasensitive sample quantitation via selected reaction monitoring using CITP/CZE-ESI-triple quadrupole MS. *Anal. Chem*. **2012**, 84, 10395-10403.

[31] Jorgenson, J. W.; Lukacs, K. D. Capillary zone electrophoresis. *Science* **1983,** 222(4621), 266-72.

[32] Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates, J. R.; 3rd. In-line separation by capillary electrophoresis prior to analysis by top-down mass spectrometry enables sensitive characterization of protein complexes. *J. Proteome Res.* **2014**, 13, 6078-6086.

[33] 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.

[34] Moini, M. Simplifying CE-MS operation. 2. Interfacing low-flow separation techniques to mass spectrometry using a porous tip. *Anal. Chem.* **2007**, 79, 4241-4246.

[35] Mellors, J. S.; Gorbounov, V.; Ramsey, R. S.; Ramsey, J. M. Fully integrated glass microfluidic device for performing high-efficiency capillary electrophoresis and electrospray ionization mass spectrometry. *Anal. Chem.* **2008**, 80, 6881-6887.

[36] 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.

[37] Magdeldin, S.; Moresco, J. J.; Yamamoto, T.; Yates, J. R.; 3rd. Off-Line Multidimensional Liquid Chromatography and Auto Sampling Result in Sample Loss in LC/LC-MS/MS. *J. Proteome Res.* **2014**, 13, 3826-3836.

[38] Heemskerk, A. A.; Busnel, J. M.; Schoenmaker, B.; Derks, R. J.; Klychnikov, O.; Hensbergen, P. J.; Deelder, A. M.; Mayboroda, O. A. Ultra-low flow electrospray ionization-mass spectrometry for improved ionization efficiency in phosphoproteomics. *Anal. Chem.* **2012**, 84, 4552-4559.

[39] Busnel, J. M.; Schoenmaker, B.; Ramautar, R.; Carrasco-Pancorbo, A.; Ratnayake, C.; Feitelson, J. S.; Chapman, J. D.; Deelder, A. M.; Mayboroda, O. A. High capacity capillary electrophoresis-electrospray ionization mass spectrometry: coupling a porous sheathless interface with transient-isotachophoresis. *Anal. Chem.* **2010**, 82, 9476-9483.

[40] Zhang, Z.; Sun, L.; Zhu, G.; Yan, X.; Dovichi, N. J., Integrated strong cation-exchange hybrid monolith coupled with capillary zone electrophoresis and simultaneous dynamic pH junction for large-volume proteomic analysis by mass spectrometry. *Talanta* **2015**, 138, 117-122.

[41] 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.

[42] Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J. Bottom-up proteomics of Escherichia coli using dynamic pH junction preconcentration and capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry. *Anal. Chem.* **2014**, 86, 6331-6336.

[43] Aebersold, R.; Morrison, H. D. Analysis of dilute peptide samples by capillary zone electrophoresis. *J. Chromatogr.* **1990**, 516, 79–88.

[44] 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**, 72,1242–1252.

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

[46] Ludwig, K. R.; Sun, L.; Zhu, G.; Dovichi, N. J.; Hummon, A. B. Over 2300 phosphorylated peptide identifications with single-shot capillary zone electrophoresis-tandem mass spectrometry in a 100 min separation. *Anal. Chem.* **2015**, 87(19), 9532-7.

[47] 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.

[48] 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.

[49] Kazarian, A. A, Hilder, E. F.; Breadmore, M. C. Online sample pre-concentration via dynamic pH junction in capillary and microchip electrophoresis. *J. Sep. Sci.* **2011**, 34, 2800–2821.

[50] Imami, K.; Monton, M. R.; 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.

[51] Hasan, M. N.; Park, S. H.; Oh, E.; Song, E. J.; Ban, E.; Yoo, Y. S. Sensitivity enhancement of CE and CE-MS for the analysis of peptides by a dynamic pH junction. *J. Sep. Sci.* **2010**, 33, 3701–3709.

[52] Cao, C. X.; Fan, L. Y.; Zhang, W. Review on the theory of moving reaction boundary, electromigration reaction methods and applications in isoelectric focusing and sample pre-concentration. *Analyst* **2008**, 133, 1139–1157.

[53] Ptolemy, A. S.; Britz-McKibbin, P. New advances in on-line sample preconcentration by capillary electrophoresis using dynamic pH junction. *Analyst* **2008**, 133, 1643–1648.

[54] Zhang, Z.; Sun, L.; Zhu, G.; Cox, O, F.; Huber, P. W.; Dovichi, N. J. Nearly 1000 Protein Identifications from 50 ng of Xenopus laevis Zygote Homogenate Using

Online Sample Preparation on a Strong Cation Exchange Monolith Based Microreactor Coupled with Capillary Zone Electrophoresis. *Anal. Chem.* **2016,** 88(1), 877-82.

[55] Yan, X.; Sun, L.; Zhu, G.; Cox, O. F.; Dovichi, N. J. Over 4100 protein identifications from a Xenopus laevis fertilized egg digest using reversed-phase chromatographic prefractionation followed by capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry analysis. *Proteomics* **2016**, 16(23):2945-2952.

[56] Magdeldin, S.; Moresco, J. J.; Yamamoto, T.; Yates, J. R.; 3rd. Off-Line Multidimensional Liquid Chromatography and Auto Sampling Result in Sample Loss in LC/LC-MS/MS. *J. Proteome Res.* **2014**, 13(8):3826-36.

[57] Sun, L.; Zhang, Y.; Tao, D.; Zhu, G.; Zhao, Q.; Wu, Q.; Liang, Z.; Yang, L.; Zhang, L.; Zhang, Y. SDS-PAGE-free protocol for comprehensive identification of cytochrome P450 enzymes and uridine diphosphoglucuronosyl transferases in human liver microsomes. *Proteomics* **2012**, 12, 3464-3469.

[58] Sun, L.; Tao, D.; Han, B.; Ma, J.; Zhu, G.; Liang, Z.; Shan, Y.; Zhang, L.; Zhang, Y. Ionic liquid 1-butyl-3-methyl imidazolium tetrafluoroborate for shotgun membrane proteomics. *Anal. Bioanal. Chem*. **2011**, 399, 3387-3397.

[59] Tao, D.; Qiao, X.; Sun, L.; Hou, C.; Gao, L.; Zhang, L.; Shan, Y.; Liang, Z.; Zhang, Y. Development of a highly efficient 2-D system with a serially coupled long column and its application in identification of rat brain integral membrane proteins with ionic liquids-assisted solubilization and digestion. *Proteome Res.* **2011**, 10, 732-738.

[60] Wang, F.; Dong, J.; Ye, M.; Jiang, X.; Wu, R.; Zou, H. Online multidimensional separation with biphasic monolithic capillary column for shotgun proteome analysis. *J. Proteome Res.* **2008**, 7, 306-310.

[61] Wang, F.; Dong, J.; Jiang, X.; Ye, M.; Zou, H. Capillary trap column with strong cation-exchange monolith for automated shotgun proteome analysis. *Anal. Chem.* **2007**, 79, 6599-6606.

CHAPTER 3. Strong cation exchange-reversed phase liquid chromatography-capillary zone electrophoresis-tandem mass spectrometry (SCX-RPLC-CZE-MS/MS) platform for deep bottom-up proteomics and phosphoproteomics

Part of this chapter was adapted from Anal. Chim. Acta 2018, 1012, 1-9; and Anal. Chem. 2019, 91(3), 2201-2208 with permission

3.1. SCX-RPLC-CZE-MS/MS platform with high peak capacity for deep bottom-up proteomics

3.1.1. Introduction

The state-of-the-art 2D-LC-MS/MS has approached over 8000 protein IDs from mammalian cell lines or tissues in 1-3 days of mass spectrometer time [1-5]. The draft human proteome containing 84% of the total annotated protein-coding genes in humans has also been generated using 2D-LC-MS/MS [6]. Over 2,000 LC-MS runs were performed for the draft human proteome, but the median protein sequence coverage was still only 28% [6]. The typical median protein sequence coverage of deep bottom-up proteomics datasets is around 25% or lower. The low sequence coverage impedes the confident identification of protein isoforms.

Alternative separation techniques that are orthogonal to LC for peptide separation will be very useful to further improve the number of peptide IDs from complex proteomes in bottom-up proteomics experiments, boosting the protein sequence coverage. CZE-MS/MS has been suggested as an alternative to LC-MS/MS for bottom-up proteomics [7-18]. CZE separates peptides based on their size-to-charge ratios and it is orthogonal to LC for peptide separation. CZE-MS/MS and LC-MS/MS are complementary in protein/peptide ID from complex proteome

digests [7-11]**.** CZE tends to identify small, basic and hydrophilic peptides compared with RPLC-MS. In addition, the migration time of peptides from CZE-MS can be predicted more accurately and easily than their retention time from commonly used RPLC-MS [19]. The electrophoretic mobilities of peptides in CZE mainly relate to their size (molecular mass) and charge, which are relatively easy to be determined accurately. The retention of peptides in RPLC can be affected by various factors, *e.g.*, hydrophobic, hydrogen-bond, and ion-pairing interactions. Modeling those factors is very difficult. Recently, Krokhin *et al.* developed a simple model for CZE and approached a very good correlation (R^2 ~0.995) between the experimental and predicted migration time of peptides in CZE based on a large-scale peptide dataset [19]. The capability for accurate prediction of peptide migration time in CZE makes CZE-MS become a powerful tool for bottom-up proteomics because it can help us further evaluate the confidence of peptide ID from the database search and even guide the database search.

Although CZE-MS has many valuable features for bottom-up proteomics, the number of protein IDs from complex proteomes using CZE-MS/MS is still much lower than the state of the art using 2D-LC-MS/MS. Much effort has been made to improve the CZE-MS for large-scale proteomics [7,15,20,21]. Sun *et al.* approached 2,000 protein and 10,000 peptide IDs from a human cell line digest using single shot CZE-MS/MS with a neutrally coated separation capillary and an Orbitrap Fusion mass spectrometer [7]. FASS was used to improve the sample loading volume to 100 nL and a 1-meter long neutrally coated separation capillary was employed to improve the peak capacity to about 300 [7]. Yan *et al.* coupled RPLC prefractionation to CZE-MS/MS for bottom-up proteomics of *Xenopus* embryos, resulting in the identification of over 4,000 proteins [20]. For each CZE-MS/MS run, about 50 nL of the sample

was injected for analysis. Faserl *et al.* coupled RPLC prefractionation to CZE-MS/MS for quantitative proteomics of yeast, leading to the identification of over 3,000 proteins [15]. A 1.5-mg yeast digest was used as the starting material and the sample loading volume for CZE-MS/MS was 40 nL. Very recently, Faserl *et al.* approached 6,000 protein IDs from a human cell line proteome digest by RPLC prefractionation and sequential sample injection based CZE-MS/MS with 2 mg of peptides as the starting material [21]. The sample loading volume of CZE-MS/MS was 25 nL.

In order to further improve the CZE-MS/MS for significantly deeper proteome coverage with a reasonable mass of initial protein material, we need to improve the sample loading volume of CZE-MS/MS and meantime boost the overall peak capacity of the system. The improvement in both sample loading volume and peak capacity can evidently benefit the identification of low abundant proteins. We showed that dynamic pH junction based CZE-MS/MS could approach both micro-liter scale sample loading volume and high peak capacity (up to 380) for analysis of complex peptide or protein mixtures [22,23]. In this work, we coupled online SCX-RPLC prefractionation to the dynamic pH junction based CZE-MS/MS for deep bottom-up proteomics. The orthogonal SCX-RPLC-CZE platform approached a very high peak capacity (~7,000). Because of the high peak capacity and the large sample loading volume of CZE (~0.5 µL per run), the SCX-RPLC-CZE-MS/MS system identified 8,200 protein groups and 65,000 unique peptides from a mouse brain proteome digest.

3.1.2. Experimental

3.1.2.1. Reagents and chemicals

All reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. LC/MS grade water, methanol, ACN, HPLC grade AA, FA, and HF were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide was ordered from Acros Organics (NJ, USA). Fused silica capillaries (50 µm i.d./360 µm o.d.) were bought from Polymicro Technologies (Phoenix, AZ). The 30 kDa Centrifugal Filter Units with Ultracel-30 membrane were purchased from Merck Millipore (Burlington, MA). Complete, mini protease inhibitor cocktail was purchased from Roche (Indianapolis, IN).

3.1.2.2. Preparation of the linear polyacrylamide-coated capillary for CZE

The inner wall of the separation capillaries for CZE was coated with LPA based on references [22] and [24] in order to reduce the EOF. The separation capillaries for CZE used in this study were coated with LPA. First, a bare fused silica capillary (50 µm i.d./360 µm o.d.) was sequentially flushed with 1 M sodium hydroxide, water, 1 M hydrochloric acid, water and methanol. Then the capillary was flushed with nitrogen for 4 h. Next, the capillary was filled with 50% (v/v) 3-(trimethoxysiyl)propyl-methacrylate in methanol, and kept at room temperature for 24 h with both ends sealed with silica rubber. After that, the capillary was flushed with methanol to remove the excess reagent, followed by nitrogen flushing for 24 h.

The solution of 5% (w/v) APS in water and the solution of 40 mg/mL of acrylamide in water were prepared. 3.5 µL of the 5% APS solution was added to 500 µL of the acrylamide solution. The mixed solution was degassed under nitrogen for

15 min. Then the pretreated capillary was filled with the mixed solution and incubated in 50 ^{o}C water bath for 35 min with sealed ends. After incubation, the redundant reagents were flushed out with water. Lastly, one end of the coated capillary (~5 mm long) was etched with HF to reduce the outer diameter of the capillary to ~70 µm based on the protocol in reference [14]. The capillary was stored at room temperature before use.

3.1.2.3. Sample preparation

The mouse brain tissue samples originated from two 6-month-old BL/6 male mice were kindly provided by Professor Chen Chen's group at Department of Animal Science, Michigan State University. All animal-related experiments were conducted strictly following the guidelines defined by the Institutional Animal Care and Use Committee of Michigan State University.

The mouse brains were cut into small pieces and washed with PBS multiple times for the removal of blood. Then the sample was suspended in 8 M urea in 100 mM NH_4HCO_3 (pH 8) with complete protease inhibitor. The suspended sample was homogenized using a Homogenizer 150 (Fisher Scientific, Pittsburgh, PA) on ice until it was completely homogenized. The sample was then sonicated using a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 5 min. The lysate was aliquoted to 1.7-mL Eppendorf tubes and centrifuged at 10,000 g for 5 min. A small aliquot of the supernatants was subjected to BCA assay for protein concentration measurement. The remaining supernatants were subjected to acetone precipitation. The supernatants and cold acetone were mixed at the ratio of 1:4 (v/v) and stored in -20 ^{o}C overnight followed by centrifugation at 10,000 g for 5 min. The supernatants were discarded. Cold acetone was added to the Eppendorf tubes to wash the pellets

again. After centrifugation, the supernatants were discarded and the pellets were air dried in the chemical hood for several minutes, followed by storage at -80 ° C before use.

Filter-aided sample preparation (FASP) method was used for protein digestion [31]. ~500 µg of mouse brain proteins were dissolved in 125 µL of 2% sodium dodecyl sulfate (SDS) (w/v) and 100 mM NH_4HCO_3 solution (pH 8). The solution was heated at 90 °C for 10 min for protein denaturation. Then, 1 µL of DTT solution (1 M in 100 mM NH_4HCO_3) was added to the sample solution and the mixture was heated at 80 °C for 20 min for protein reduction. After that, 2.5 µL of 1 M IAA solution in 100 mM NH_4HCO_3 was added to the sample solution and the mixture was kept at room temperature for 10 min (in dark) for protein alkylation. Then 1 µL of 1 M DTT solution was added to the sample solution in order to react with the excess IAA. The sample was then mixed with 375 µL of 8 M urea in 100 mM NH_4HCO_3. Next, the mixture (~500 µL) was added onto two 30-kDa centrifugal filter units (250 µL/unit), followed by centrifugation at 14,000 g for 15 min. The proteins on the membrane were washed with 250 µL of 8 M urea in 100 mM NH_4HCO_3 three times. Next, the proteins were washed with 100 mM NH_4HCO_3 three times to remove urea. Finally, 150 µL of 100 mM NH_4HCO_3 (pH 8.0) was loaded on each membrane and 8 µL of trypsin solution (1 µg/µL) was added to each unit. The filter units were gently vortexed for 5 min to mix the trypsin and proteins. After that, the filter units were kept in a 37 °C water bath for 12 h for tryptic digestion. After digestion, the units were centrifuged at 15,000 g for 15 min, and the flow-through containing the peptides was collected. To increase peptide recovery from the membrane, the membrane was further washed with 150 µL of 100 mM NH_4HCO_3. The flow-through from those two steps were combined and the peptide solution was acidified with FA, followed by peptide

desalting with C18 SPE columns (Waters, Milford, MA). The eluates from the C18 SPE columns were lyophilized with a vacuum concentrator (Thermo Fisher Scientific) and stored at -80 °C until use. In total, we performed the FASP workflow described above three times, and prepared ~1.5 mg of mouse brain proteome digests for the following online SCX-RPLC and high pH RPLC fractionation.

3.1.2.4. Online SCX-RPLC fractionation of a mouse brain proteome digest

An Agilent Infinity II HPLC system with a quaternary pump was used for the experiment. A SCX trap column (Zorbax 300SCX, 4.6 mm i.d. × 12.5 mm length, 5 µm particles, Agilent Technologies) and a C18 RP column (Zorbax 300Extend-C18, 2.1 mm i.d. × 150 mm length, 3.5 µm particles, Agilent Technologies) were directly connected with a PEEK tubing and two fittings for online 2D-LC separation. 0.1% FA in water (mobile phase A), 0.1% FA in ACN (mobile phase B) and 890 mM ammonium acetate solution (pH=2.88) (mobile phase C) were used for separation. Mobile phase A and C were used for the generation of different salt concentrations for step-wise elution of peptides from the SCX column to the RPLC column. Mobile phases A and B were used for gradient separation of peptides with RPLC.

A 500-µg mouse brain proteome digests dissolved in mobile phase A were injected into a sample loop and loaded onto the SCX column by pushing the sample through the system with mobile phase A at 0.3 mL/min flow rate for 5 min. The peptides trapped on the SCX column were eluted in a step-wise fashion with different concentrations of ammonium acetate for 20 min at a flow rate of 0.3 mL/min. After each salt step elution, the eluted peptides from the SCX were captured on the RPLC column, followed by RPLC gradient separation at 0.3 mL/min for 90 min: 0-20 min, 2%B; 20-22 min, 2-6%B; 22-67 min, 6-40%B; 67-72 min, 40-80% B;

72-77 min, 80%B; 77-80 min, 80-2%B; 80-90 min, 2%B. 40 fractions were collected for each RPLC run from 25 min to 71 min. From 25-31 min and 65-71 min, we collected each fraction every 2 min; from 31 min to 65 min, we collected one fraction per min. We named the fractions based on the order of retention time from 1 to 40. Then we combined the fraction N and fraction N+20 to generate 20 fractions.

We performed two SCX-RPLC fractionation experiments. In the first experiment, we eluted the peptides from SCX with three different concentrations of ammonium acetate: 150 mM, 350 mM, and 890 mM. In total, we got 60 SCX-RPLC fractions (3 salt steps× 20 fractions/salt step) from this experiment. In the second experiment, we used two salt steps for peptide elution from the SCX with salt concentrations of 250 mM and 890 mM. In total, we collected 40 fractions for the second experiment. All of the collected fractions were lyophilized and stored at -80 ^{o}C for the following CZE-MS/MS experiments.

3.1.2.5. High-pH RPLC fractionation of the mouse brain proteome digest

The same Agilent Infinity II HPLC system was used for high pH RPLC fractionation. A C18 RP column (Zorbax 300Extend-C18, 2.1 mm i.d. × 150 mm length, 3.5 µm particles, Agilent Technologies) was used for separation. Mobile phase A (5 mM NH_4HCO_3 in water, pH 9) and mobile phase B (5 mM NH_4HCO_3 in 80% ACN, pH 9) were used to generate a gradient for peptide separation.

500-µg mouse brain proteome digest was injected onto the RP column for the experiment. The flow rate was 0.3 mL/min. The gradient was as follow: 0-5 min, 2% B; 5-7 min, 2-10% B; 7-67 min, 10-50% B; 67-69 min, 50-100 % B; 69-79 min, 100% B; 79-80 min, 100-2% B; 80-90 min, 2% B. In total, 60 fractions were collected from 7 min to 67 min, one fraction per min. We named the fractions based on the order

of retention time from 1 to 60. Then we combined the fraction N and fraction N+30 to generate 30 fractions. Those fractions were then lyophilized and stored at -80 ^{o}C for low pH RPLC-MS/MS.

3.1.2.6. CZE-ESI-MS/MS

For CZE-ESI-MS/MS, a commercialized electro-kinetically pumped sheath flow CE-MS interface (CMP Scientific, Brooklyn NY) was employed for coupling CZE to MS [25,26]. An ECE-001 CE autosampler (CMP Scientific) was used for the automated operation of CZE. The ESI emitter was pulled from a borosilicate glass capillary (1.0 mm o.d., 0.75 mm i.d.) with a Sutter P-1000 flaming/brown micropipette puller. The orifice of the ESI emitter was 20-40 µm. The BGE was 5% (v/v) AA with pH 2.4 and the sheath buffer was 0.2% (v/v) FA containing 10% (v/v) methanol. The etched end of the separation capillary was introduced into the ESI emitter, and the distance between the end of the capillary and the orifice of the ESI emitter was ~300 µm. The distance between the orifice of the emitter and the inlet of the mass spectrometer was ~2.0 mm. The voltage applied to the sample injection end of the capillary was 30 kV and ~2.2 kV was applied at the interface for ESI.

The 60 SCX-RPLC fractions from the three salt step experiment were dissolved in 4 µL of 10 mM NH_4HCO_3 (pH 8.0) for CZE-MS/MS. A 71-cm LPA-coated capillary (50-µm i.d., 360-µm o.d.) was used for CZE. Each fraction was injected into the capillary with 5-psi pressure for 63 s, corresponding to about 500-nL sample injection volume. After that, 30 kV was applied at the injection end for CZE separation for 50 min, followed by capillary flushing with BGE using 10 psi for 10 min. The 20 fractions from the second salt step (350 mM ammonium acetate) were further diluted from ~3.5 µL to 6 µL with 50 mM NH_4HCO_3 (pH 8). From those 20 fractions, we performed

CZE-MS/MS analysis again using a 92-cm long LPA-coated capillary (50-µm i.d., 360-µm o.d.). For those analyses, 5 psi for 87 s was used for sample injection, corresponding to ~500-nL sample injection volume. The separation was performed with 30-kV voltage for 90 min, followed by BGE flushing with 10 psi for 15 min.

The 40 SCX-RPLC fractions from the two-salt-step experiment were dissolved in 4 µL of 50 mM NH_4HCO_3 (pH 8). A 94-cm long LPA-coated capillary (50-µm i.d., 360-µm o.d.) was used for CZE. The sample was injected into the capillary using 5 psi for 92 s, corresponding to about 500-nL sample injection volume. Next, 30 kV was applied at the injection end for separation for 92 min, followed by BGE flushing with 10 psi for 13 min.

A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for all of the experiments. A Top10 DDA method was used. The mass resolution was set to 60,000 (at m/z 200) for both full MS scans and MS/MS scans. For full MS scans, the target value was 3E6, the maximum injection time was 50 ms and the scan range was 300 to 1500 m/z. For MS/MS scans, the target value was 1E5 and the maximum injection time was 110 ms. The ten most abundant ions in an MS spectrum with intensity higher than 1E5 were sequentially isolated in the quadrupole with an isolation window as 2 m/z, followed by fragmentation in the HCD cell with a normalized collision energy of 28%. Dynamic exclusion was applied and it was set to 30 s. Only ions with charge states as two or higher were considered for fragmentation.

3.1.2.7. RPLC-ESI-MS/MS

The 30 high-pH RPLC fractions were subjected to low pH RPLC-ESI-MS/MS analysis. An EASY-nLC™ 1200 system (Thermo Fisher Scientific) was used for

RPLC separation. Each fraction was dissolved in 10 µL of 0.1% (v/v) FA and 2% (v/v) ACN. 3 µL of the sample was loaded onto a C18 pre-column (Acclaim PrepMapTM 100, 75-µm i.d. × 2 cm, nanoviper, 3 µm particles, 100 Å, Thermo Scientific). Then, the peptides were separated on a C18 separation column (Acclaim PrepMapTM 100, 75-µm i.d. × 50 cm, nanoviper, 2 µm particles, 100 Å, Thermo Scientific) at a flow rate of 200 nL/min. Mobile phase A (2% (v/v) ACN in water containing 0.1% (v/v) FA) and mobile phase B (80% (v/v) ACN and 0.1% (v/v) FA) was used to generate the gradient for separation. For separation, a 90-min gradient was used: 0-70 min, 8-40% B; 70-72 min, 40-100% B; 72-90 min, 100% B. The LC system required another 30 min for column equilibration between runs. Therefore, one LC-MS run required about 2 h.

The same Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for the RPLC-MS/MS experiments. The spray voltage was set to 1.8 kV. The other detailed parameters were the same as CZE-MS/MS described above.

3.1.2.8. Data analysis

Proteome Discoverer 2.1 software (Thermo Fisher Scientific) was used for analyses of RAW files. Sequest HT database search engine was used for database search. The mouse proteome database (UP000000589) downloaded from UniProt (http://www.uniprot.org/) was used as the database. Both the forward and reversed databases were used for database search in order to evaluate the FDRs [27,28]. The enzyme was set as trypsin. The maximum number of missed cleavages was set as 2. The mass tolerances of precursor ions and fragment ions were set as 20 ppm and 0.05 Da, respectively. Oxidation (methionine) and deamination (Asparagine or Glutamine) were chosen as the dynamic modifications. The carbamidomethylation

(cysteine) was set as the static modification. The peptide IDs were filtered with peptide confidence as high, corresponding to less than 1% FDR. Protein grouping was enabled, and the strict parsimony principle was applied.

The grand average of hydropathy (GARVY) value of peptides was calculated with GARVY Calculator (http://www.gravy-calculator.de/). Isoelectric points (pIs) of peptides were calculated using the “Compute pI/Mw” tool in ExPASy (http://web.expasy.org/compute_pi/). The gene ontology (GO) information of proteins was observed using the DAVID bioinformatics resources 6.8 (https://david.ncifcrf.gov/) [29,30].

3.1.3. Results and discussion

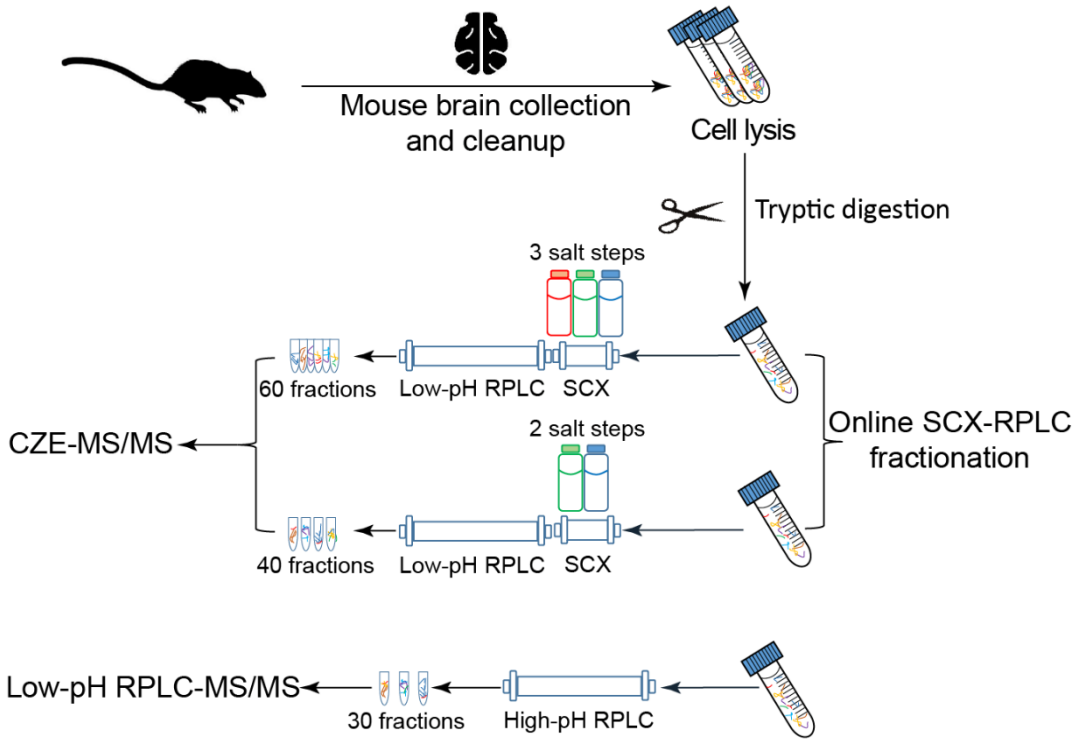

Figure 3.1. Experimental design of the work

As shown in **Figure 3.1**, proteins were extracted from mouse brains and were digested into peptides with trypsin based on the FASP method [31]. Three aliquots of mouse brain digests (500-µg peptides/aliquot) were used for prefractionation and MS/MS analysis. Two aliquots were fractionated by online SCX-RPLC. The peptides were trapped on an SCX trap column first, followed by step-wise elution from the SCX trap column to the RPLC column using three or two salt steps. The eluates were further separated by RPLC. In total, 60 fractions were collected from the three-salt-step SCX-RPLC experiment and 40 fractions were collected from the two-salt-step experiment. All of the fractions were analyzed by the CZE-MS/MS in 60 h (for the three-salt-step experiment) and 70 h (for the two-salt-step experiment). The dynamic pH junction method was used for on-line stacking of peptides to improve the

sample loading capacity of CZE [32,33]. The sample loading volume for each CZE-MS/MS run was about 500 nL. The third aliquot of the mouse brain digest was fractionated by high-pH RPLC into 30 fractions, and those fractions were analyzed by low-pH RPLC-MS/MS in 60 h.

3.1.3.1. SCX-RPLC-CZE-MS/MS for deep bottom-up proteomics of the mouse brain

Figure 3.2 presents the results from the mouse brain proteome digest using SCX-RPLC-CZE-MS/MS with three-salt-step elution (150 mM, 350 mM and 890 mM ammonium acetate, pH 2.88). The 60 SCX-RPLC fractions were analyzed by CZE-MS/MS with a 71-cm LPA-coated separation capillary in 60 h (1 h/fraction), leading to the identification of over 7,000 protein groups and 40,000 unique peptides, **Figure 3.2A**. The LC fractions from the second salt step (350 mM) made a significantly higher contribution to the overall peptide IDs than those from other two salt steps. We made two conclusions based on this preliminary experiment. First, we should be able to boost the overall protein/peptide IDs via improving the analyses of the twenty fractions from the 350-mM salt step. Second, we need to change the concentration of the ammonium acetate for peptide elution from SCX in the following experiments to maximize the protein/peptide IDs.

We further analyzed the twenty LC fractions from the 350-mM salt step with CZE-MS/MS based on a much longer LPA-coated separation capillary (92 cm *vs.* 71 cm). The long separation capillary produced much more protein group and unique peptide IDs than the short capillary, **Figure 3.2B**, boosting the protein group and unique peptide IDs from 6,000 to 7,100 and from about 27,000 to over 35,000, respectively. The improvement in protein and peptide IDs is most likely due to the

much wider separation window from the long separation capillary(55 min *vs.* 30 min), leading to more tandem mass spectra. As shown in **Figure 3.2C**, the CZE-MS system using the long separation capillary produced reasonably narrow peaks of peptides with the full width at half maximum (FWHM) ranging from 7.2 s to 36 s. The number of theoretical plates, on average, was around 240,000. The peak capacity of the CZE-MS run in **Figure 3.2C** was estimated to be around 170 based on the FWHM of the three selected peptides. We decided to use the long separation capillary-based CZE-MS/MS for following experiments due to the much better protein/peptide IDs, although the long separation capillary required a longer time for each CZE-MS/MS run compared with the short capillary (1.75 h *vs.* 1 h).

Next, we tried to improve the overall protein/peptide IDs via changing the concentration of the ammonium acetate for peptide elution from SCX based on the preliminary data from the three-salt-step experiment. We fractionated another 500-µg mouse brain peptides with SCX-RPLC into 40 fractions based on two salt steps (250 mM and 890 mM ammonium acetate, pH 2.88). The SCX-RPLC fractions were analyzed by CZE-MS/MS with a 94-cm separation capillary in 70 h (1.75 h/fraction). We increased the concentration of NH_4HCO_3 (pH 8.0) in the sample buffer from 10 mM to 50 mM in order to improve the stacking performance of the dynamic pH junction method [22,23,34]. As shown in **Figure 3.3**, the first and second salt steps made comparable contributions to the overall unique peptide IDs. In total, the platform identified nearly 8,200 protein groups and 65,000 unique peptides from the mouse brain proteome (**Figure 3.3**), representing the largest proteomics dataset using CZE-MS/MS. CZE-MS/MS analysis of the fractions from the first salt step alone produced nearly 7,000 protein group IDs in 35 h. The data clearly suggest that CZE-MS/MS has the capability for deep sequencing of complex proteomes.

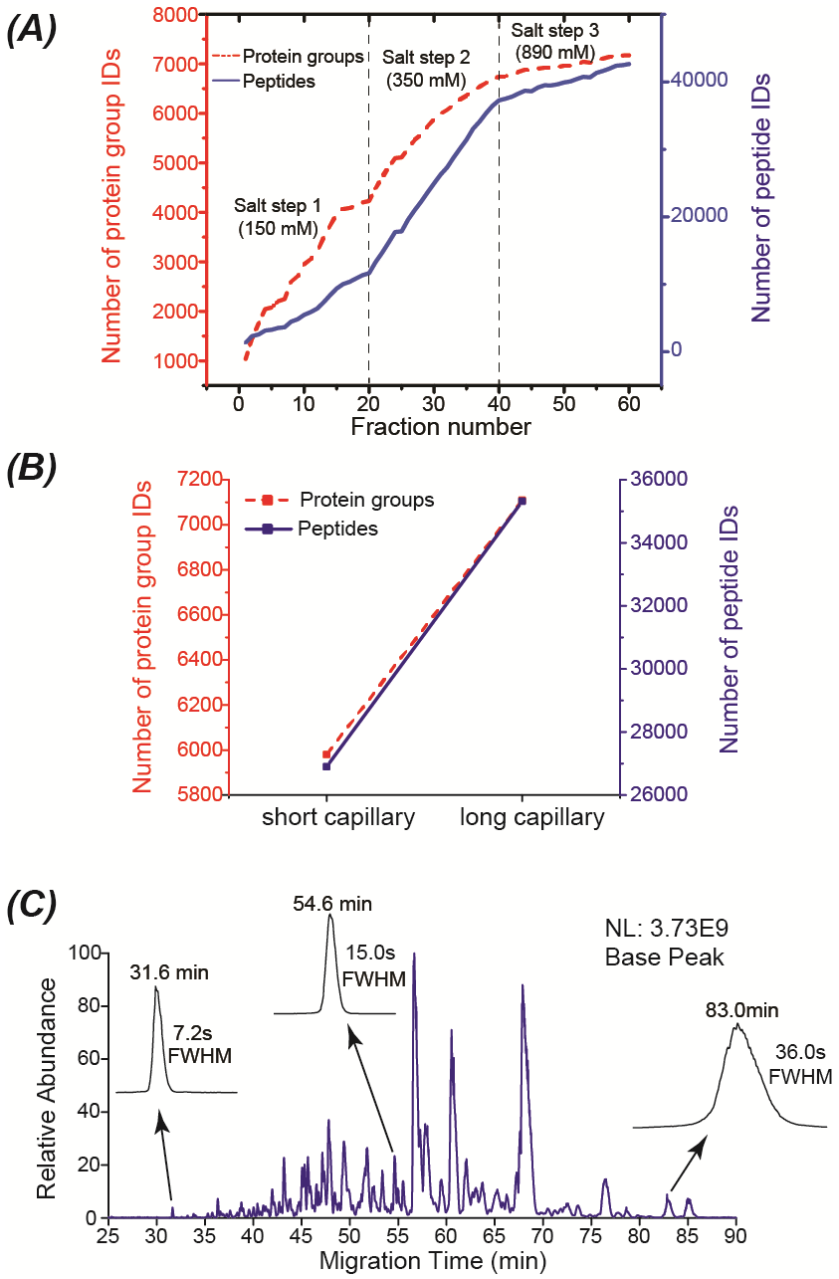

Figure 3.2. Summary of the results from the mouse brain proteome digest using SCX-RPLC-CZE-MS/MS. Three salt steps were employed for step-wise elution of peptides from the SCX to the RPLC. (A) The accumulated numbers of protein group and unique peptide IDs *vs.* the number of fractions. A 71-cm separation capillary was used for CZE-MS/MS. (B) Comparison of the number of protein group and unique peptide IDs from the twenty LC fractions corresponding to the second salt step analyzed by the CZE-MS/MS with a 71-cm separation capillary (short) or a 92-cm separation capillary (long). (C) An electropherogram of one SCX-RPLC fraction analyzed by CZE-MS/MS with the 92-cm separation capillary. The migration time and the full width at half maximum (FWHM) of three peptides were shown in the figure.

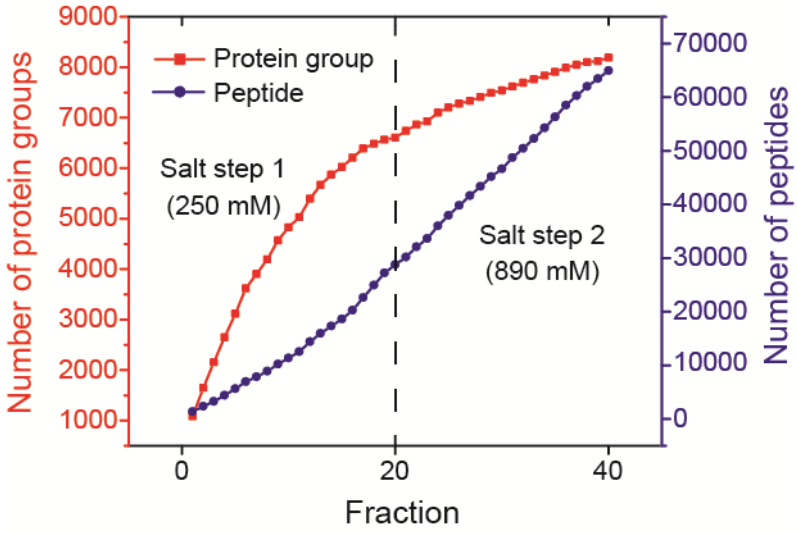

Figure 3.3. The accumulated numbers of protein group and unique peptide IDs from the mouse brain proteome digest *vs.* the number of SCX-RPLC fractions. Two salt steps were employed for step-wise elution of peptides from the SCX to the RPLC. CZE-MS/MS with a 94-cm separation capillary was used for analysis of the 40 SCX-RPLC fractions.

We attributed the large numbers of protein and peptide IDs from the experiment to two main reasons. First, the CZE-MS/MS system was capable of loading over 10% of the analytes in each LC fraction for analysis based on the dynamic pH junction stacking, 500 nL injection volume *vs.* 4 µL total sample volume. The large sample loading volume guaranteed the identification of low abundant proteins in the sample. Second, the SCX, RPLC, and CZE are orthogonal for separation of peptides based on their charge, hydrophobicity, and size-to-charge ratio. The orthogonal three-dimensional separation platform produced high peak capacity. We chose five CZE-MS runs and calculated their peak capacity based on five randomly chosen peptides with medium abundance. The peak capacity per CZE-MS run ranged from 175 to 250 based on the FWHM of those five peptides. Therefore, we estimated the overall peak capacity of the SCX-RPLC-CZE platform as at least 7,000 (175 × 40 fractions), representing the highest peak capacity of the CZE based platform until now for separation of a complex proteome digest.

The SCX-RPLC-CZE-MS/MS system combined the advantages of SCX, RPLC, and CZE-MS/MS. SCX has high sample loading capacity; RPLC can desalt the peptides and provide high-resolution separation of peptides; CZE can easily approach high separation efficiency for peptides and CZE-MS/MS can provide highly sensitive identification of peptides. [8,10,13,14]. The whole platform is straightforward and no sample cleanup is required between SCX-RPLC fractionation and CZE-MS/MS. In addition, the μ_{ef} of peptides in CZE has been predicted accurately using a simple model based on the size (molecular mass) and charge of peptides [19], which is invaluable for evaluating the confidence of peptide ID from the database search and even guiding the database search. The large-scale proteomic dataset from CZE-MS/MS presented in this work will be very useful for further evaluating and improving the model for prediction of μ_{ef} of peptides [19].

3.1.3.2. Comparison of SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS for deep sequencing of the mouse brain proteome

Much effort has been made for comparing CZE-MS/MS and RPLC-MS/MS for bottom-up proteomics, and the results clearly showed the good complementarity of those two methods for protein/peptide ID from complex proteomes [7-11,13,18-20,35]. In general, CZE-MS/MS tended to identify small, basic and hydrophilic peptides compared with RPLC-MS/MS, most likely due to the relatively weak retention of those peptides on RPLC column. However, the highest number of protein and peptide IDs using CZE-MS/MS in those previous works was only about 4,000 and 20,000, respectively. It is still not clear whether the good complementarity between CZE- and RPLC-MS/MS in protein/peptide ID still exists or not for dramatically larger proteomic datasets.

Here we further employed 2D-LC-MS/MS (high pH RPLC-low pH RPLC) for deep sequencing of the mouse brain proteome, resulting in the identification of 8,900 protein groups and 70,000 unique peptides in 60 h of mass spectrometer time. The data represents the capability of the state-of-the-art 2D-LC-MS/MS for deep sequencing of complex proteomes. Our SCX-RPLC-CZE-MS/MS identified 8,200 protein groups and 65,000 unique peptides in 70 h using the same amount of peptides as the starting material. This is the first time that CZE-MS/MS showed its capability to approach comparable performance to the state-of-the-art 2D-LC-MS/MS for deep proteomic sequencing.

We then compared the SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS on the scale of 8,000 protein groups and 65,000 peptides. The two techniques had good complementarity at both protein and peptide levels. Combination of both techniques improved the number of protein group ID to over 9700, which was nearly 10% higher than that from 2D-LC-MS/MS alone. The two techniques had even more significant complementarity at the peptide level, **Figure 3.4A**. Combining the data from those two methods resulted in the identification of over 100,000 unique peptides, which was over 40% higher than that from 2D-LC-MS/MS alone. The median sequence coverage of the overlapped proteins between CZE-MS/MS and LC-MS/MS (~7,000 proteins) was ~22% based on the LC-MS/MS data alone, and it was boosted to ~30% by combining CZE-MS/MS and LC-MS/MS data. The data clearly indicate that combination of the SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS can significantly improve the sequence coverages of identified proteins.

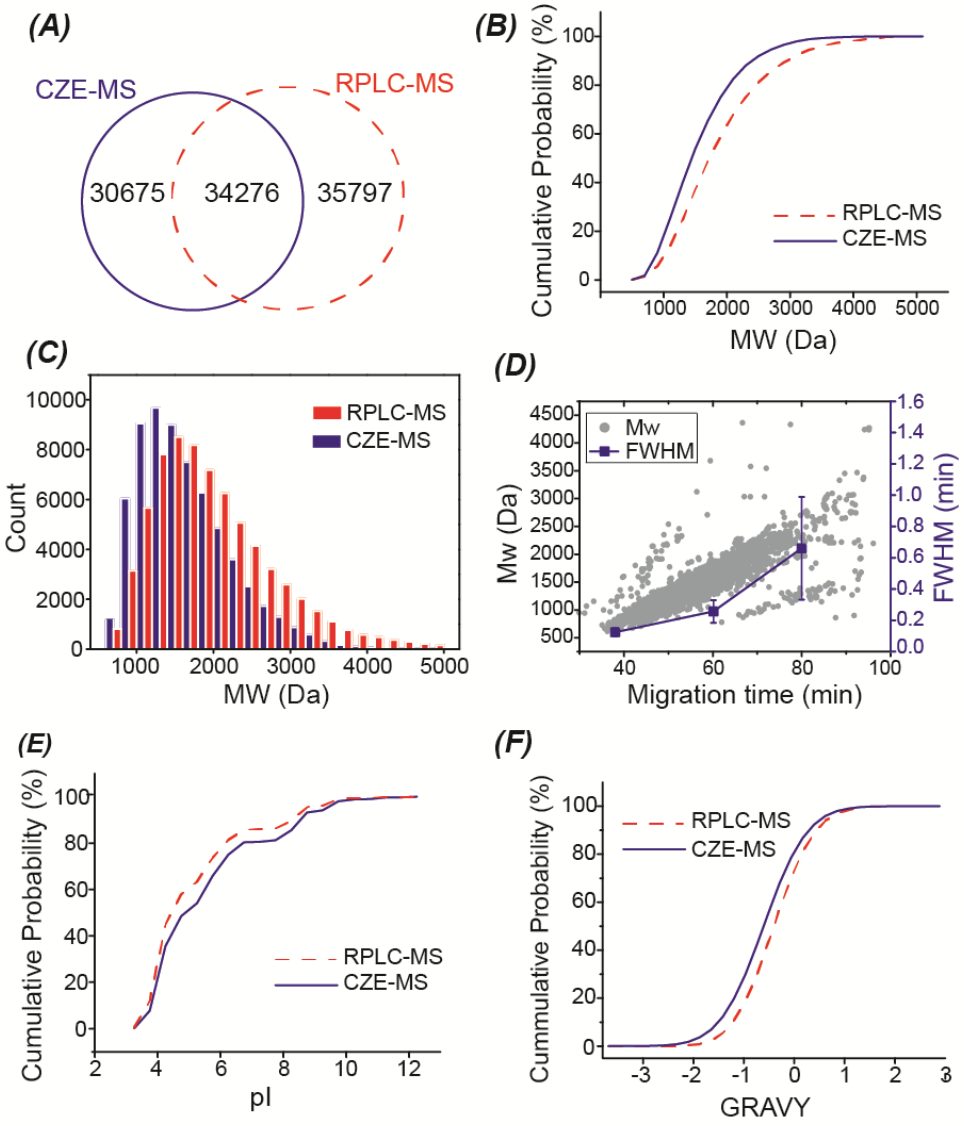

Figure 3.4. Comparison of SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS in terms of the identified peptides from the mouse brain proteome digest. (A) Overlap of identified peptides. (B) Cumulative distribution of molecular weight (MW) of identified peptides. (C) Bar graph of the MW distribution of the identified peptides. (D) Correlation between migration time and MW, migration time and FWHM of identified peptides from one random CZE-MS run. The FWHM of peptides at the three different migration time were calculated based on five randomly chosen peptides. The mean and the standard deviations of the FWHM of those five peptides were shown in the figure. (E) Cumulative distribution of the isoelectric point (pI) of identified peptides. (F) Cumulative distribution of the grand average of hydropathy (GRAVY) value of the identified peptides. Negative GRAVY values indicate hydrophilic; Positive GRAVY values signify hydrophobic.

Next, we investigated the physicochemical properties of the identified peptides from the SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS. The SCX-RPLC-CZE-MS/MS tended to identify small peptides compared with 2D-LC-MS/MS, **Figure 3.4B** and **Figure 3.4C.** One reason is that those small peptides tend to have weak retention on RPLC column, and are most likely washed out during the sample loading step [8]. Another possible reason relates to CZE. As shown in **Figure 3.4D**, larger peptides tend to have slower migration in the CZE separation capillary; The peptides with longer migration time tend to have wider peaks due to more significant diffusion in the capillary. Therefore, the relatively large peptides tend to have obviously wider peaks than the small peptides in CZE, leading to a more significant overlap of peptide peaks and more serious ionization suppression. As shown in **Figure 3.4E** and **Figure 3.4F**, the SCX-RPLC-CZE-MS/MS also tended to identify basic peptides and hydrophilic peptides. Basic peptides have more positive charges in an acidic buffer than acidic peptides, and they are more hydrophilic. Hydrophilic peptides can not be captured and separated well on the RPLC column. The different prefractionation methods used in those two platforms (SCX-RPLC *vs.* high pH RPLC) might also contribute to the differences in peptide IDs. In summary, the SCX-RPLC-CZE-MS/MS tended to identify small, basic and hydrophilic peptides, which agreed well with the data in the literature [8,10,11,19,35].

We also compared the identified proteins using SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS in terms of their gene ontology (GO) information. When we performed the comparison based on all of the identified proteins, we observed that those two platforms agreed well in GO information of identified proteins, **Figure 3.5.** We further performed a biological process enrichment analysis of the genes that were uniquely identified by the SCX-RPLC-CZE-MS/MS (847 genes) or 2D-LC-MS/MS (1,476

genes). Surprisingly, we observed that those uniquely identified genes from those two platforms had dramatically different biological process enrichment profiles. The genes uniquely identified by the SCX-RPLC-CZE-MS/MS were enriched in potassium ion transmembrane transport, regulation of angiogenesis, bone development, covalent chromatin modification, and positive regulation of I-kappaB kinase/NF-kappaB signaling. The genes uniquely identified by 2D-LC-MS/MS were enriched in the regulation of gene expression, ribosome biogenesis, DNA methylation, nucleosome assembly, transcription, and rRNA processing. The results clearly suggest that the combination of SCX-RPLC-CZE-MS/MS and 2D-LC-MS/MS not only can boost the sequence coverage of proteins but also can improve our ability for more comprehensive characterization of biological processes in cells.

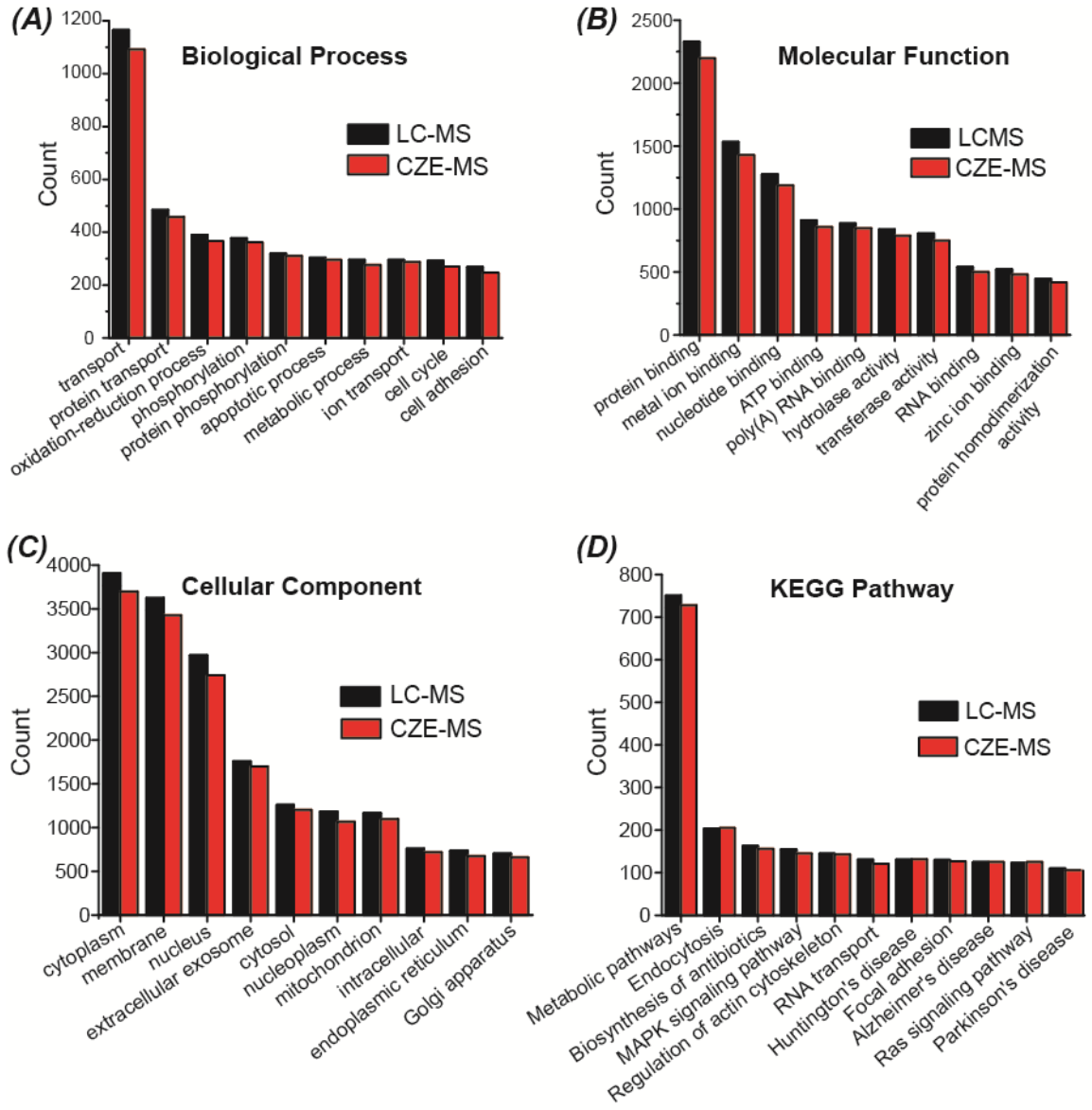

Figure 3.5. Gene Ontology (GO) information of identified proteins using the SCX-RPLC-CZE-MS/MS (CZE-MS) and 2D-LC-MS/MS (LC-MS). DAVID bioinformatics resources 6.8 (https://david.ncifcrf.gov/) was used to get the GO information of proteins. The GO terms were sorted by the number of proteins (Count). The top10 or top11 GO terms were used for the figure. (A) Biological process; (B) Molecular function; (C) Cellular component; (D) KEGG pathway.

3.1.4. Conclusions

In this work, for the first time, we established an SCX-RPLC-CZE-MS/MS platform for deep bottom-up proteomics, leading to the identification of 8,200 protein groups and 65,000 unique peptides from a mouse brain proteome digest. The data represents the largest bottom-up proteomics dataset using CZE-MS/MS. The orthogonal SCX-RPLC-CZE platform produced a high peak capacity of ~7,000, representing the highest peak capacity of the CZE based platform until now for separation of a complex proteome digest. The SCX-RPLC-CZE-MS/MS and the state-of-the-art 2D-LC-MS/MS showed good complementarity in protein and peptide IDs based on the comparisons performed on the scale of 8,000 proteins and 65,000 unique peptides.

We expect that the number of protein/peptide IDs from complex proteomes using the SCX-RPLC-CZE-MS/MS platform can be further significantly improved via simply increasing the number of SCX-RPLC fractions. In order to speed up the analysis of those large numbers of SCX-RPLC fractions, the sequential sample injection based CZE-MS/MS can be employed. [21,36,37].

3.1.5. Acknowledgments

We thank the Prof. Chen Chen's group at Department of Animal Science, Michigan State University for kindly providing the mouse brain samples for our research. The research was funded by Michigan State University.

3.2. SCX-RPLC-CZE platform for large-scale phosphoproteomics with the production of over 11000 phosphorylated peptides from the colon carcinoma HCT116 cell line

3.2.1. Introduction

Protein phosphorylation is a key reversible post-translational modification in nature, and it is involved in various cellular processes such as transcriptional and translational regulation, cellular signaling, metabolism, and cell differentiation [38]. Global site-specific characterization of protein phosphorylation allows us to gain insights into the regulatory role of phosphorylation in fundamental biological processes. MDLC-MS/MS (*e.g.,* SCX-RPLC-MS/MS) is routinely used for large-scale phosphoproteomics and it can identify over 10,000 phosphorylation events per study [39-47]. More than 50,000 distinct phosphorylated peptides have been reported from a single human cancer cell line using MDLC-MS/MS [40].

Based on statistical estimates, there are over half a million potential phosphorylation sites in the human proteome [40,48,49]. We need to boost the peptide separation to reach a deeper coverage of the human phosphoproteome. Since the proteomics community has made tremendous efforts in improving MDLC-MS/MS for phosphoproteomics in the last 20 years, we argue that an alternative separation method that is complementary to the LC for phosphorylated peptide separation will be very useful for deep phosphoproteomics.

CZE is a powerful method for the separation of biomolecules (*e.g.,* peptides and proteins) and it can have extremely high separation efficiency [13,14,50-54]. CZE-MS/MS has attracted great attention for proteomics recently because of the

improvements in the CE-MS interface [25,26,55,56], the sample stacking method [32,33,57], and the high-quality coating on the inner wall of the separation capillary [24].

CZE-MS/MS has some unique features for phosphoproteomics. First, CZE-MS/MS and RPLC-MS/MS can sample different pools of the phosphorylated peptides in cells due to the different separation mechanisms of CZE and RPLC (size-to-charge ratio *vs.* hydrophobicity) [9, 58,59]. The combination of these two methods can boost the phosphoproteome coverage significantly. Second, CZE can separate the phosphorylated and unphosphorylated forms of peptides due to their significant difference in charge. This feature reduces the interference of phosphorylated peptide ID from unphosphorylated peptides [9,58]. Third, the migration time of peptides in CZE can be predicted easily and accurately [19]. If we can generate a large phosphorylated peptide dataset using CZE-MS/MS, we can build a simple model to predict the migration time of phosphorylated peptides. This unique feature of CZE-MS/MS makes it a powerful tool for phosphoproteomics because the predicted migration time of phosphorylated peptides can be used to evaluate their ID confidence from a database search and even guide the database search.

Few papers have been published on using CZE-MS/MS for phosphoproteomics. We previously coupled CZE to a Q-Exactive mass spectrometer via an electro-kinetically pumped sheath flow CE-MS interface for phosphoproteomics of a human cell line [9]. 2,300 phosphorylated peptides were identified with single-shot CZE-MS/MS in 100 min, and the data suggested the high potential of CZE-MS/MS for large-scale phosphoproteomics. Recently, Faserl *et al.* investigated the sheathless CE-MS interface-based CZE-MS/MS for large-scale phosphoproteomics [58]. They identified over 5,000 phosphorylated peptides by coupling RPLC fractionation to the

CZE-MS/MS. To boost the number of phosphorylated peptide IDs using the CZE-MS/MS, the loading capacity and the separation window of CZE need to be improved. Recently, we developed a novel CZE-MS/MS system with a micro-liter scale sample loading volume and hours of separation window, opening the door to using CZE-MS/MS for large-scale proteomics [22,23]. The CZE-MS/MS system employed a 1-meter separation capillary with a high-quality neutral coating on its inner wall for eliminating the electroosmotic flow, an optimized dynamic pH junction method for highly efficient online stacking of peptides and proteins, the improved electro-kinetically pumped sheath flow CE-MS interface [25] and a Q-Exactive HF mass spectrometer.

We recently coupled SCX-RPLC fractionation to the CZE-MS/MS for deep proteomics of a mouse brain, leading to extremely high peak capacity for peptide separation and 8,200 protein IDs [60]. Motivated by the high peak capacity of the SCX-RPLC-CZE system for peptide separation, in this work, we applied the SCX-RPLC-CZE-MS/MS system for large-scale phosphoproteomics of HCT116 colon cancer cells. We had three goals in this work. First, boost the number of phosphorylated peptide IDs from a human cell line using CZE-MS/MS. The large phosphorylated peptide dataset will be useful for building a model for predicting the migration time of phosphorylated peptides. Second, we were interested in reaching a better understanding of CZE for the separation of phosphorylated and unphosphorylated forms of peptides. Third, we wished to investigate the difference between our CZE-MS/MS data and the literature LC-MS/MS data regarding the phosphosite motifs. We speculated that the good complementarity between CZE-MS/MS and RPLC-MS/MS for peptide IDs might result in significant differences in phosphosite motifs and found that the data supported our hypothesis.

3.2.2. Experimental

3.2.2.1. Materials and reagents

All reagents were bought from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. LC/MS grade water, FA, methanol, ACN, HPLC grade AA and HF were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide was obtained from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ).

3.2.2.2. Cell Growth Conditions

The human colon carcinoma cell line HCT 116 was obtained from American Type Culture Collection (ATCC). The cells were grown in RPMI 1640 cell culture medium (Life Technologies) supplemented with 10% fetal bovine serum (FBS) (Thermo Scientific). The provider assured authentication of the cell line by cytogenetic analysis. In addition, the cell line was validated by short tandem repeat (STR) analysis within the last two years.

3.2.2.3. Sample Preparation and phosphorylated peptide enrichment

A lysis buffer with 8 M urea with 75 mM NaCl, 50 mM Tris-HCl (pH 8.2), 10 mM sodium pyrophosphate, 1 mM PMSF, 1 mM Na_3VO_4, 1 mM NaF, 1 mM β-glycerophosphate, and 1 EDTA-free protease inhibitor cocktail was prepared. HCT116 colon cancer cells were cultured to 70% confluence followed by cell lysis with the lysis buffer. A small proportion of the cell lysates were subjected to Bicinchoninic acid assay for protein concentration measurement. Three mg of extracted protein was subjected to denaturation at 37 °C for 1 h, reduction with 5 mM

DTT at 37 °C for 1 h, and alkylation with 14 mM IAA for 30 min at room temperature. The alkylation was terminated by adding 5 mM DTT for 25 min. The sample was then diluted with 25 mM Tris-HCl buffer (pH 8.2) with 1 mM $CaCl_2$. Trypsin was added to the sample for overnight digestion at 37 °C. Phosphorylated peptides in the desalted digest were enriched with TiO_2 beads in 1:4 peptides to beads ratio based on the references [61,62]. After enrichment, the phosphorylated peptides were desalted, lyophilized and stored at -80 °C before use.

3.2.2.4. SCX-RPLC fractionation

An SCX-RPLC online fractionation was performed based on reference [60] with some minor modifications. Briefly, a 4.6 mm i.d. × 12.5 mm length SCX trap column (Zorbax 300SCX, Agilent Technologies) and a 2.1 mm i.d. × 150 mm length C18 RP column (Zorbax 300Extend-C18, Agilent Technologies) were connected directly for online 2D-LC fractionation. An Agilent Infinity II HPLC system was used for the experiment. 0.1% FA in water, 0.1% FA in ACN, and 890 mM ammonium acetate solution (pH = 2.88) were used as mobile phase A, B, C for separation, respectively. Mobile phase A and C were used for stepwise elution of peptides from the SCX column. Mobile phases A and B were used to generate a linear gradient for the RPLC separation of peptides.

Roughly 200-µg phosphorylated peptides were dissolved in mobile phase A and then loaded onto the SCX column with mobile phase A at a flow rate of 0.3 mL/min for 20 min. The phosphorylated peptides retained on the SCX column were eluted stepwise by two different concentrations of ammonium acetate solution: 150 mM and 890 mM. Then, each SCX eluate was captured on the RPLC column. RPLC gradient separation was performed at a 0.3 mL/min flow rate for 70 min: 0-5 min, 2%B; 5-7

min, 2-8% B; 7-47 min, 8-40% B; 47-49 min, 40-80%; 49-59 min, 80% B; 59-60 min, 80-2% B; 60-70 min, 2%B. 42 fractions were collected (1 fraction/ min) from 6 to 48 min for each salt step elution and the fractions were named based on the elution order. From fraction 2 to fraction 41, fractions were combined by the following rule: fraction N + fraction (N+20). Fraction 1 was combined with the mixture of fraction 2 and fraction 22, and fraction 42 was combined with the mixture of fraction 21 and fraction 41. In total, there were 40 fractions (20 fractions/salt step x 2 salt steps) collected, and they were lyophilized and stored at -80 °C before use.

3.2.2.5. CZE-MS/MS

An ECE-001 CE autosampler (CMP Scientific, Brooklyn, NY) and a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) were coupled with the third-generation electro-kinetically pumped sheath flow CE-MS interface (an EMASS-II CE-MS interface, CMP Scientific) [25]. A borosilicate glass capillary (1.0 mm o.d., 0.75 mm i.d.) was pulled with a Sutter P-1000 flaming/brown micropipette puller to make an electrospray emitter. The opening of the emitter was 20-40 µm.

A 95-cm long fused silica capillary (50 µm i.d., 360 µm o.d.) was used for CZE separation. The inner wall of the capillary was coated with LPA based on reference [63]. One end of the LPA coated capillary was etched with hydrofluoric acid based on reference [14] to reduce the outer diameter to less than 100 µm. The BGEfor CZE was 5% (v/v) AA (pH 2.4) and the sheath buffer for electrospray was 10% (v/v) methanol and 0.2% (v/v) FA in water. The etched end of the capillary was introduced into the electrospray emitter, and the distance between the etched end and the orifice of the emitter was ~300 µm. The distance between the emitter orifice and the

inlet of the mass spectrometer was ~2 mm. 2.2 kV voltage was applied for electrospray ionization.

The 40 LC fractions were redissolved in 5 µL of 50 mM NH_4HCO_3 (pH 8) for CZE-MS/MS. For sample injection, approximately 200 nL or 300 nL of each sample was injected into the capillary for analysis. Then, 30 kV voltage was applied at the injection end for 5400 seconds for CZE separation, followed by capillary flushing with the BGE for 900 seconds under a 5-psi pressure.

A Q-Exactive HF mass spectrometer was used in CZE-MS/MS. A DDA method was employed. The mass resolution was 60,000 (at m/z 200) for both full MS scans and MS/MS scans. The automatic gain control targets were set to 3E6 and 1E5 for full MS scans and MS/MS scans, respectively. For full MS scans, the maximum injection time was 50 ms with a scan range of 300 to 1500 *m/z*. For MS/MS scans, the maximum injection time was set to 110 ms. The top ten most abundant ions were sequentially isolated with a 2-m/z isolation window for fragmentation with 28% normalized collision energy. The dynamic exclusion was set to 40 s. Ions with charges higher than 1 and lower than 8 were selected for fragmentation.

3.2.2.6. Data analysis

Proteome Discoverer 2.2 software (Thermo Fisher Scientific) was used for data analysis. Sequest HT was used for the database search [63]. The human database (UP000005640) was downloaded from UniProt (http://www.uniprot.org/). All raw files were searched against both the forward database and a decoy (reverse) database to estimate the FDR [28]. Maximum two missed cleavage sites were allowed for peptide identification, and the peptide length was set to 6 to 144 amino acid residues. The mass tolerances of precursor and fragments were 20 ppm and

0.05 Da, respectively. Oxidation (methionine) and phosphorylation (serine, threonine, and tyrosine) were set as dynamic modifications. Acetylation at the protein N-terminal was chosen as a dynamic modification. Carbamidomethylation (cysteine) was set as a static modification. The peptide ID was filtered with confidence as high, corresponding to a 1% FDR. Protein grouping was enabled, and the strict parsimony principle was applied. The phosphoRS that integrated into the workflow was used to evaluate the confidence of the phosphosite localization [64].

MaxQuant 1.5.5.1 [65] was also used for the database search to compare phosphorylated peptide IDs and phosphosite motifs obtained from our CZE-MS/MS data with the literature data. The Andromeda search engine was used to search the MS/MS spectra [66]. The same human database used in the Proteome Discoverer search was used. The peptide mass tolerances of the first search and main search were 20 and 4.5 ppm, respectively. The fragment ion mass tolerance was 20 ppm. Trypsin was selected as the protease. The variable and static modifications were the same as the Proteome Discoverer search. The minimum length of a peptide was set to 7. The FDRs were 1% for both peptide and protein IDs. For phosphorylated peptide identifications, the phosphosite localization probability should be better than 0.75.

An online available GRAVY calculator (http://www.gravy-calculator.de/) was used to calculate the grand average of hydropathy (GRAVY) values of peptides. Online version of SSRCalc (http://hs2.proteome.ca/SSRCalc/SSRCalcX.html) was used to calculate the hydrophobicity indexes for peptides [67]. Molecular weights and isoelectric points of identified peptides were calculated using the “Compute pI/Mw” tool in ExPASy (http://web.expasy.org/compute_pi/). Motif-x (http://motif-x.med.harvard.edu/motif-x.html) was used to extract motifs from the data sets,

default settings were used except MS/MS was chosen as foreground format, and the human proteome was chosen as the organism [68]. Motif alignment was performed with WebLogo3 (http://weblogo.threeplusone.com/create.cgi).

3.2.2.7. Observed and predicted electrophoretic mobility of peptides

The data from LC fraction 8 was used for the μ_{ef} analysis. Only peptides having no variable modifications except for single phosphorylation were used for the analysis. Peptides' observed μ_{ef} (μ_{ef} observed) was determined using migration times (t_M, min) - time of MS/MS acquisition of the most intense tandem spectra for each unique peptide identification. We assumed that the electroosmotic flow (EOF) at 5% (v/v) acetic acid (pH 2.4) in the BGE was very low and mapped t_M into μ_{ef} using the equation for their experimental conditions (a 95 cm long capillary at 280 volts/cm):

μ_{ef} observed = $\frac{95}{65 \times t_m \times 280}$ (units of $cm^2 * V^{-1} * s^{-1}$)

Sequence-Specific Retention Calculator (SSRCalc) CZE model reported previously was used to predict the μ_{ef} of peptides [67]. While peptide charge and mass are the main parameters in determining mobility value, we introduced corrections for several sequence-specific features affecting corrected charge value (*Zc*) applied for calculations:

μ_{ef} predicted = $3.069 + 386 \times \ln\left(\frac{1+0.35 \times Zc}{Mc^{0.411}}\right) + OFFSET(\frac{Zc}{N})$

where 3.069 and 386 are empirical coefficients applied to align modeling output with experimentally measured values; *Zc* – peptide charge at pH 2.4, corrected using thirteen residue and sequence specific coefficients; *Mc* = (0.66**M* + 0.34**N**110.9), corrected mass to reflect the influence of different amino acid size; *M* is the molecular weight of peptides; *N* is the peptide length; OFFSET is a polynomial empirical function of *Zc*/*N* to correct prediction for peptides with extremely high and low mobility values.

3.2.3. Results and discussion

As shown in **Figure 3.6A**, 3 mg of HCT-116 cell proteins were digested into peptides with trypsin, followed by phosphorylated peptide enrichment using TiO_2 beads based on references [61,62]. The phosphorylated peptides were fractionated with online SCX-RPLC into 40 fractions based on the charge and hydrophobicity of phosphorylated peptides. Each LC fraction was analyzed by dynamic pH junction based CZE-MS/MS [22], and CZE separates peptides based on their size-to-charge ratios. The SCX, RPLC, and CZE are orthogonal for peptide separation. As shown in **Figure 3.6B**, a 2-min RPLC eluate was further separated by CZE into a 50-min window. As shown in **Figure 3.6C**, the correlation between m/z and migration time of peptides from the database search is complicated but, in general, peptides with higher *m/z* tend to migrate slower in the capillary during the CZE separation.

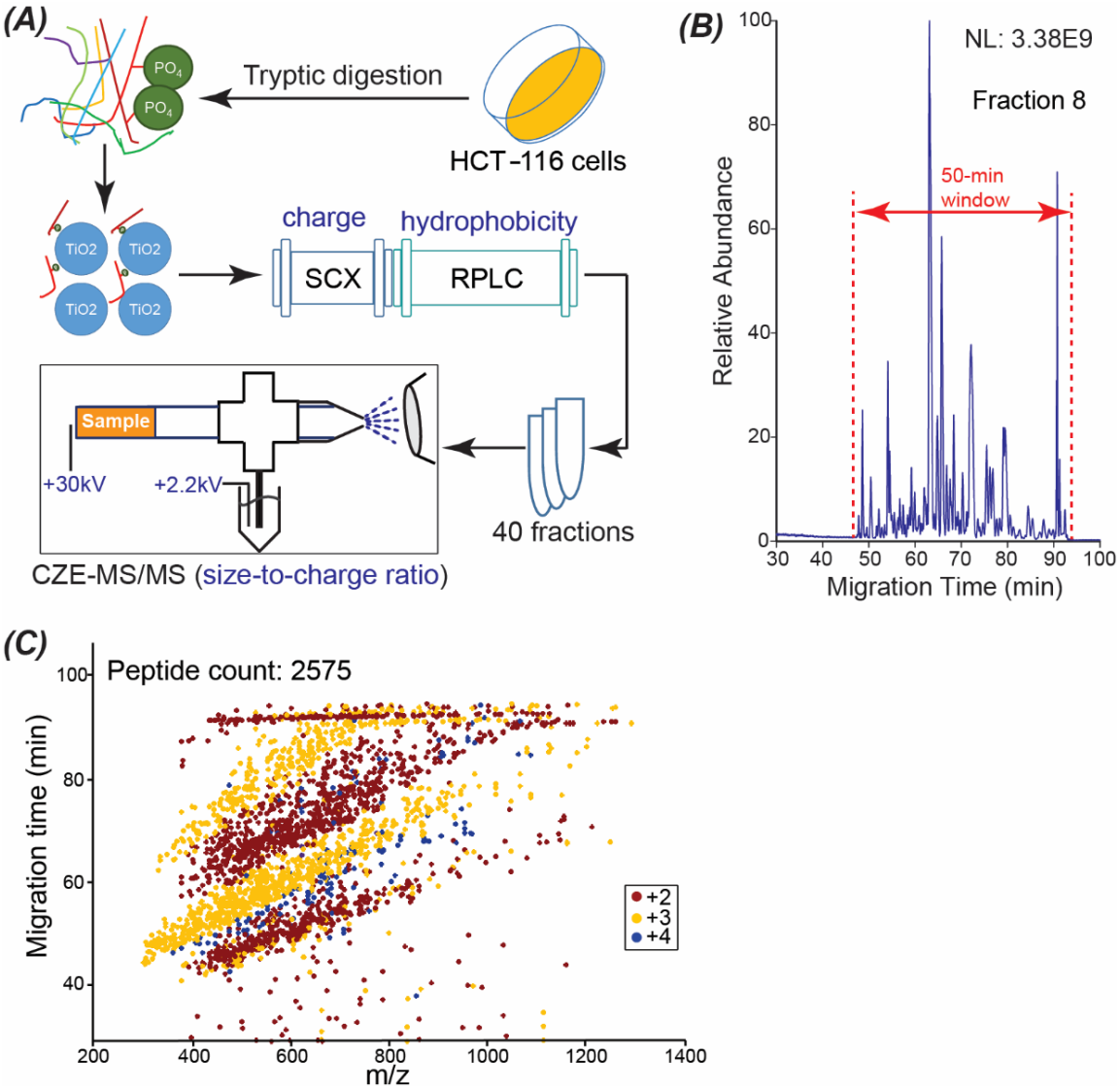

Figure 3.6. (A) The experimental design of the work. (B) Base peak electropherogram of one RPLC fraction (fraction 8) after CZE-MS/MS analysis. (C) Mass-to-charge ratio (m/z) *vs.* migration time of peptides identified by CZE-MS/MS from the RPLC fraction 8.

3.2.3.1. Large-scale phosphoproteomics of the HCT-116 cell line using SCX-RPLC-CZE-MS/MS

CZE-MS/MS analyses of the 40 SCX-RPLC fractions produced 6,502 protein IDs, 33,301 peptide IDs, and 11,555 phosphopeptides with a peptide-level 1% FDR. Proteome discoverer 2.2 was used for the peptide and protein IDs. 10,029 phosphorylated peptides were identified with phosphosite localization probability better than 95%. To our knowledge, our phosphorylated peptide dataset represents the largest phosphoproteomics data so far using CZE-MS/MS. In the literature, we reached 2,300 phosphorylated peptide IDs with single-shot CZE-MS/MS in 100 min [9] and Faserl *et al.* approached over 5,000 phosphorylated peptide IDs using an RPLC-CZE-MS/MS system in about 60 hours [58]. In this work, we identified 11,555 phosphorylated peptides using the SCX-RPLC-CZE-MS/MS in 67 hours. All three studies employed the Proteome Discoverer platform for data analysis. Our system improved the number of phosphorylated peptide IDs by 100% compared with Faserl's work with a comparable instrument time. We noted that the phosphorylated peptide identification efficiency decreased drastically from the single-shot CZE-MS/MS data (1400 phosphorylated peptides/hour) [9] to the RPLC-CZE-MS/MS data (90 phosphorylated peptides/hour) [58] and our SCX-RPLC-CZE-MS/MS data (170 phosphorylated peptides/hour).

We also noted that the specificity of our TiO_2 enrichment was low (about 35%) regarding the ratio between phosphorylated peptide IDs and total peptide IDs. We believe the number of phosphorylated peptide IDs can be improved significantly with a better phosphorylated peptide enrichment procedure. The large-scale phosphorylated peptide dataset produced in this work will be useful for building a simple model for accurate prediction of migration time of phosphorylated peptides

[19]. The accurately predicted migration time of phosphorylated peptides could be used to evaluate the confidence of their IDs from a database search and even further guide the database search.

The SCX with two salt step elution (150 mM (salt step 1) and 890 mM (salt step 2) ammonium acetate solution, pH = 2.88) separated the phosphorylated peptides well, and only 618 out of the 11,555 phosphorylated peptides were overlapped between those two salt steps, **Figure 3.7A**. Phosphorylated peptides in salt step 2 tend to have higher charge states (**Figure 3.7B**), have higher molecular weights (**Figure 3.7C**), and be more basic (**Figure 3.7D**) compared with that in salt step 1. The number of phosphorylated peptide IDs per LC fraction ranges from 300 to 700 for most of the fractions, and the distribution is moderately uniform, **Figure 3.7E**. In CZE, phosphorylated peptides tend to migrate significantly slower than unphosphorylated peptides in the separation capillary, **Figure 3.7F**. This feature makes CZE-MS/MS useful for phosphoproteomics because the interference of phosphorylated peptide IDs from unphosphorylated peptides can be reduced.

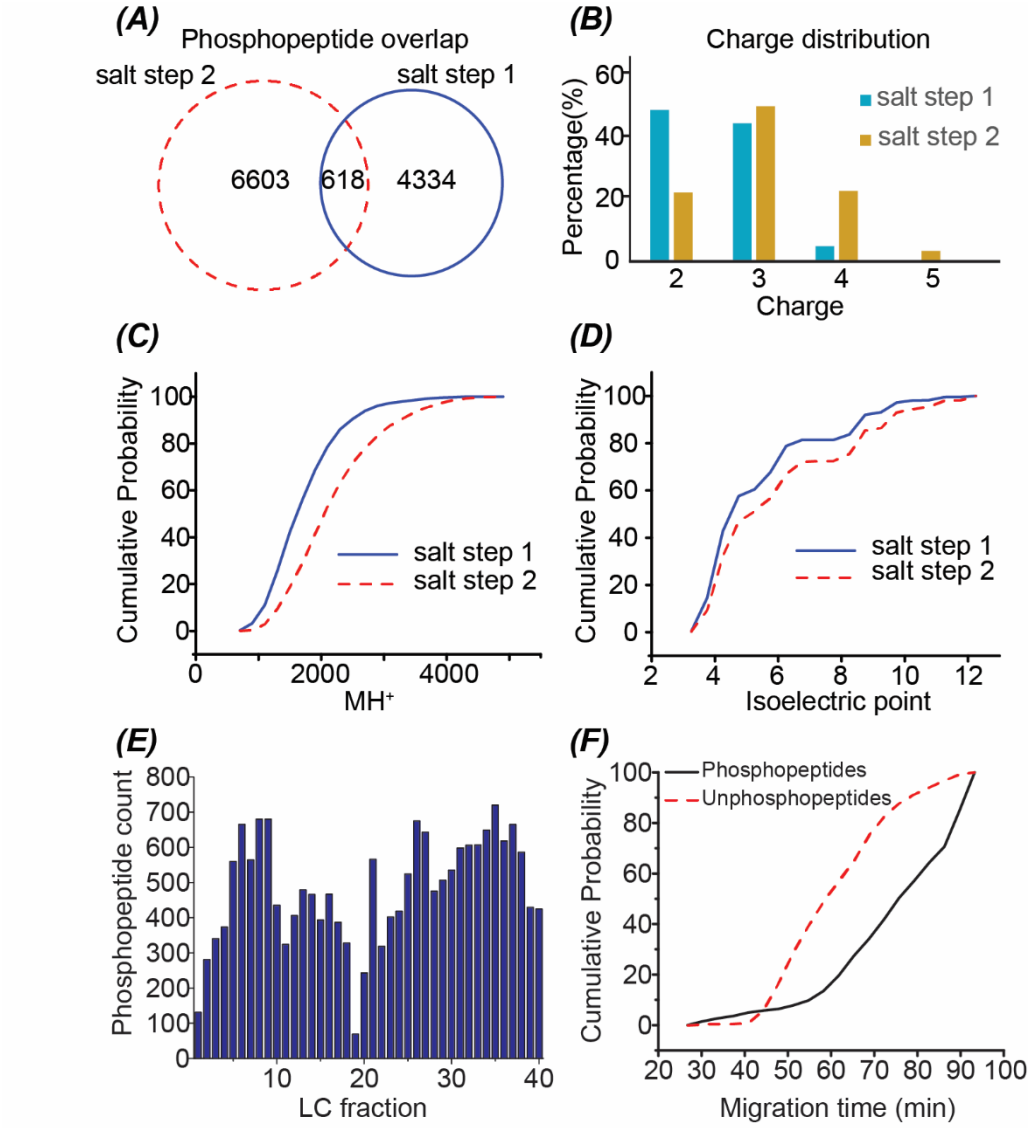

Figure 3.7. Summary of the phosphorylated peptide IDs using the SCX-RPLC-CZE-MS/MS. (A) Overlap of the identified phosphorylated peptides from the two salt steps of the SCX. Salt step 1 and 2 used 150 mM and 890 mM ammonium acetate solution (pH = 2.88) for peptide elution, respectively. (B) The charge distribution of identified phosphorylated peptides in the two salt steps. (C) Cumulative distribution of mass of identified phosphorylated peptides in the two salt steps. (D) Cumulative distribution of pI of identified phosphorylated peptides in the two salt steps. The pI was calculated based on the peptide sequence. (E) The number of phosphorylated peptide IDs across the 40 LC fractions. (F) Cumulative distribution of migration time of identified phosphorylated peptides and unphosphorylated peptides in one LC fraction (fraction 8).

3.2.3.2. Investigating the effect of phosphorylation on electrophoretic mobility of peptides

Phosphopeptides tend to migrate significantly slower than their unphosphorylated forms under acidic conditions used for CZE separations and in normal polarity. The addition of one phosphoryl group reduces the overall positive charge of peptides by one charge unit, thus resulting in a drastic decrease in μ_{ef}.

As shown in **Figure 3.8A** and **3.8B**, the phosphorylated forms of peptides QGGGGGGGSVPGIER and AGELTEDEVER migrate much slower than their unphosphorylated forms and their migration time difference (Δ time) is about 20 min. We noted that the Δ time should be larger than 20 min because we started to flush the capillary by applying a 5-psi pressure at 90 min. As shown in **Figure 3.8C**, the doubly phosphorylated form of the peptide AAKLSEGSQPAEEEEDQETPSR migrates slower than the singly phosphorylated form due to the one more negative charge. Their Δ time should be much larger than 3 min because they were both pushed out of the capillary by the pressure.

We manually analyzed the data from six LC fractions regarding the Δ time between unphosphorylated and singly phosphorylated forms of peptides. We obtained 200 pairs of peptides and their altered migration (Δ) time in CZE. As shown in **Figure 3.8D**, for all the 200 pairs of peptides, the singly phosphorylated forms migrate slower than the corresponding unphosphorylated forms, which is demonstrated by the positive Δ time values. For about 70% of the peptide pairs, the Δ time ranges from 10 to 30 min. We reached two conclusions here. First, for the majority of peptides studied, the addition of one phosphoryl group onto the peptide can drastically slow down its migration in the capillary during CZE. Second, adding

one phosphoryl group to different peptides influences their migration to various extents.

We further investigated the effect of phosphorylation on μ_{ef} of peptides by comparing the observed and predicted mobility values of phosphopeptides and unphosphopeptides. We used the data from one LC fraction (fraction 8) for this task. The observed and predicted μ_{ef} of peptides were calculated using the methods described in the "Experimental section". Application of non-modified SSRCalc CZE model (without considering the effect of negatively charged phosphoryl groups) was used to illustrate the effect of phosphorylation, **Figure 3.8E**. Mobility of unphosphopeptides follows SSRCalc prediction (R^2 =0.99) [19], whereas addition of one phosphoryl group decreases mobility dramatically, **Figure 3.8E**. After removing all the phosphopeptides (+2) and some of the phosphopeptides (+3) with mobility lower than 6.2×10^{-5} $cm^2*V^{-1}*s^{-1}$, we obtained reasonably good linear correlations between observed and predicted μ_{ef} values within each group of phosphopeptides ($R^2 \geq 0.94$), **Figure 3.8E**. We noted that the observed mobilities of peptides were obviously lower than their predicted mobility. We attributed the phenomenon to the dynamic pH junction sample stacking method used in the CZE experiments, which slowed down the mobility of peptides in the capillary.

First attempts have been made to adapt SSRCalc CZE model to prediction of phosphopeptides' mobility values. The corrected charge (*Zc)* of phosphopeptides has been modified to improve correlation for the entire set of peptides shown in **Figure 3.8E** (phosphopeptides and unphosphopeptides). We found that *Zc* values had to be adjusted by -0.91 and by -1.0 for +5/+4 and +3 phosphopeptides, respectively. **Figure 3.8F** shows prediction accuracy (R^2 ~0.99) for combined set of peptides, identical to the collection of unphosphopeptides in **Figure 3.8E**. This data

indicate that the charge shift is indeed very close to the expected contribution from one phosphoryl group. We need to note that much larger phosphopeptide datasets with confident assignments of modification site are needed for the development of the sequence-dependent model for mobility prediction and for a better understanding of how phosphorylation influences μ_{ef} of peptides. Similar to the effect of acidic Asp/Glu residues reported before [19], we anticipate that N-terminal positioning of phosphate (or in close proximity to other positively charged groups) will result in a larger decrease in mobility.

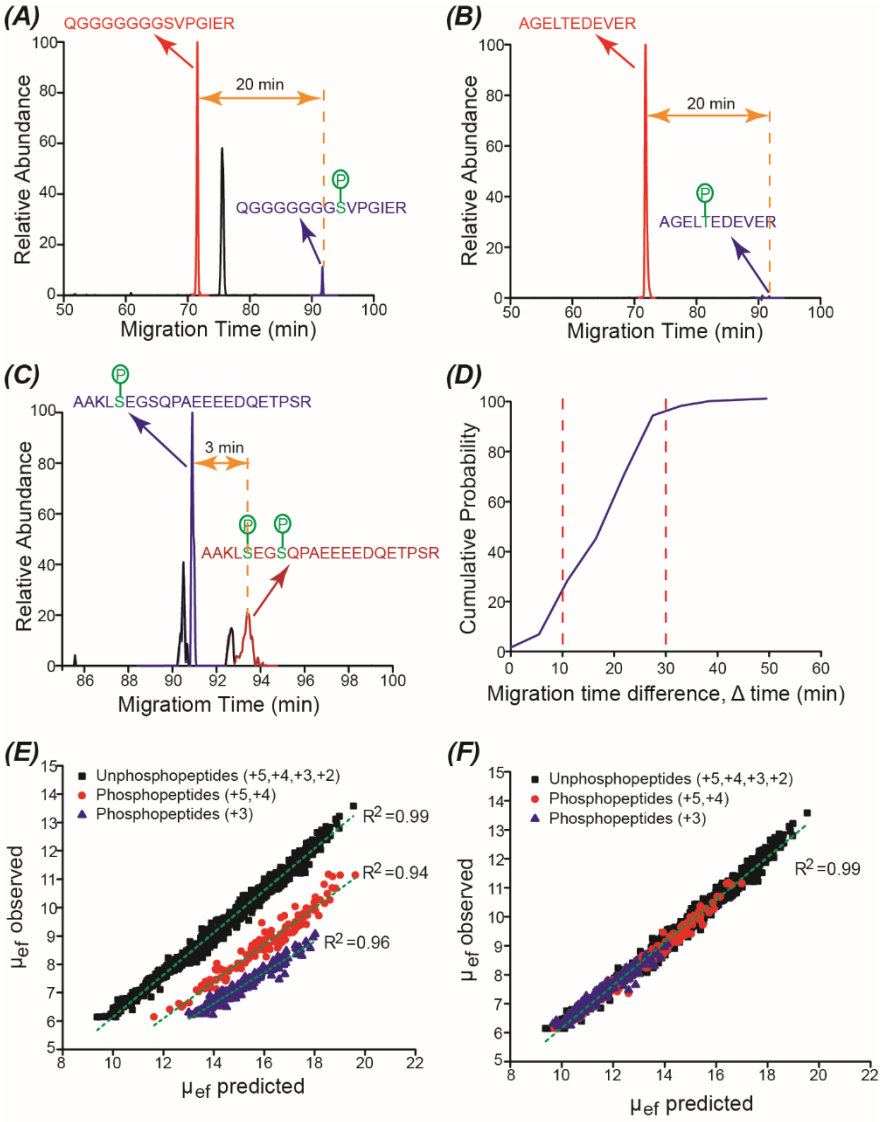

Figure 3.8. (A) Extracted ion electropherogram (EIE) of phosphorylated and unphosphorylated forms of the peptide QGGGGGGGSVPGIER. (B) EIE of phosphorylated and unphosphorylated forms of the peptide AGELTEDEVER. (C) EIE of singly phosphorylated and doubly phosphorylated forms of the peptide AAKLSEGSQPAEEEEDQETPSR. (D) Cumulative distribution of the migration time difference (Δ time) between unphosphorylated and singly phosphorylated forms of peptides. The figure was based on the data from six LC fractions. (E) Correlations between observed and predicted electrophoretic mobility (μ_{ef}) of unphosphopeptides and phosphopeptides with one phosphoryl group. The non-modified SSRCalc CZE model [31] was used to highlight the effect of phosphorylation. (F) Correlation between observed and predicted μ_{ef} of peptides using the modified SSRCalc CZE model. μ_{ef} x 10^5 ($cm^2*V^{-1}*s^{-1}$) is shown in (E) and (F). The peptides' charges in (E) and (F) are shown for non-modified peptide sequences (counting the number of lysine, arginine, and histidine residues, plus positively charged N-terminus).

3.2.3.3. Comparing our phosphoproteome dataset from CZE-MS/MS with an LC-MS/MS dataset in literature

Recently, Kubiniok *et al.* performed deep phosphoproteomics of HCT116 cells using TiO_2 enrichment, SCX-RPLC-MS/MS and MaxQuant software for data analysis [69]. We compared the HCT116 phosphoproteomics datasets from our SCX-RPLC-CZE-MS/MS with Kubiniok's SCX-RPLC-MS/MS. In order to make a fair comparison, we reanalyzed our data with MaxQuant software and filtered the data with the same criteria as Kubiniok *et al.* 6,221 phosphopeptides were identified using MaxQuant software, and only 45% of these phosphopeptides were covered by that identified in the Kubiniok's work, suggesting good complementarity between those two platforms for phosphopeptide IDs. The result here agrees well with the data in the literature that CZE-MS/MS and RPLC-MS/MS are well complementary for peptide and phosphopeptide IDs [9,58,60]. Further analyses of the physicochemical properties of identified phosphopeptides demonstrated that CZE-MS/MS tended to identify basic, small and hydrophilic phosphopeptides compared with LC-MS/MS **Figure 3.9**; these data agree with reports in the literature [8,35,60]. The data highlights that CZE-MS/MS can make a significant contribution to phosphoproteomics by improving the phosphoproteome coverage.

We further analyzed the phosphopeptides exclusively identified in our work or Kubiniok's work regarding the phosphosite motifs using the Motif-x, **Figure 3.10**. Interestingly, we observed significantly different motif logos between those two datasets for both phosphoserine and phosphothreonine. The corresponding phosphosites from the phosphopeptides exclusively identified in our work tend to be surrounded by acidic amino acids (glutamic acid and aspartic acid) after the phosphosites and basic amino acids (lysine and arginine) before the phosphosites

compared to that in Kubiniok's work, **Figure 3.10**. The data further highlights the value of CZE-MS/MS for phosphoproteomics for not only improving the phosphoproteome coverage but also providing more insight into the phosphosite motifs.

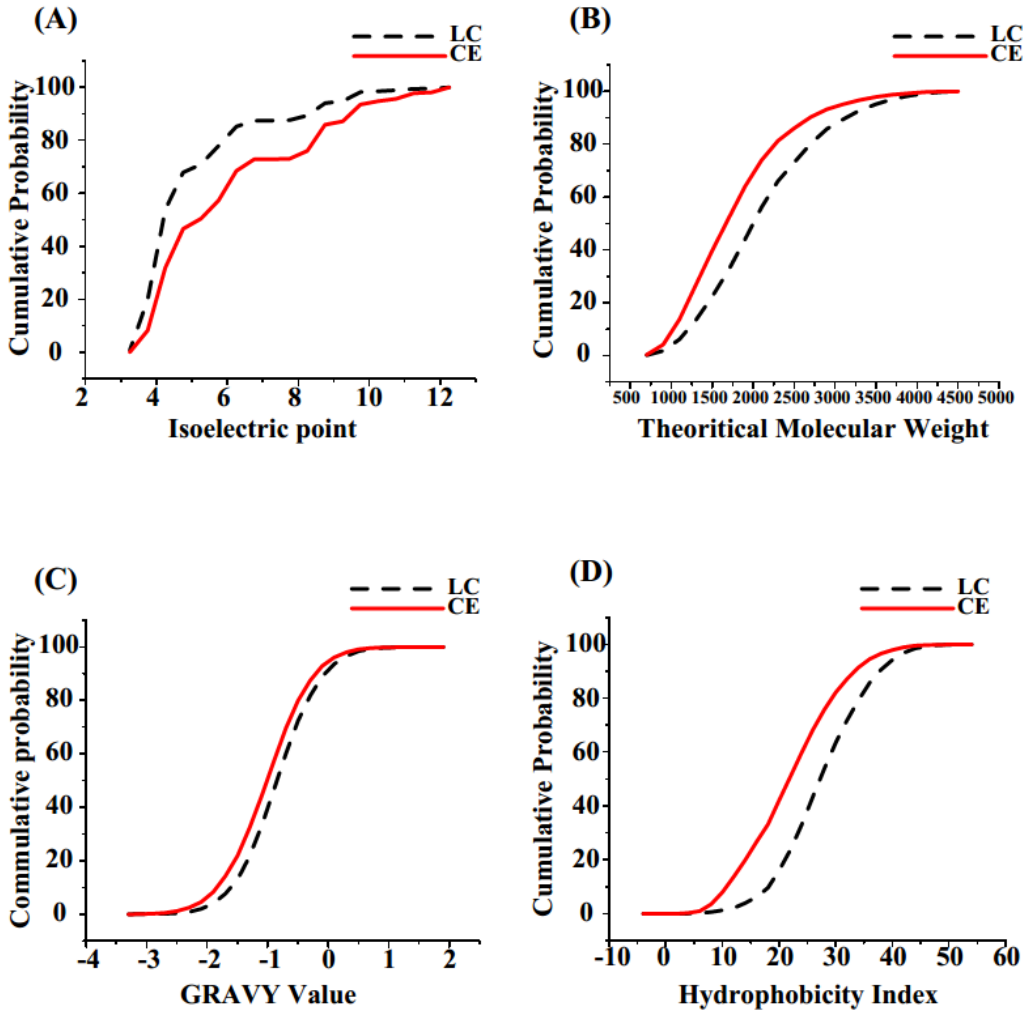

Figure 3.9. Physicochemical properties of phosphopeptides identified by the SCX-RPLC-CZE-MS/MS in this work (CE) and by SCX-RPLC-MS/MS in reference [69] (LC). Cumulative distributions of (A) isoelectric point, (B) theoretical molecular weight, (C) GRAVY value and (D) hydrophobicity index of peptides. For GRAVY value, negative values demonstrate hydrophilic peptides and positive values indicate hydrophobic peptides. For hydrophobicity index, a larger value indicates more hydrophobic.

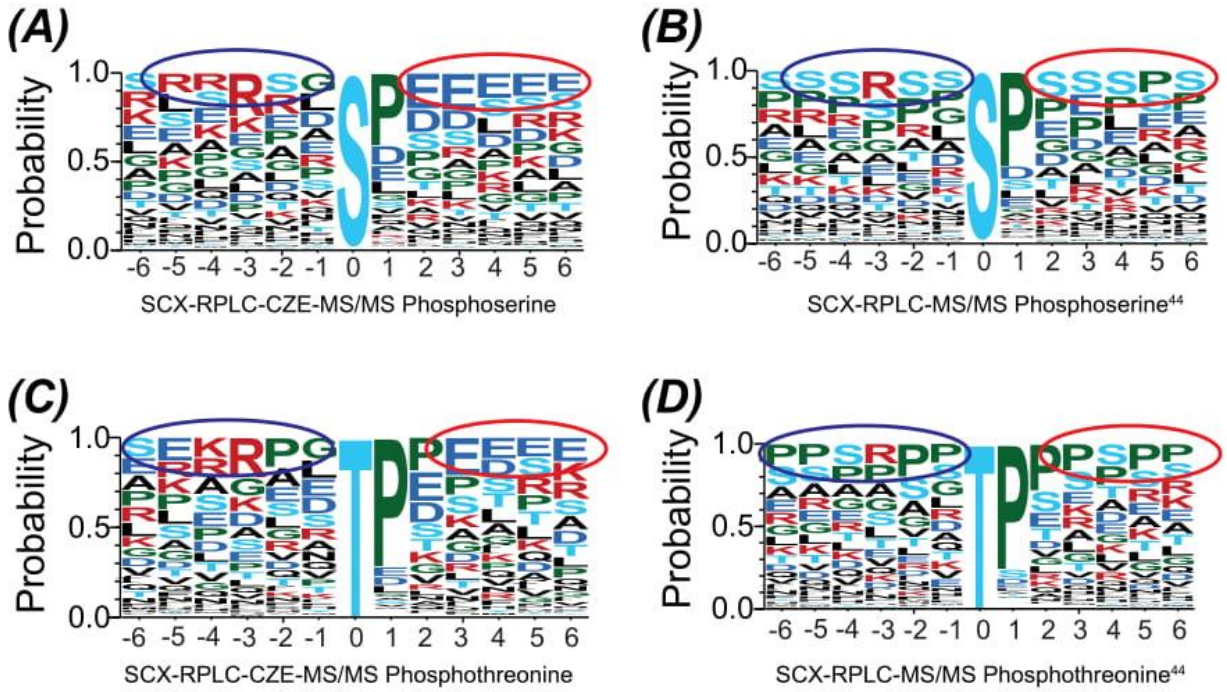

Figure 3.10. Summary of the phosphosite motif data from the SCX-RPLC-CZE-MS/MS in this work and from the SCX-RPLC-MS/MS in reference [69]. Motif-x (http://motif-x.med.harvard.edu/motif-x.html) was used to extract motifs from the data sets. Motif alignment was performed with WebLogo3 (http://weblogo.threeplusone.com/create.cgi). Motif logo of the phosphoserine (A) and phosphothreonine (C) based on the phosphorylated peptides exclusively identified in the SCX-RPLC-CZE-MS/MS data. Motif logo of the phosphoserine (B) and phosphothreonine (D) based on the phosphorylated peptides exclusively identified in the SCX-RPLC-MS/MS data.

3.2.4. Conclusions

An SCX-RPLC-CZE-MS/MS platform was employed for large-scale phosphoproteomics of the HCT116 cell line with the production of 11,555 phosphopeptide IDs. The dataset represents the largest phosphoproteome data so far using CZE-MS/MS. We are working on building a simple model based on the phosphoproteome dataset generated here for accurate prediction of phosphopeptide migration time in CZE. Our preliminary modeling attempts demonstrate that, similar to the unmodified tryptic peptides, the μ_{ef} of phosphopeptides can be accurately predicted. We expect that the predicted migration time of phosphopeptides will be useful to improve the confidence of phosphopeptide IDs from the database search and even guide the database search.

We expect that the number of phosphopeptide IDs from biological samples using CZE-MS/MS can be significantly boosted via several improvements. First, the phosphopeptide enrichment procedure can be dramatically improved. In this work, the specificity of phosphopeptide enrichment was only 35%. Over 80% and even 90% phosphopeptide enrichment specificity should be approachable with an optimized procedure based on the data in the literature [70,71]. Second, the separation system can be improved. Recently, we developed a high-resolution nanoflow RPLC-CZE-MS/MS system for deep and highly sensitive bottom-up proteomics with the production of 60,000 peptide IDs with only 5-µg of peptides as the starting material [72]. We expect significant improvements in both the number of phosphopeptide IDs and sensitivity will be achieved by using the nanoRPLC-CZE-MS/MS platform.

3.2.5. Acknowledgments

We thank the support from the Michigan State University and the National Institute of General Medical Sciences, National Institutes of Health (NIH), through Grant R01GM125991. Amanda B. Hummon thanks the support from the National Institutes of Health (R01GM110406), and the National Science Foundation (CAREER Award, CHE-1351595). Modeling studies were supported by grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05963 – Oleg Krokhin).

REFERENCES

REFERENCES

[1] Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins *Mol. Cell. Proteomics*, **2012**, 11, M111.014050

[2] Mertins, P. Qiao, J.W.; Patel, J.; D Udeshi, N.; Clauser, K.R.; Mani, D.R.; Burgess, M.W.; Gillette, M.A.; Jaffe, J.D.; Carr, S.A. Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat. Methods* **2013**, 10, pp. 634-637

[3] Ding, C.; Jiang, J.; Wei, J.; Liu, W.; Zhang, W.; Liu, M.; Fu, T.; Lu, T.; Song, L.; Ying, W.; Chang, C.; Zhang, Y.; Ma, J.; Wei, L.; Malovannaya, A.; Jia, L.; Zhen, B.; Wang, Y.; He, F.; Qian, X.; Qin, J. A fast workflow for identification and quantification of proteomes. *Mol. Cell. Proteomics* **2013**, 12, pp. 2370-2380

[4] Kelstrup, C.D.; Jersie-Christensen, R.R.; Batth, T.S.; Arrey, T.N.; Kuehn, A.; Kellmann, M.; Olsen, J.V. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J. Proteome Res.* **2014** 13, pp. 6187-6195.

[5] Zhao, Q.; Fang, F.; Shan, Y.; Sui, Z.; Zhao, B.; Liang, Z.; Zhang, L.; Zhang, Y. In-Depth Proteome Coverage by Improving Efficiency for Membrane Proteome Analysis. *Anal. Chem.* **2017**, 89(10), 5179-5185.

[6] Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, 509, 575-581.

[7] Sun, L.; Hebert, A. S.; Yan, X.; Zhao, Y.; Westphall, M. S.; Rush, M. J.; Zhu, G.; Champion, M. M.; Coon, J. J.; Dovichi, N. J. Over 10,000 peptide identifications from the HeLa proteome by using single-shot capillary zone electrophoresis combined with tandem mass spectrometry. *Angew Chem. Int. Ed. Engl.* **2014**, 8, 53(50), 13931-3.

[8] Faserl, K.; Sarg, B.; Kremser, L.; Lindner, H. Optimization and evaluation of a sheathless capillary electrophoresis-electrospray ionization mass spectrometry

platform for peptide analysis: comparison to liquid chromatography-electrospray ionization mass spectrometry. *Anal. Chem.* **2011**, 83, 7297-7305.

[9] Ludwig, K. R.; Sun, L.; Zhu, G.; Dovichi, N. J.; Hummon, A. B. Over 2300 phosphorylated peptide identifications with single-shot capillary zone electrophoresis-tandem mass spectrometry in a 100 min separation. *Anal. Chem.* **2015**, 87, 9532-9537.

[10] Wang, Y.; Fonslow, B. R.; Wong, C. C.; Nakorchevsky, A.; Yates, J. R.; 3rd. Improving the comprehensiveness and sensitivity of sheathless capillary electrophoresis-tandem mass spectrometry for proteomic analysis. *Anal. Chem.* **2012**, 84, 8505-8513.

[11] Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J. Single-shot proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with production of more than 1250 Escherichia coli peptide identifications in a 50 min separation. *Anal. Chem.* **2013**, 85, 2569-2573.

[12] Sun, L.; Zhu, G.; Yan, X.; Zhang, Z.; Wojcik, R.; Champion, M. M.; Dovichi N. J. Capillary zone electrophoresis for bottom-up analysis of complex proteomes. *Proteomics* **2016**, 16(2), 188-96.

[13] Lombard-Banek, C.; Moody, S. A.; Nemes, P. Single-Cell Mass Spectrometry for Discovery Proteomics: Quantifying Translational Cell Heterogeneity in the 16-Cell Frog (Xenopus) Embryo. *Angew. Chem. Int. Ed.* **2016**, 55, 2454-2458.

[14] 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.

[15] Faserl, K.; Kremser, L.; Müller, M.; Teis, D.; Lindner, H. H. Quantitative proteomics using ultralow flow capillary electrophoresis-mass spectrometry. *Anal. Chem.* **2015**, 87, 4633-4640.

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

[17] Guo, X.; Fillmore, T. L.; Gao, Y.; Tang, K. Capillary Electrophoresis-Nanoelectrospray Ionization-Selected Reaction Monitoring Mass Spectrometry via a True Sheathless Metal-Coated Emitter Interface for Robust and High-Sensitivity Sample Quantification. *Anal. Chem.* **2016**, 88, 4418-4425.

[18] Zhang, Z.; Dovichi, N. J. Optimization of mass spectrometric parameters improve the identification performance of capillary zone electrophoresis for single-shot bottom-up proteomics analysis. *Anal. Chim. Acta* **2018**, 1001, 93-99.

[19] Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem.* **2017**, *89*, 2000-2008.

[20] Yan, X.; Sun, L.; Zhu, G.; Cox, O. F.; Dovichi, N. J. Over 4100 protein identifications from a Xenopus laevis fertilized egg digest using reversed-phase chromatographic prefractionation followed by capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry analysis. *Proteomics* **2016**, 16(23):2945-2952.

[21] Faserl, K.; Sarg, B.; Sola, L.; Lindner, H. H. Enhancing Proteomic Throughput in Capillary Electrophoresis-Mass Spectrometry by Sequential Sample Injection. *Proteomics* **2017**, 17(22).

[22] Chen, D.; Shen, X.; Sun, L. Capillary zone electrophoresis-mass spectrometry with microliter-scale loading capacity, 140 min separation window and high peak capacity for bottom-up proteomics. *Analyst.* **2017**, *142*, 2118-2127.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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*, 207-14.

[29] Huang, da. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 2009, 4(1), 44-57.

[30] Huang, da. W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009,** 37(1), 1-13.

[31] Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, 6(5), 359-62.

[32] Aebersold, R.; Morrison, H. D. Analysis of dilute peptide samples by capillary zone electrophoresis. *J. Chromatogr.* **1990**, *516*, 79-88.

[33] 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*, 1242-52.

[34] Imami, K.; Monton, M. R.; 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.

[35] Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J. Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry as an alternative proteomics platform to ultraperformance liquid chromatography-electrospray ionization-tandem mass spectrometry for samples of intermediate complexity. *Anal. Chem.* **2012**, 84, 1617-1622.

[36] Boley, D. A.; Zhang Z.; Dovichi, N. J. Multisegment injections improve peptide identification rates in capillary zone electrophoresis-based bottom-up proteomics. *J. Chromatogr. A* **2017**,1523, 123-126.

[37] Garza, S.; Moini, M. Analysis of complex protein mixtures with improved sequence coverage using (CE-MS/MS)n. *Anal. Chem*. **2006**, 78(20), 7309-16.

[38] Graves, J. D.; Krebs, E. G. Protein phosphorylation and signal transduction. *Pharmacol. Ther.* **1999**, *82*, 111-121.

[39] Yue, X.; Lukowski, J. K.; Weaver, E. M.; Skube, S. B.; Hummon, A. B. Quantitative Proteomic and Phosphoproteomic Comparison of 2D and 3D Colon Cancer Cell Culture Models. *J. Proteome Res.* **2016**, *15*, 4265-4276.

[40] Sharma, K.; D'Souza, R. C.; Tyanova, S.; Schaab, C.; Wiśniewski, J. R.; Cox, J.; Mann, M. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **2014**, *8*, 1583-94.

[41] Yue, X. S.; Hummon, A. B. Combination of multistep IMAC enrichment with high-pH reverse phase separation for in-depth phosphoproteomic profiling. *J. Proteome Res.* **2013**, *12*, 4176-86.

[42] 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*, 260-71.

[43] Song, C.; Ye, M.; Han, G.; Jiang, X.; Wang, F.; Yu, Z.; Chen, R.; Zou, H. Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphorylated peptides. *Anal. Chem.* **2010**, *82*, 53-6.

[44] Erickson, B. K.; Jedrychowski, M. P.; McAlister, G. C.; Everley, R. A.; Kunz, R.; Gygi, S. P. Evaluating multiplexed quantitative phosphorylated peptide analysis on a hybrid quadrupole mass filter/linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **2015**, *87*, 1241-9.

[45] Phanstiel, D. H.; Brumbaugh, J.; Wenger, C. D.; Tian, S.; Probasco, M. D.; Bailey, D. J.; Swaney, D. L.; Tervo, M. A.; Bolin, J. M.; Ruotti, V.; Stewart, R.; Thomson, J. A.; Coon, J. J. Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nat. Methods* **2011**, *8*, 821-7.

[46] Peuchen, E. H.; Cox, O. F.; Sun, L.; Hebert, A. S.; Coon, J. J.; Champion, M. M.; Dovichi, N. J.; Huber, P. W. Phosphorylation Dynamics Dominate the Regulated Proteome during Early Xenopus Development. *Sci. Rep.* **2017**, *7*, 15647.

[47] Wang, F.; Song, C.; Cheng, K.; Jiang, X.; Ye, M.; Zou, H. Perspectives of comprehensive phosphoproteome analysis using shotgun strategy. *Anal. Chem.* **2011**, *83*, 8078-85.

[48] Boersema, P. J.; Foong, L. Y.; Ding, V. M.; Lemeer, S.; van Breukelen, B.; Philp, R.; Boekhorst, J.; Snel, B.; den Hertog, J.; Choo, A. B.; Heck, A. J. In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphorylated peptide immunoaffinity purification and stable isotope dimethyl labeling. *Mol. Cell. Proteomics* **2010**, *9*, 84-99.

[49] Ubersax, J. A.; Ferrell, J. E. Jr. Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 530-41.

[50] Jorgenson, J. W.; Lukacs, K. D. Capillary zone electrophoresis. *Science* **1983**, *222*, 266-72.

[51] Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates, J. R. 3rd. In-line separation by capillary electrophoresis prior to analysis by top-down mass spectrometry enables sensitive characterization of protein complexes. *J. Proteome Res.* **2014**, *13*, 6078-86.

[52] Zhang, Z.; Hebert, A. S.; Westphall, M. S.; Qu, Y.; Coon, J. J.; Dovichi, N. J. Production of Over 27 000 Peptide and Nearly 4400 Protein Identifications by Single-Shot Capillary-Zone Electrophoresis-Mass Spectrometry via Combination of a Very-Low-Electroosmosis Coated Capillary, a Third-Generation Electrokinetically-Pumped Sheath-Flow Nanospray Interface, an Orbitrap Fusion Lumos Tribrid Mass Spectrometer, and an Advanced-Peak-Determination Algorithm. *Anal. Chem.* **2018**, doi: 10.1021/acs.analchem.8b02991.

[53] Busnel, J. M.; Schoenmaker, B.; Ramautar, R.; Carrasco-Pancorbo, A.; Ratnayake, C.; Feitelson, J. S.; Chapman, J. D.; Deelder, A. M.; Mayboroda, O. A. High capacity capillary electrophoresis-electrospray ionization mass spectrometry: coupling a porous sheathless interface with transient-isotachophoresis. *Anal. Chem.* **2010**, *82*, 9476-83.

[54] Cheng, Y. F.; Wu, S. L.; Chen, D. Y.; Dovichi, N. J. Interaction of capillary zone electrophoresis with a sheath-flow cuvette detector. *Anal. Chem.* **1990**, *62*, 496–503.

[55] Moini, M. Simplifying CE-MS operation. 2. Interfacing low-flow separation techniques to mass spectrometry using a porous tip. *Anal. Chem.* **2007**, *79*, 4241-6.

[56] 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*, 1130-7.

[57] Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J. Bottom-up proteomics of Escherichia coli using dynamic pH junction preconcentration and capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry. *Anal. Chem.* **2014**, *86*, 6331-6.

[58] Faserl, K.; Sarg, B.; Gruber, P.; Lindner, H. H. Investigating capillary electrophoresis-mass spectrometry for the analysis of common post-translational modifications. *Electrophoresis* **2018**, *39*, 1208-1215.

[59] Sarg, B.; Faserl, K.; Kremser, L.; Halfinger, B.; Sebastiano, R.; Lindner, H. H. Comparing and combining capillary electrophoresis electrospray ionization mass spectrometry and nano-liquid chromatography electrospray ionization mass spectrometry for the characterization of post-translationally modified histones. *Mol. Cell. Proteomics* **2013**, *12*, 2640-56.

[60] 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.

[61] Li, Q. R.; Ning, Z. B.; Tang, J. S.; Nie, S.; Zeng, R. Effect of peptide-to-TiO2 beads ratio on phosphorylated peptide enrichment selectivity. *J. Proteome Res.* **2009**, *8*, 5375-81.

[62] Yue, X.; Schunter, A.; Hummon, A. B. Comparing multistep immobilized metal affinity chromatography and multistep TiO2 methods for phosphorylated peptide enrichment. *Anal. Chem.* **2015**, *87*, 8837-44.

[63] Eng, J. K.; McCormack, A. L.; Yates, J. R.; 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976-89.

[64] Taus, T.; Köcher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **2011**, *10*, 5354-62.

[65] Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367-72.

[66] Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794-805.

[67] Krokhin, O. V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.* **2006**, *78*, 7785-95.

[68] Schwartz, D.; Gygi, S. P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* **2005**, *23*, 1391-8.

[69] Kubiniok, P.; Lavoie, H.; Therrien, M.; Thibault, P. Time-resolved Phosphoproteome Analysis of Paradoxical RAF Activation Reveals Novel Targets of ERK. *Mol. Cell. Proteomics* **2017**, *16*, 663-679.

[70] Zhou, H.; Ye, M.; Dong, J.; Corradini, E.; Cristobal, A.; Heck, A. J.; Zou, H.; Mohammed, S. Robust phosphoproteome enrichment using monodisperse microsphere-based immobilized titanium (IV) ion affinity chromatography. *Nat. Protoc.* **2013**, *8*, 461-80.

[71] Iliuk, A. B.; Martin, V. A.; Alicie, B. M.; Geahlen, R. L.; Tao, W. A. In-depth analyses of kinase-dependent tyrosine phosphoproteomes based on metal ion-functionalized soluble nanopolymers. *Mol. Cell. Proteomics* **2010**, *9*, 2162-72.

[72] Yang, Z.; Shen, X.; Chen, D.; Sun, L. Microscale Reversed-Phase Liquid Chromatography/Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Deep and Highly Sensitive Bottom-Up Proteomics: Identification of 7500 Proteins with Five Micrograms of an MCF7 Proteome Digest. *Anal. Chem.* **2018**, *90*, 10479-10486.

CHAPTER 4. Quantitative proteomics of zebrafish early-stage embryos with the quantification of 5000 proteins

4.1. Introduction

Embryology is a field studying the maturation of sex cells, fertilization, and embryonic development. An important branch of embryology, i.e., teratology, is dedicated to understanding birth defects. Centers for Disease Control and Prevention (CDC) has previously reported that around 3% of the newborn were affected by birth defects annually [1]. Till now, the understanding of the underlying reasons for various congenital disorders is still very limited. Investigation of early embryogenesis is desired since it can provide insights into the origin of birth defects.

Danio rerio, also known as zebrafish, is a widely used vertebrate model organism in embryological developmental biology due to its feasibility in handling, the high genomic similarity with humans, and rapid development [2-4]. Zebrafish embryos with diameters of ~0.7 mm are easy to handle and compatible with high-throughput multi-well plate assays. The transparency of zebrafish embryos also allows direct observation during early embryonic development. Moreover, the close relationship between the human and zebrafish genomes has been revealed by Howe *et al*. 70% of human genes were found to have at least one zebrafish orthologous gene and over 80% of human disease-related genes had at least one zebrafish orthologue [5]. These findings highlight the value of zebrafish as a model organism for evaluations such as disease progress and environmental toxicology [3,4,6,7]. Also, zebrafish embryos develop rapidly. In the first 24 hours, zebrafish embryos can

reach the stage where human embryos require one month to develop. The rapid development of zebrafish embryos can effectively shorten the timespan for research.

The first 24-hour zebrafish embryonic development comprises 5 key periods: zygote, cleavage, blastula, gastrula, and segmentation periods [8]. In the zygote period, embryos are single cells. In the cleavage period, the cells, or the blastomeres, divide every 15 min. At the eighth cycle (128-cell stage), the blastodisc becomes ball-like indicating the start of the blastula period. Before the tenth cleavage cycle (512-cell stage), the cells divide rapidly. The lengthening of the cell cycle denotes the onset of mid-blastula transition (MBT), in which the transcription of zygotic DNA is initiated, namely the zygotic genome activation (ZGA), and the cells start to be motile. Later in the gastrula period, the cells undergo initial differentiation and rearrangement and result in the formation of primary germ layers. In the segmentation period, the embryos continue to differentiate. The first 24-hour embryonic development of zebrafish can simulate the first month of human embryogenesis which is the critical period for the formation of developmental abnormalities.

There has been a considerable number of research on transcriptomic dynamic changes during early embryogenesis [9-14]. However, changes at the transcriptomic level during early embryogenesis cannot fully represent the changes at the protein level due to two reasons. First, the zygotic mRNA is silent before ZGA. Second, studies on the correspondence between mRNA and protein have shown a poor correlation between the expression levels of mRNA and protein [15]. Besides the expression level difference, the modulations of the gene expression by complicated post-transcriptional regulations and the influences of various post-translational modifications on gene functions cannot be reflected by transcriptome-level

information. Therefore, proteome-level information would be invaluable for the understanding of early embryogenesis.

BUP is a powerful tool for studying protein dynamics globally. However, the existing BUP-based zebrafish proteomics databases have either limited time-resolution or limited quantitative information in the first 24-hour zebrafish embryogenesis [16-19]. In 2006, Tay *et al.* profiled protein expression during early embryogenesis at 10 different time points using 2D-PAGE and MALDI-TOF, **Figure 4.1** [16]. Due to technological limitations, only 108 protein expression profiles were examined. On the other hand, three later studies on zebrafish proteome during early embryogenesis only employed embryos from less than 5 stages [17-19]. None of these databases contains information around the MBT stage, **Figure 4.1**. Limited coverage at embryonic developmental stages or insufficient protein ID numbers of these existing zebrafish proteome dynamic databases failed to provide sophisticated insights into zebrafish early embryogenesis.

In a typical BUP workflow, LC-MS/MS is the platform of choice in the delineation of complex proteomes [20-23]. Recently, CZE-MS/MS has shown its potential as an alternative platform for BUP with complementary peptide identification to LC-MS/MS [24-28]. In the present study, we coupled isobaric tag for relative and absolute quantitation (iTRAQ) chemistry with both CZE-MS/MS and LC-MS/MS to profile protein expression levels of zebrafish embryos across four stages during the maternal-to-zygotic transition (MZT) between fertilization and 6 hours post fertilization (hpf). We quantified nearly 5,000 proteins across the four embryonic stages with biological replicates. The clusters of protein expression profiles clearly indicated the important events that happened during the first 6 hours of zebrafish embryos' life. We also observed the wave-like expression patterns of tens of

transcription factors. Furthermore, the loss of function study of one important transcription factor, Nanog, demonstrated its important role in regulating early zebrafish embryogenesis.

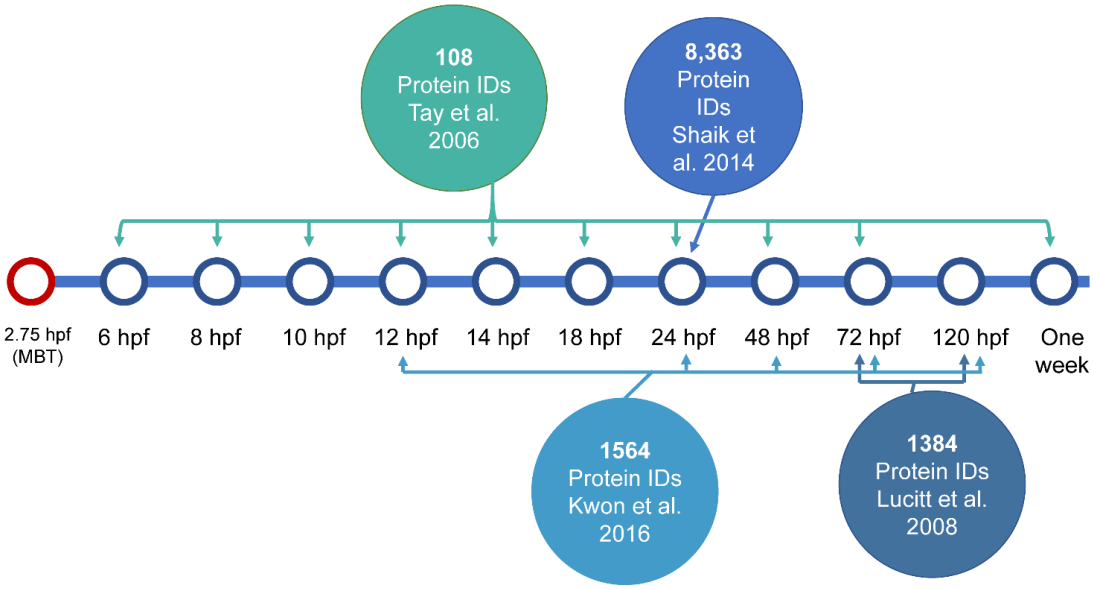

Figure 4.1. Schematic representation of existing zebrafish proteome databases.

4.2. Experimental

4.2.1. Material

All reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. LC/MS grade water, FA, methanol, ACN, HPLC grade AA and HF were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide was obtained from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were bought from Polymicro Technologies (Phoenix, AZ). Mammalian Cell-PE LB™ buffer for cell lysis was purchased from G-Biosciences (St. Louis, MO). Complete, mini protease inhibitor cocktail (provided in EASYpacks) and PhosSTOP phosphatase

inhibitor were purchased from Roche (Indianapolis, IN). Morpholino was purchased from Gene Tools, LLC (https://www.gene-tools.com/).

4.2.2. Zebrafish maintenance and breeding

Zebrafish are housed in a ZMod self-enclosed system. Water temperature was set at 28.5 °C. pH, ammonia and nitrites were checked every day. Twenty-five percent of the water in the system was changed daily. Each holding tank has 2-liter water capacity and has a maximum number of 12 adult fish per tank. Tanks are made of polypropylene and fish were moved to a clean tank approximately every 4 weeks. Fish were on a 14h light/10h dark light cycle. Adult fish were fed brine shrimp twice a day, and new brine shrimps were prepared daily. Newborn fish were fed using the GEMMA larval diet using the company recommendations and were transitioned to brine shrimp 30 days after fertilization.

One day before the breeding day, a male and a female were placed in the breeding tank, with a sieve separator. On the breeding day, soon after the lights come on, the separator was removed, and fish were briefly allowed to breed naturally. After natural breeding was observed, the fish were separated so the sperm (milt) and eggs can be manually collected for in vitro fertilization (IVF). Eggs and milt were collected manually. After breeding activity, the male and female were placed back into tanks and are rested for at least one month before the next round of breeding or experimental handling. **The maintenance and breeding of zebrafish were performed by the Cibelli group in the Department of Animal Science at Michigan State University.**

4.2.3. Embryo collection

Approximate 100 embryos at each of 4 different early stages (64-cells, 256-cells, Dome, 50%-epiboly) were collected and split into three 1.7 mL centrifuge tubes (~ 30 embryos/tube). The redundant liquid was carefully removed with 200-µL pipette. The tubes were immediately cooled down by liquid nitrogen and stored at -80 °C before use.

4.2.4. Sample preparation

Collected embryos in each tube were suspended in 700 µL of mammalian cell-PE LB™ cell lysis buffer (G-biosciences) containing complete protease inhibitor (Roche) and phosphatase inhibitor (Roche). After homogenization for 1 min on ice, all the tubes were sonicated for 10 min on ice using a Branson Sonifier 250 (VWR Scientific, Batavia, IL). The lysates were then centrifuged at 12,000 g for 5 min. The supernatants were collected and a small portion was used to measure the protein concentration with BCA assay. Based on the measured concentration, 200 µg of protein samples at each stage were purified by acetone precipitation: 1 volume of protein sample was mixed with 4 volumes of cold acetone and the mixtures were kept at -20 °C overnight. The tubes were then centrifuged at 10,000 g for 5 min and the supernatants were discarded. The pellets were simply washed using 500 µL of cold acetone and re-centrifuged. The supernatants were discarded, and the protein pellets were placed in a chemical hood for 1~2 min until they are dry (Caution: Do NOT overly dry!). The protein samples were stored at -80 °C before use.

All the protein pellets were suspended in 100 µL of 2% SDS (w/v) and 100 mM NH_4HCO_3 solution and vortexed and sonicated. To denature the protein, tubes were kept at 90 °C for 20 min. To reduce the disulfide bonds between cysteine amino acid

residues, 6 µL of 0.1 M of DTT was added to each tube and tubes were kept at 80 °C for 20 min. Protein alkylation with 15 µL of 0.1 M IAA was done at room temperature in dark for 20 min. 6 µL of 0.1 M DTT was added to each tube to react with the residue IAA. The sample was then mixed with 125 µL 8 M urea in 100 mM NH_4HCO_3. Then the mixture of each tube was loaded onto a 30-kDa centrifugal filter unit (250 µL/unit) followed by centrifugation at 12,000 g for 10 min. The proteins on the membrane were washed with 250 µL of 8 M urea in 100 mM NH_4HCO_3 three times. Next, the proteins were washed with 100 mM NH_4HCO_3 three times to remove urea. Finally, 150 µL of 100 mM NH_4HCO_3 (pH 8.0) was loaded on each membrane and 7 µL of trypsin solution (1 µg/µL) was added to each unit. The filter units were gently vortexed for 5 min to mix the trypsin and proteins. After that, the filter units were kept in a 37 °C water bath overnight for tryptic digestion.

After digestion, the units were centrifuged at 12,000 g for 15 min, and the flow-through containing the peptides was collected. To increase peptide recovery from the membrane, the membrane was further washed with 150 µL of 100 mM NH_4HCO_3. FA was used to acidify the protein digests to terminate digestion (0.5% FA final), followed by desalting using C18 SPE columns (50 mg beads, 250 µg peptide capacity) (Waters, Milford, MA), the elutes from desalting were lyophilized with a vacuum concentrator (Thermo Fisher Scientific) (Caution: Do NOT overly dry!). The digests were stored at -80 °C before use.

4.2.5. iTRAQ labeling

The lyophilized digests were dissolved in 70 µL of 500 mM Triethylammonium bicarbonate buffer (Dilute the 1 M buffer with water, make sure the pH >7.5). Then withdraw 35 µL of the solution and added 50 µL of isopropanol to each iTRAQ

reagent vial. The iTRAQ reagent (60-70 uL) was transferred to each sample and incubated for 2 h. 50 µL of 100 mM Tris-HCl (pH 8.0) buffer was added to each tube followed by incubation at room temperature for 40 min to block the residue iTRAQ reagent. Then the digests were mixed and acidified by formic acid to 0.5% formic acid final concentration. The sample was lyophilized to remove the organic solvent. When there was ~200 µL solution left, the lyophilization was stopped. 600 µL of 0.5% FA was added to the sample (total volume 800 µL). The samples were desalted with two C18 SPE columns (100 mg beads, 500 µg capacity, Waters, Milford, MA) and lyophilized. The sample was re-dissolved in 0.1% FA, 2% ACN (800 µL) and kept at -20 °C before use.

4.2.6. High-pH RPLC fractionation for iTRAQ labeled zebrafish embryo digest

An Agilent Infinity II HPLC system was used for high pH RPLC fractionation. A Zorbax 300Extend-C18 RP column (2.1 mm i.d. × 150 mm length, 3.5 µm particles, Agilent Technologies) was used for separation. Mobile phase A was 5 mM NH_4HCO_3 in water with pH 9 and mobile phase B was 5 mM NH_4HCO_3 in 80% ACN with pH 9. Gradient elution was used for peptide separation.

800-µg iTRAQ-labeled zebrafish embryo digest was injected onto the RP column for each experiment. The flow rate was 0.3 mL/min. The gradient was as follow: 0-5 min, 2%B; 5-8 min, 2-10%B; 8-108 min, 10-50%B; 108-110 min, 50-100%B; 110-120 min, 100%B; 120-122 min, 100-2% B; 122-132 min, 2% B. In total, 100 fractions were collected from 8 min to 108 min, one fraction per min. We named the fractions based on the order of retention time from 1 to 100. Then we combined the fraction N and fraction N+50 to generate 50 fractions. Those fractions were then lyophilized and stored at -80 ^{o}C for low pH RPLC-MS/MS.

4.2.7. LC-MS/MS

An EASY-nLC™ 1200 system (Thermo Fisher Scientific) was used for RPLC separation. Each of 50 high-pH RPLC fraction was dissolved in 10 µL of 0.1% (v/v) FA and 2% (v/v) ACN. 2 µL of the sample was injected onto a C18 pre-column (Acclaim PrepMapTM 100, 75-µm i.d. × 2 cm, nanoviper, 3 µm particles, 100 Å, Thermo Scientific). Then, the loaded peptides were separated on a C18 separation column (Acclaim PrepMapTM 100, 75-µm i.d. × 50 cm, nanoviper, 2 µm particles, 100 Å, Thermo Scientific) at a flow rate of 200 nL/min. Mobile phase A was 2% (v/v) ACN in water containing 0.1% (v/v) FA), and mobile phase B was 80% (v/v) ACN and 0.1% (v/v) FA. For separation, a 150-min gradient was used: 0-80 min, 8-30% B; 80-120 min, 30-55% B; 120-135 min, 55-100% B, 135-150 min, 100% B. The LC system required another 30 min for column equilibration between runs. Therefore, one LC-MS run required about 2 h.

A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for the RPLC-MS/MS experiments. The spray voltage was set to 1.8 kV. A Top10 DDA method was used. The mass resolution was set to 60,000 (at m/z 200) for both full MS scans and MS/MS scans. For full MS scans and MS/MS scans, the target value was 3E6 and 1E5, the maximum injection time was 50 ms and 110 ms, respectively. Scan range for MS scans was 300 to 1500 m/z. For MS/MS scans, the isolation window was 2 m/z. Fragmentation in the HCD cell was performed with a normalized collision energy of 28%. The fixed first mass was set to 100 m/z. Dynamic exclusion was applied and it was set to 30 s. Ions with unassigned or +1 charge states were not considered for fragmentation.

4.2.8. CZE-MS/MS

An ECE-001 capillary electrophoresis autosampler (CMP Scientific, Brooklyn, NY) was used for CZE separation. A commercialized electro-kinetically pumped sheath flow interface (CMP Scientific, Brooklyn, NY) [29] was employed for coupling CZE to MS.

A one-meter-long LPA coated capillary (50 μm i.d., 360 μm o.d.) was made in house based on ref. 30. A Sutter instrument P-1000 flaming/brown micropipette puller was used to pull a borosilicate glass capillary (1.0 mm o.d., 0.75 mm i.d., and 10 cm length). The resultant electrospray emitter has an opening with a diameter of 30 to 40 mm. The background electrolyte (BGE) of CZE was 5% (v/v) acetic acid and the sheath buffer was 0.2% (v/v) FA containing 10% (v/v) methanol. 4 μL of leftover samples in RPLC-MS/MS analysis was withdrawn and lyophilized. The dry sample was redissolved in 3 μL of 50 mM NH_4HCO_3 (pH 8.0). The sample was injected using 5-psi pressure for 95 seconds. 30 kV was applied at the injection end for CZE separation and around 2 kV was applied in the sheath buffer vial for electrospray. For all the CZE–MS runs, the separation was performed for 90 min followed by 10 min wash BGE for 10 min at 5 psi pressure.

4.2.9. Data analysis

All raw files were analyzed by MaxQuant 1.5.5.1 software with the Andromeda search engine. The zebrafish proteome (ID: UP000000437) downloaded from UniProt (http://www.uniprot.org/) was used for database search. iTRAQ on lysine and N-terminus was selected as peptide labels. The peptide mass tolerances of the first search and main search were 20 and 4.5 ppm, respectively. The fragment ion mass tolerance was 20 ppm. Trypsin was selected as the protease. The dynamic

modification was oxidation on methionine and acetylation on protein N-terminus, and carbamidomethyl on cysteine was set as static modifications. The minimum length of a peptide was set to 7. Match between runs was enabled, The FDRs were 1% for both peptide and protein IDs.

The data processing was performed using Perseus software [31]. Average intensities of every two biological replicates were normalized to that at 64-cell for each protein ID. All values were then transferred to log(2) values. Cluster analysis was performed on Graphical Proteomics Data Explorer (GProX) [32]. The number of clusters was set to 9 with an upper limit of 0.4 and a lower limit of -0.5.

4.2.10. Morpholino injection

The Nanog morpholino (MO) (sequence: CTGGCATCTTCCAGTCCGCCATTT-C) and the tp53 MO (sequence**:** TCAATTCTTGCAAAGCAATGGCGCA) were dissolved in sterile water to 0.3 mM. A Sutter instrument P-1000 flaming/brown micropipette puller was used to make the glass needles for microinjection. The injected volume was estimated by measuring the diameter of the droplet. Unless stated otherwise, approximately 1 nL of MO was injected into the yolk region in each embryo at the one-cell stage.

4.3. Results and discussion

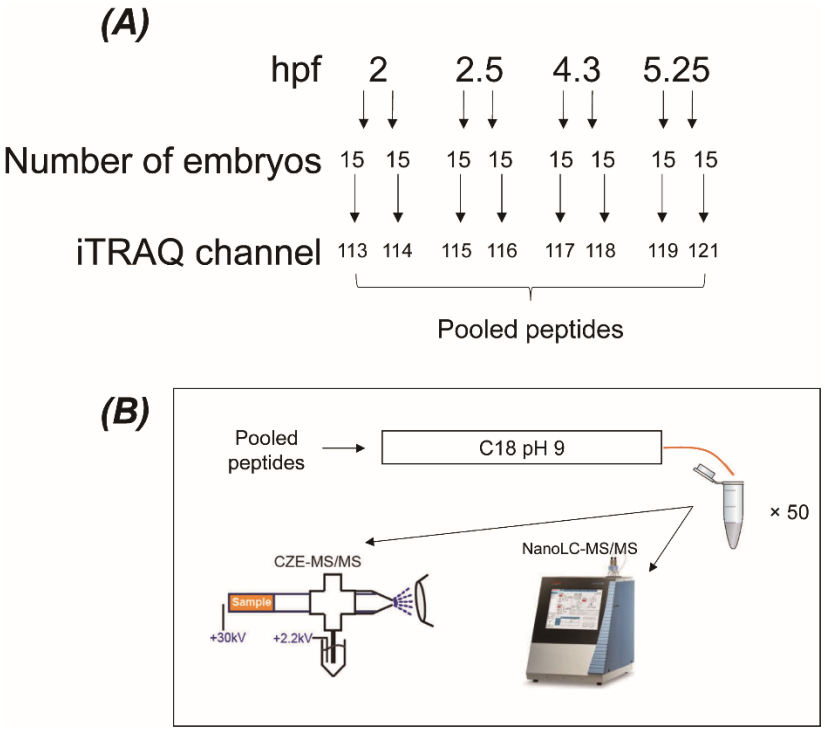

Figure 4.2. Experimental design. Design of three iTRAQ labeling experiment (A). Separation strategy of pooled peptides (B).

We performed an iTRAQ experiment on zebrafish embryos during early embryonic development. Stages of embryos employed, and the design of labeling are shown in **Figure 4.2A**. In the experiment, embryos from 64-cell (2 hpf), 256-cell (2.5 hpf), Dome (4.3 hpf), 50%-epiboly (5.25 hpf) stages were examined to study proteome dynamic around the MBT stage (512-cell, 2.75 hpf). 30 embryos at each of the four stages were equally split and labeled by separate iTRAQ channels as biological duplicates. After enzymatic digestion of proteins, the resultant peptides from different embryonic stages were labeled by 8-channel iTRAQ reagents. The labeled peptides were pooled together. **Figure 4.2B** shows the separation strategy for pooled iTRAQ-labeled peptides. Pooled peptides were first offline fractionated by high pH (pH=9) RPLC to 50 fractions. Each of the fractions was separated into two portions. One portion was subjected to nanoRPLC-MS/MS, and the other was

subjected to CZE-MS/MS. Both platforms employed a Q Exactive HF mass spectrometer for analysis.

4.3.1. Deep proteome analysis of zebrafish embryos during early embryogenesis

After 83-hour analysis, CZE-MS/MS acquired 77,226 peptide-separtrum-matches (PSMs) accounting for 20,888 peptide IDs and 3,009 protein IDs. The 50 nanoRPLC-MS/MS runs cost approximately 150 hours. 107,331 PSMs were acquired by 50 nanoRPLC-MS/MS with 22,895 peptide IDs and 3545 protein IDs. As expected, good orthogonality was found between CZE-MS/MS and nanoRPLC-MS/MS as demonstrated by the small overlap in peptide IDs, **Figure 4.3A**. Analysis of all raw files from two platforms approached 5,031 protein IDs, 67% and 42% higher than using CZE-MS/MS or nanoRPLC-MS/MS alone, respectively, **Figure 4.3B**. The nanoRPLC-MS/MS has an average sequence coverage of 18.0%. After combining the results from two platforms, the average sequence coverage was boosted to 25.7%. These results underline the improvement in both proteome coverage and sequence coverage caused by the complementarily between CZE-based and RPLC-based platforms. Our data represents the largest quantitative dataset of zebrafish embryos during early embryogenesis to date.

Before MBT, protein expression relies on maternally deposited mRNAs since zygotic gene transcription is silent [33]. After MBT, gene expression transforms from the modulation of maternal mRNA to the exclusive control of the zygotic genome. This process is known as ZGA. Proteins play pivotal roles in the escape of zygotic genome silencing. A crucial modulatory factor for the silence of zygotic genes is the restricted systematic regulatory network constructed by transcription factors (TFs).

By binding to specific DNA sequences, TFs can either trigger or suppress certain gene expression in time- and cell type-specific fashions [34,35]. A zebrafish proteome database covering stages around MBT has been desired to understand the programmed regulation during the ZGA. After comparing our data with the TF list predicted by protein sequence in Animal Transcription Factor DataBase (AnimalTFDB) 3.0(http://bioinfo.life.hust.edu.cn/AnimalTFDB/ #!/) [36], we found 118 annotated TFs in our dataset.

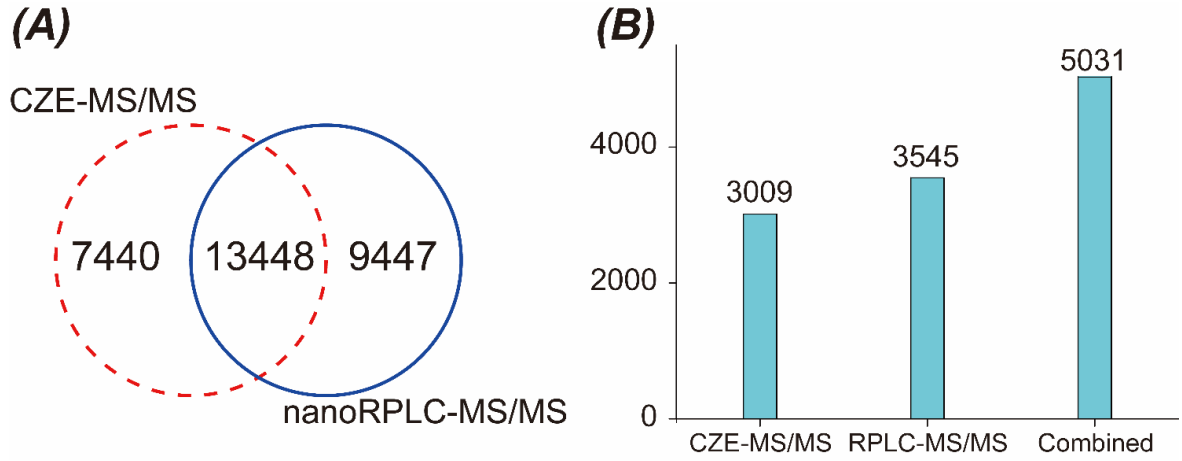

Figure 4.3. Comparison of CZE-MS/MS and RPLC-MS/MS in terms of the IDs from zebrafish embryo proteome digest. (A) Overlap of peptides. (B) Column plot of protein ID numbers of CZE-MS/MS, RPLC-MS/MS and combination of two platforms.

We performed the gene ontology (GO) analysis and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of over 5,000 annotated proteins. We plotted out the 11 most abundant elements in each analysis, **Figure 4.4A-D**. Important embryonic development-associated events like metabolic process [37], cell cycle and protein folding [38] are also found in the top 11 biological processes (BP) events, **Figure 4.4A**. Events associated with ZGA such as translation and proteolysis are the top 2 and 4 abundant BP events, respectively. Genes related to nucleotide binding, RNA binding, transferase activity, and structural constituent of ribosome are found highly expressed during the early stages of embryonic

development via molecular function (MF) analysis, **Figure 4.4B**. These genes may be involved in the ZGA-associated transcription, translation, and chromatin structural remodeling. The nucleus, as the cell sub-compartment where zygotic transcription occurs, has the second largest ratio in the cellular component (CC) analysis, **Figure 4.4C**. Similarly, ZGA-associated pathways like RNA transport, ribosome, and Ubiquitin mediated proteolysis are enriched in the KEGG analysis, **Figure 4.4D**. The GO and KEGG pathway analysis reflect the important events during zebrafish early embryonic development and highlighted the onset of ZGA.

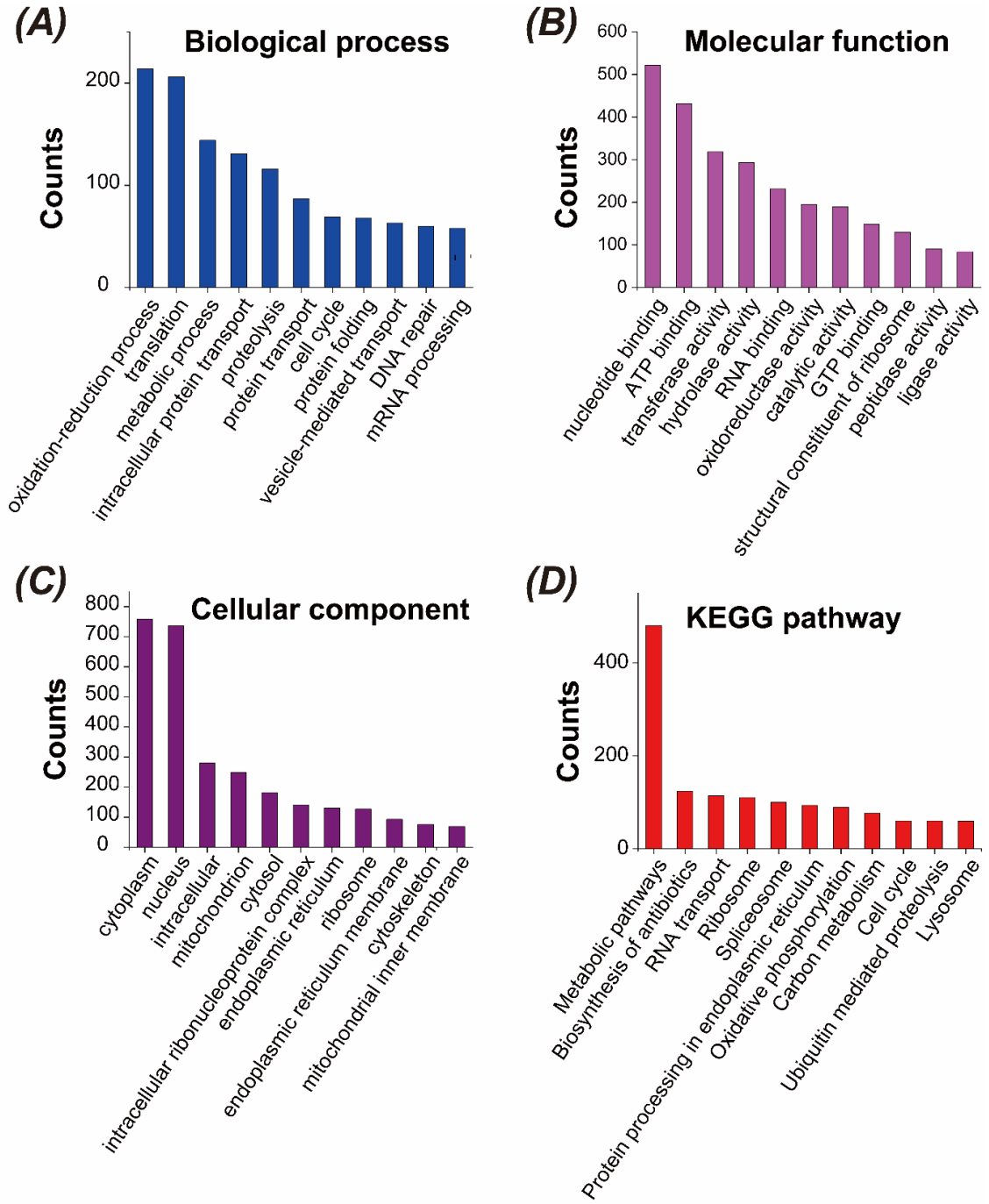

Figure 4.4. Gene Ontology (GO) information of identified proteins using both CZE-MS and LC-MS. DAVID bioinformatics resources 6.8 (https://david.ncifcrf.gov/) was used to get the GO information of proteins. The GO terms were sorted by the number of proteins (Count). The top11 GO terms were used for the figure. (A) Biological process; (B) Molecular function; (C) Cellular component; (D) KEGG pathway.

4.3.2. Cluster analysis

We performed cluster analysis on expression dynamics for all quantified proteins. Proteins were clustered to group 0 in which no significant abundance changes were found on the group members and 9 other groups that have shown significant changes in abundance, **Figure 4.5**. All the protein expression level data were normalized to the 64-cell-stage data. Protein expression levels higher than 1.3 or lower than 0.7 compared to the 64-cell stage were considered as proteins with significant changes in abundance, corresponding to 0.4- and -0.5-fold changes in log(2) values, respectively. Group 0 contains 3,978 proteins that had no significant changes in expression levels when Group 1 to 9 have a total of 1,053 proteins with varied numbers of proteins shown in **Figure 4.5**. To have a better understanding of BP events that proteins are involved in for each cluster, BP enrichment was also performed, **Figure 4.6**.

Proteins in cluster 1 are upregulated in early embryogenesis, **Figure 4.5**. They initially are the products of maternally deposited transcripts before MBT, and their expression levels are further heightened by the expression of maternal or both maternal and zygotic transcripts after MBT. Gene products in this cluster are involved in nucleosome assembly, a process where histones and DNAs are constructed together and form nucleosomes, and ribosomal RNA processing, which matures ribosomal RNA for translation of proteins, **Figure 4.6**. We also manually found proteins associated with mRNA splicing (u2af2b, ddx5, cstf3), and histone modification (naa40) in cluster 1. Uplift of these proteins before MBT indicates that they may modulate gene expression and promote protein translation. Proteins in cluster 1 may also promote the escape from zygotic gene repression such as Nanog, a well-known TF. Nanog serves as an activator for several hundred genes in ZGA

and it is required to initiate the zygotic development in zebrafish [39]. Similar trajectories of Nanog expression during early embryogenesis were proven by immunoaffinity assays in the literature [40,41]. Interestingly, kidney development related proteins are also enriched in this cluster suggesting certain types of organogenesis may initiate from the very early stages of embryonic development.

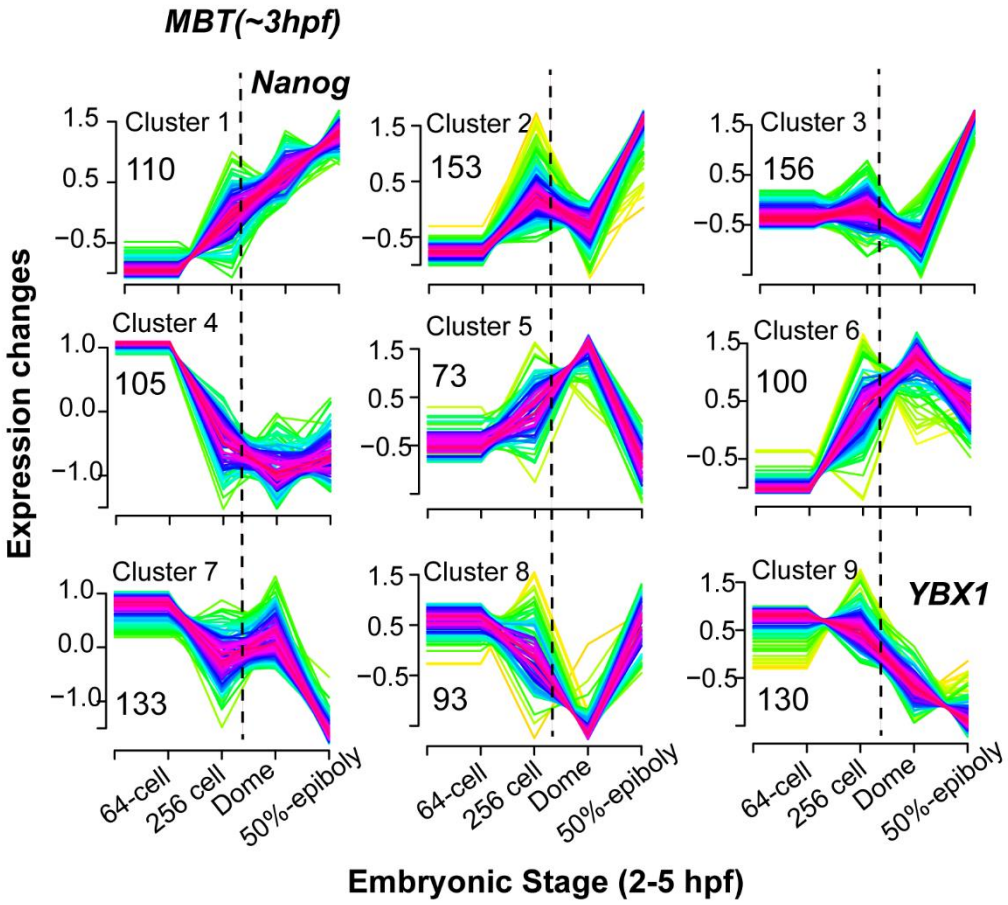

Figure 4.5. Cluster analysis of quantified proteins.

Cluster 4 and 9 are comprised of over 200 proteins that are maternally deposited and gradually decreased after fertilization. The difference between these two clusters is the expression behaviors after MBT. The decline of proteins in cluster 4 is attenuated after ZGA whereas proteins in cluster 9 have a consistent declining rate throughout early embryonic stages. Cluster 9 contains proteins that negatively

regulate translation initiation, **Figure 4.6**, such as Y-box-binding protein 1 (gene name: ybx1). Y-box-binding protein 1 represses global translation in oocytes [42]. Negative regulators of protein translation may contribute to safeguarding the maturation of oocytes by controlling gene expression in a time-specific manner [42]. Proteins in cluster 4 are majorly involved in protein transportation. Transportation of proteins, for example from the cytoplasm to the nucleus, is indispensable to cell processes like differentiation and control of gene expression. An interesting sample in cluster 4 is another TF called Forkhead box k2, which activates the Wnt-beta catenin pathway by transporting dishevelled proteins into the nucleus [43]. Unlike Nanog, Forkhead box k2 can act as both transcription activator and repressor [44,45]. We speculate Forkhead box k2 serves as a global transcription repressor like Y-box-binding protein 1 before MBT. But when gene transcription initiates, Forkhead box k2 became an activator for pathways. More studies are needed to achieve a better understanding of the function of Forkhead box k2 in regulating early embryogenesis.

The other clusters display significant trajectory changes after MBT. Cluster 2 enriches proteins engaging in Kupffer's vesicle and erythrocyte development. Early research proposed that Kupffer's vesicle is a zebrafish-specific organ that guides the development of the heart, brain and gut [46]. Our data suggest erythropoiesis is one of the earliest types of cell-differentiation, and initial organogenesis takes place very soon after ZGA.

Proteins in cluster 3 are rapidly upregulated after MBT. A good number of proteins (e.g., mrpl51, rps4x, rps15, rpl39, rpl14, rplp1, rps9, rps15a) in this cluster are involved in gene expression-related events including mRNA splicing, ribosome biogenesis, and translation, **Figure 4.6**. Since the assembly of translation machinery

is required for the translation of zygotic mRNA, it is rational that the ribosomal proteins are highly expressed after ZGA.

Cluster 5 contains proteins that are responsible for RNA and protein transports. In this cluster, we found CHMP1, a key protein involved in the formation of multivesicular bodies (MVBs) [47]. MVB is a type of endosome facilitating protein degradation by trafficking proteins to lysosomes or extracellular spaces [48]. CHMP1's upregulation before MBT and the following decline after MBT indicates the CHMP1-mediated formation for MVBs has an relationship with the maternal protein abundances. Therefore, MVBs may be involved in maternal protein degradation. Besides protein transportation, CHMP1 was found to affect chromatin structure and consequentially gene expression [49].

Deoxyribonucleotide biosynthesis is the major enriched BP event in cluster 6. The uplift of the related proteins as shown in cluster 6 implies the requirement for a large amount of deoxyribonucleotide during rapid cell division prior to MBT. However, the time for each cell division cycle increases after MBT thus reduces the need for deoxyribonucleotide biosynthesis marked by the decline of the associated proteins.

Proteins in cluster 7 majorly engage in nucleotides regulation. CCR4-NOT transcription complex subunit 2 is a member of the CCR4-NOT protein complex which is linked to mRNA degradation [50,51]. The CCR4-NOT subunit 2’s sustaining level of before MBT and rapid decline suggest its engagement in maternal mRNA decay.

Similar to cluster 7, cluster 8 has enriched BP events involved in nucleotide processing, **Figure 4.6**. The subunits 4 and 7 of the RNA exosome, a general

regulator of mRNA turnover, are found in this cluster. Besides, some genes in cluster 8 (mrpl14, polr2ea, and rps26l), and cluster 3 (mrpl51, rps4x, rps15, rpl39, rpl14, rplp1, rps9, rps15a) are either ribosomal proteins or translation-related. Since assemble of translation machinery is required for the translation of zygotic mRNA, it is rational that the ribosomal proteins are upregulated right after ZGA.

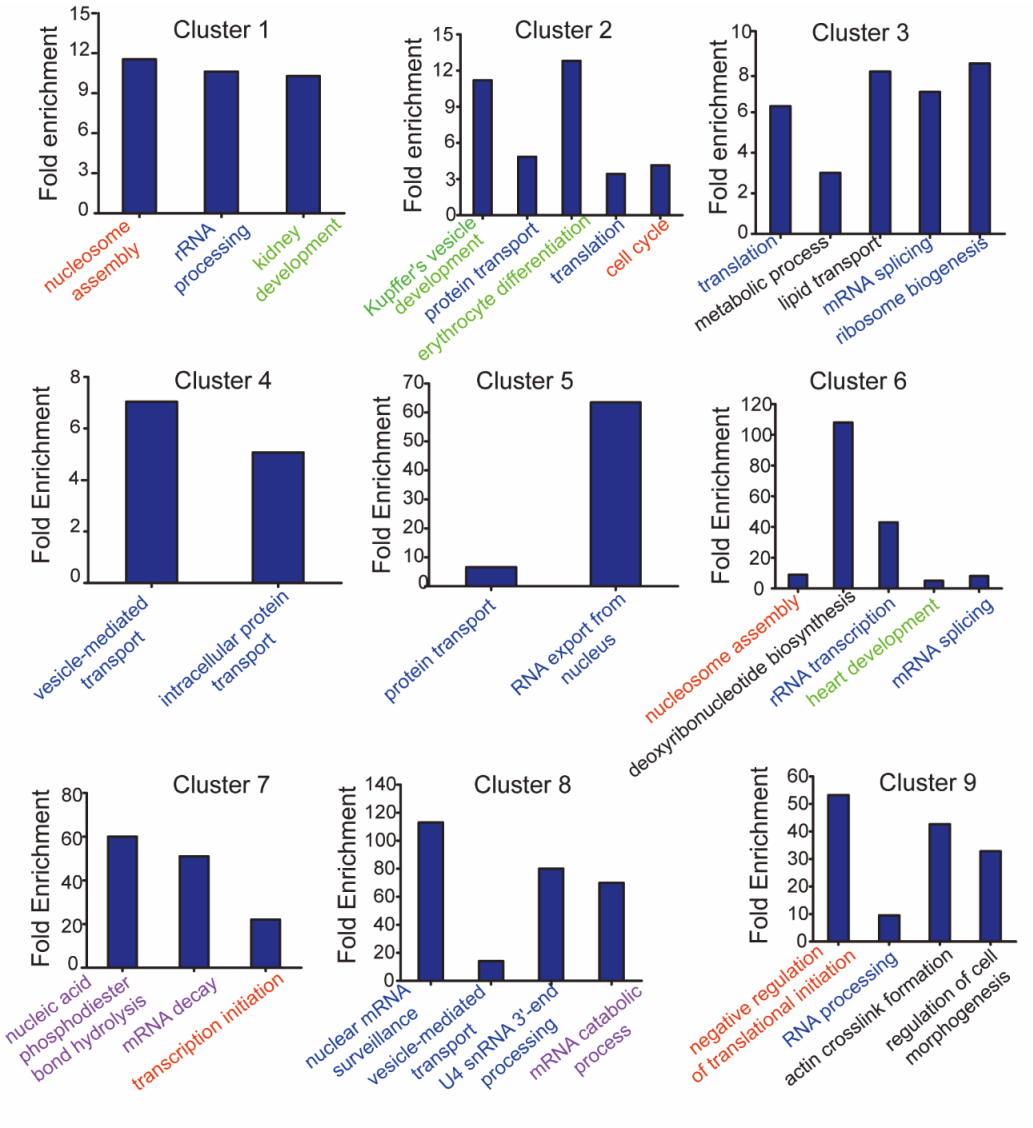

Figure 4.6. Biological process enrichment of proteins in each cluster.

4.3.3. Transcription factors expression dynamics

By correctly controlling gene expression, TFs serve as essential components in the restricted regulatory system that controls many biological processes, including crucial activities in early embryogenesis such as cell differentiation and organogenesis [52]. Aberrant expressions of TFs are attributed to one-third of human developmental disorders [53]. Understanding TF functions may enrich our knowledge of the cause of birth defects. Computational prediction of new TF identification and function was performed by mapping DNA-binding domains (DBDs) to the known TFs [36, 54]. We observed that 32 out of 108 identified TFs have significant abundances changes across the four stages, **Figure 4.7**.

Previous transcriptomic research has revealed that maternal-to-zygotic transition consists of two waves of ZGA [34,35,39,55]. Certain maternal transcription regulators were found to activate the first wave of ZGA [39]. Coinciding with transcriptomic profiling, the result of our cluster analysis also illustrated waves of TF expressions. The difference between wave 1, **Figure 4.7A**, and wave 2, **Figure 4.7B**, is the stage when their primary boost in expression takes place. Proteins in wave 1 start to be rapidly upregulated before the 256-cell stage whereas proteins in wave 2 begin to be elevated after the 256-cell stage. Maternal translations in both waves are accelerated prior to ZGA indicating TFs in these two waves may be the transcriptional regulators to the first wave of ZGA, exampled by a famous first-wave transcriptional activator Nanog in wave 2. On the other hand, TFs in wave 3 with a boost in concentration after ZGA may be the products of the first wave and serve as the transcription regulators for the second ZGA wave, **Figure 4.7C**. A proposed mechanism in ZGA onset is the titration of maternal repressors such as ybx1 [34,35], which suggests proteins in **Figure 4.7D** may play inhibitory roles in ZGA.

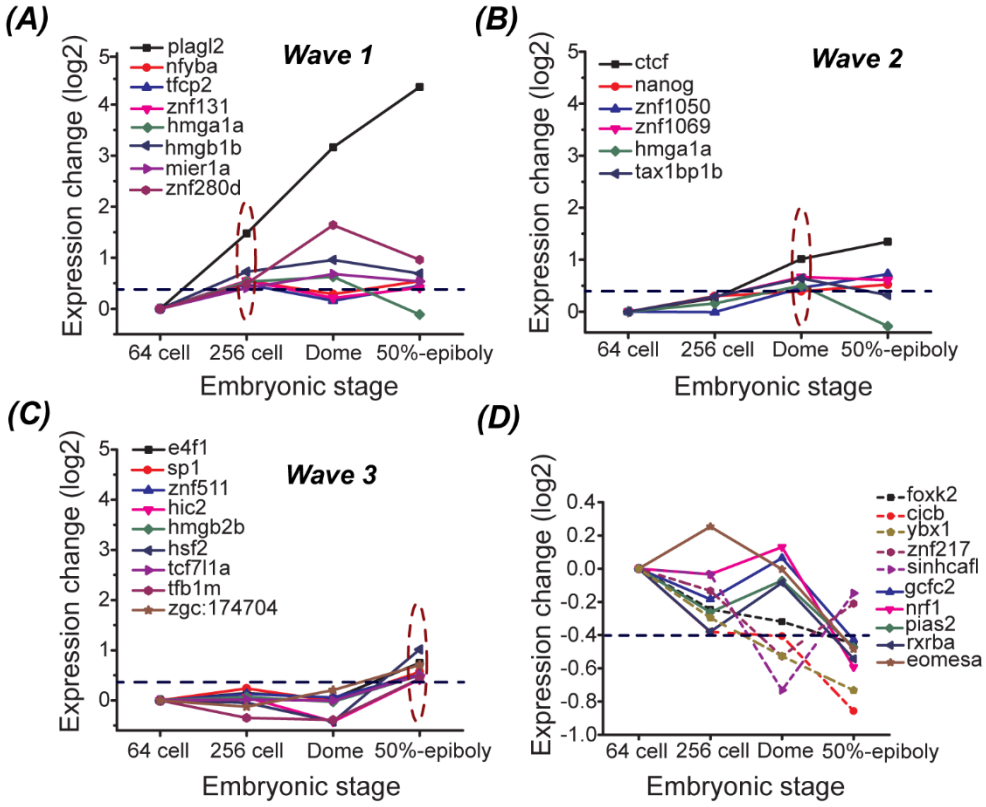

Figure 4.7. Expression profiles of 32 TFs with significant changes in abundance across the four stages. (A) Expression changes of TFs had significantly elevated levels at the 256-cell stage. (B) Expression changes of TFs had significantly elevated levels at the Dome stage. (C) Expression changes of TFs had significantly elevated levels at the 50%-epiboly stage. (D) Expression changes of TFs had a significant abundance decline.

4.3.4. Loss of function of Nanog via morpholino injection

Although rough prediction of protein functions can be made by the expression trajectory during early embryogenesis, experimental verification is inevitable for understanding proteins' function. A classic method for loss of gene function in zebrafish is the injection of the corresponding morpholino (MO) [56]. MO is a synthetic oligomer containing DNA bases that can suppress gene expression by binding to complementary RNA or single-strand DNA. To observe the phenotype of Nanog knock-down zebrafish embryos, we injected 1 nL of 0.3 mM of Nanog MO

(Sequence in experimental section) or 1 nL of culture media into embryos at the one-cell stage. We found Nanog MO-injected sample can normally undergo rapid cleavages as embryos in the control group, **Figure 4.8A** and **Figure 4.8B**. 6 hours after fertilization, a signature formation of blastoderm that margins to 50% of the diameter was found in the control embryos, **Figure 4.8C**. On contrary, the blastomeres in MO-injected cells were arrest at MBT and failed to enter gastrulation, **Figure 4.8D.** The co-injection of Nanog MO and tp53 MO led to the same embryo phenotype, **Figure 4.8E,** as Nanog MO-injected embryos**, Figure 4.8D,** indicating the cell-arrest phenotype was caused by loss of Nanog instead of the injection-associated tp53-dependent neural toxicity [57].

The loss of function study data agrees reasonably well with the literature [39,41], which highlighted the importance of maintaining the expression level of Nanog during early embryogenesis. The data also showed that we successfully downregulated the expression of Nanog through MO injection. We are working on western blot analysis of these MO and control embryos at different stages to confirm the Nanog expression profiles. We plan to perform quantitative BUP analysis of the control and MO-injected embryos across different stages during maternal to zygotic transition for elucidating the function of Nanog. The combination of gene knockdown and/or overexpression with quantitative BUP will bridge the genotypes and phenotypes and will allow us to pursue better understandings of functions of tens of transcription factors quantified in this study. The outcome from the studies will be invaluable for the developmental biology community.

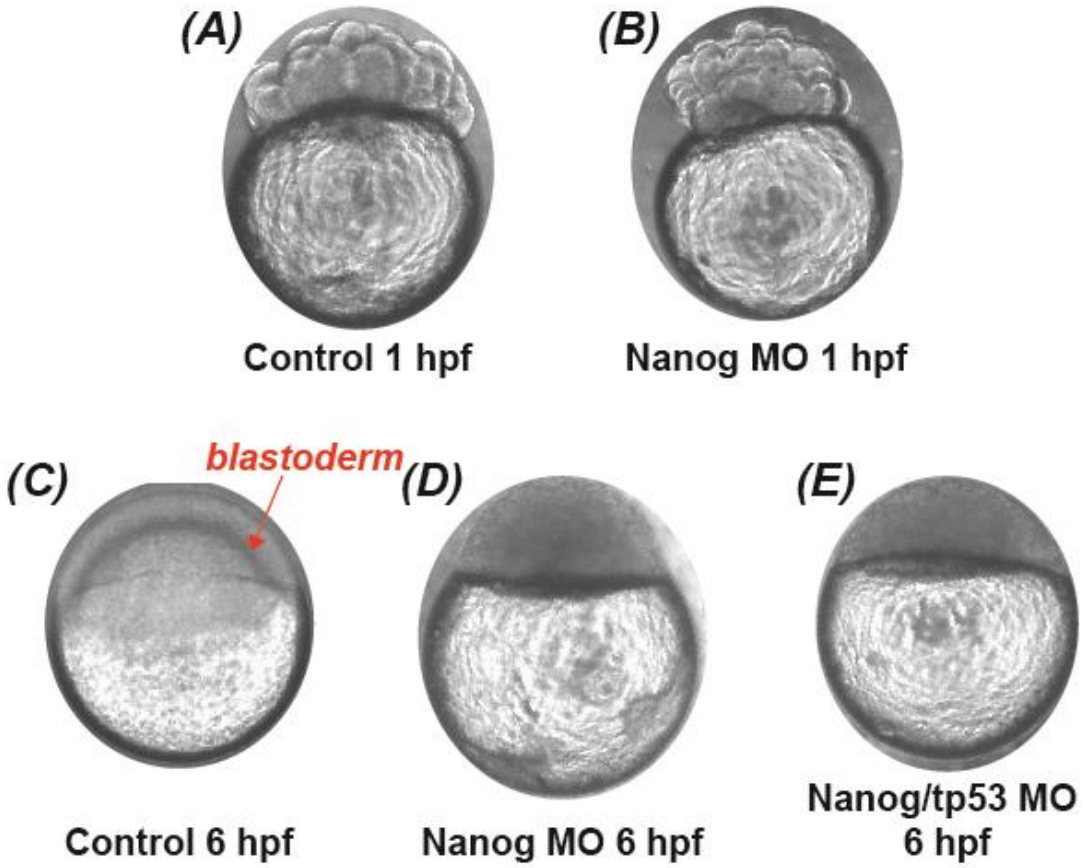

Figure 4.8. Embryos showing effects of Nanog MO. The control embryos injected culture media were observed at 1 hpf (A) and 6 hpf (C). Embryos injected with Nanog MO were observed at 1 hpf (B) and 6 hpf (D). Embryos co-injected with Nanog MO and tp53 MO were observed at 6 hpf (E)

4.4. Conclusions

In this study, we created the largest quantitative zebrafish proteome database during early embryogenesis to date. We unprecedentedly quantified nearly 5,000 proteins around the MBT. The cluster profiles precisely reflect the major events including the decline of maternal repressor and the accumulation of transcription activator which are marked with significant gene expression fluctuations. Wave-like TF expression patterns were found suggesting the specific regulatory roles of TFs to both waves of ZGA. We also demonstrated a potential method to determine protein

functions combined by loss of function assay and BUP. Overall, this work provides a draft map for future studies on zebrafish embryonic development.

4.5. Acknowledgments

We thank Prof. Jose Cibelli's group at the Department of Animal Science of Michigan State University for kindly providing the Zebrafish embryos for this project. We thank the support from the National Science Foundation (CAREER Award, DBI-1846913) and the National Institutes of Health (R01GM125991).

REFERENCES

REFERENCES

[1] Centers for Disease Control and Prevention (CDC). Update on overall prevalence of major birth defects--Atlanta, Georgia, 1978-2005. *MMWR Morb. Mortal. Wkly. Rep.* **2008,** 11, 57(1), 1-5.

[2] Horsfield, J. A. Packaging development: how chromatin controls transcription in zebrafish embryogenesis. *Biochem. Soc. Trans*. 2019, 47(2), 713-724.

[3] Zaucker, A.; Kumari, P.; Sampath, K. Zebrafish embryogenesis - A framework to study regulatory RNA elements in development and disease. *Dev. Biol.* **2020**, 457(2), 172-180.

[4] Carnovali, M.; Banfi, G.; Mariotti, M. Zebrafish Models of Human Skeletal Disorders: Embryo and Adult Swimming Together. *Biomed. Res. Int.* 2019, 1253710.

[5] Howe, K.; Clark, M. D.; Torroja, C. F.; Torrance, J.; Berthelot, C.; Muffato, M.; Collins, J. E.; Humphray, S.; McLaren, K.; Matthews, L.; McLaren, S.; Sealy, I.; Caccamo, M.; Churcher, C.; Scott, C.; Barrett, J. C.; Koch, R.; Rauch, G. J.; White, S.; Chow, W.; Kilian, B.; Quintais, L. T.; Guerra-Assunção, J. A.; Zhou, Y.; Gu, Y.; Yen, J.; Vogel, J. H.; Eyre, T.; Redmond, S.; Banerjee, R.; Chi, J.; Fu, B.; Langley, E.; Maguire, S. F.; Laird, G. K.; Lloyd, D.; Kenyon, E.; Donaldson, S.; Sehra, H.; Almeida-King, J.; Loveland, J.; Trevanion, S.; Jones, M.; Quail, M.; Willey, D.; Hunt, A.; Burton, J.; Sims, S.; McLay, K.; Plumb, B.; Davis, J.; Clee, C.; Oliver, K.; Clark, R.; Riddle, C.; Elliot, D.; Threadgold, G.; Harden, G.; Ware, D.; Begum, S.; Mortimore, B.; Kerry, G.; Heath, P.; Phillimore, B.; Tracey, A.; Corby, N.; Dunn, M.; Johnson, C.; Wood, J.; Clark, S.; Pelan, S.; Griffiths, G.; Smith, M.; Glithero, R.; Howden, P.; Barker, N.; Lloyd, C.; Stevens, C.; Harley, J.; Holt, K.; Panagiotidis, G.; Lovell, J.; Beasley, H.; Henderson, C.; Gordon, D.; Auger, K.; Wright, D.; Collins, J.; Raisen, C.; Dyer, L.; Leung, K.; Robertson, L.; Ambridge, K.; Leongamornlert, D.; McGuire, S.; Gilderthorp, R.; Griffiths, C.; Manthravadi, D.; Nichol, S.; Barker, G.; Whitehead, S.; Kay, M.; Brown, J.; Murnane, C.; Gray, E.; Humphries, M.; Sycamore, N.; Barker, D.; Saunders, D.; Wallis, J.; Babbage, A.; Hammond, S.; Mashreghi-Mohammadi, M.; Barr, L.; Martin, S.; Wray, P.; Ellington, A.; Matthews, N.; Ellwood, M.; Woodmansey, R.; Clark, G.; Cooper, J.; Tromans, A.; Grafham, D.; Skuce, C.; Pandian, R.; Andrews, R.; Harrison, E.; Kimberley, A.; Garnett, J.; Fosker, N.; Hall, R.; Garner, P.; Kelly, D.; Bird, C.; Palmer, S.; Gehring, I.; Berger, A.; Dooley, CM.; Ersan-Ürün, Z.; Eser, C.; Geiger, H.; Geisler, M.; Karotki, L.; Kirn, A.; Konantz, J.; Konantz, M.; Oberländer, M.; Rudolph-Geiger, S.; Teucke, M.; Lanz, C.; Raddatz, G.; Osoegawa, K.; Zhu, B.; Rapp, A.; Widaa, S.; Langford, C.; Yang, F.; Schuster, S. C.; Carter, N. P.; Harrow, J.; Ning, Z.; Herrero, J.; Searle, S. M.; Enright, A.; Geisler, R.; Plasterk, R. H.; Lee, C.; Westerfield, M.; de, Jong, P. J.; Zon, L. I.; Postlethwait, J. H.; Nüsslein-Volhard, C.; Hubbard, T. J.; Roest, Crollius, H.; Rogers, J.; Stemple, D. L. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **2013**, 496(7446), 498-503.

[6] Goessling, W.; Sadler, K. C. Zebrafish: an important tool for liver disease research. *Gastroenterology* **2015**, 149(6), 1361-77.

[7] Bambino, K.; Chu, J. Zebrafish in Toxicology and Environmental Health. *Curr. Top. Dev. Biol.* **2017**, 124, 331-367.

[8] Kimmel, C. B.; Ballard, W. W.; Kimmel, S. R.; Ullmann, B.; Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **1995**, 203(3), 253-310.

[9] Satija, R.; Farrell, J. A.; Gennert, D.; Schier, A. F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **2015**, 33(5), 495-502.

[10] Mathavan, S.; Lee, S. G.; Mak, A.; Miller, L. D.; Murthy, K. R.; Govindarajan, K. R.; Tong, Y.; Wu, Y. L.; Lam, S. H.; Yang, H.; Ruan, Y.; Korzh, V.; Gong, Z.; Liu, E. T.; Lufkin, T. Transcriptome analysis of zebrafish embryogenesis using microarrays. *PLoS Genet.* **2005,** 1(2), 260-76.

[11] Harvey, S. A.; Sealy, I.; Kettleborough, R.; Fenyes, F.; White, R.; Stemple, D.; Smith, J. C. Identification of the zebrafish maternal and paternal transcriptomes. *Development* 2013, 140(13),2703-10.

[12] Vesterlund, L.; Jiao, H.; Unneberg, P.; Hovatta, O.; Kere, J. The zebrafish transcriptome during early development. *BMC Dev. Biol.* **2011**, 11, 30.

[13] White, R. J.; Collins, J. E.; Sealy, I. M.; Wali, N.; Dooley, C. M.; Digby, Z.; Stemple, D. L.; Murphy, D. N.; Billis, K.; Hourlier, T.; Füllgrabe, A.; Davis, M. P.; Enright, A. J.; Busch-Nentwich, E. M. A high-resolution mRNA expression time course of embryonic development in zebrafish. *Elife* **2017**, 6, e30860.

[14] Rauwerda, H.; Pagano, J. F.; de, Leeuw, W. C.; Ensink, W.; Nehrdich, U.; de, Jong, M.; Jonker, M.; Spaink, H. P.; Breit, T. M. Transcriptome dynamics in early zebrafish embryogenesis determined by high-resolution time course analysis of 180 successive, individual zebrafish embryos. *BMC Genomics* **2017**, 18(1), 287.

[15] Koussounadis, A.; Langdon, S. P.; Um, I. H.; Harrison, D. J.; Smith, V. A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci. Rep.* **2015**, 5, 10775.

[16] Tay, T. L.; Lin, Q.; Seow, T. K.; Tan, K. H.; Hew, C. L.; Gong, Z. Proteomic analysis of protein profiles during early development of the zebrafish, Danio rerio. *Proteomics* **2006**, 6(10), 3176-88.

[17] Alli, Shaik, A.; Wee, S.; Li, R. H.; Li, Z.; Carney, T. J.; Mathavan, S.; Gunaratne, J. Functional mapping of the zebrafish early embryo proteome and transcriptome. *J. Proteome Res.* **2014**, 13(12), 5536-50.

[18] Kwon, O. K.; Kim, S. J.; Lee, Y. M.; Lee, Y. H.; Bae, Y. S.; Kim, J. Y.; Peng, X.; Cheng, Z.; Zhao, Y.; Lee, S. Global analysis of phosphoproteome dynamics in embryonic development of zebrafish (Danio rerio). *Proteomics* **2016**, 16(1), 136-49.

[19], Lucitt, M. B.; Price, T. S.; Pizarro, A.; Wu, W.; Yocum, A. K.; Seiler, C.; Pack, M. A.; Blair, I. A.; Fitzgerald, G. A.; Grosser, T. Analysis of the zebrafish proteome during embryonic development. *Mol. Cell. Proteomics* **2008**, 7(5), 981-94.

[20] Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R.; 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013,** 113(4), 2343-94.

[21] Lau, E.; Lam, M. P.; Siu, S. O.; Kong, R. P.; Chan, W. L.; Zhou, Z.; Huang, J.; Lo, C.; Chu, I. K. Combinatorial use of offline SCX and online RP-RP liquid chromatography for iTRAQ-based quantitative proteomics applications. *Mol. Biosyst.* **2011**, 7(5), 1399-408.

[22] Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The one hour yeast proteome. *Mol. Cell. Proteomics* **2014**, 13(1), 339-47.

[23] Yeung, D.; Mizero, B.; Gussakovsky, D.; Klaassen, N.; Lao, Y.; Spicer, V.; Krokhin, OV. Separation Orthogonality in Liquid Chromatography-Mass Spectrometry for Proteomic Applications: Comparison of 16 Different Two-Dimensional Combinations. *Anal. Chem.* **2020**, 92(5), 3904-3912.

[24] Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J. Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry as an alternative proteomics platform to ultraperformance liquid chromatography-electrospray ionization-tandem mass spectrometry for samples of intermediate complexity. *Anal. Chem.* **2012**, 84(3), 1617-22.

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

[26] Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem.* **2017**, 89(3), 2000-2008.

[27] Lombard-Banek, C.; Choi, S. B.; Nemes, P. Single-cell proteomics in complex tissues using microprobe capillary electrophoresis mass spectrometry. *Methods Enzymol.* **2019**, 628, 263-292.

[28] Lombard-Banek, C.; Moody, S. A.; Manzini, M. C.; Nemes, P. Microsampling Capillary Electrophoresis Mass Spectrometry Enables Single-Cell Proteomics in Complex Tissues: Developing Cell Clones in Live Xenopus laevis and Zebrafish Embryos. *Anal. Chem.* **2019**, 91(7), 4797-4805.

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

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

[31] Tyanova, S.; Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. *Methods Mol. Biol.* **2018**, 1711, 133-148.

[32] Rigbolt, K. T.; Vanselow, J. T.; Blagoev, B. GProX, a user-friendly platform for bioinformatics analysis and visualization of quantitative proteomics data. *Mol. Cell. Proteomics.* **2011**, 10(8), O110.007450.

[33] Kane DA, Kimmel CB. The zebrafish midblastula transition. *Development* **1993**, 119(2), 447-56.

[34] Schulz, K. N.; Harrison, M. M. Mechanisms regulating zygotic genome activation. *Nat. Rev. Genet.* **2019**, 20(4), 221-234.

[35] Jukam, D.; Shariati, S. A. M.; Skotheim, J. M. Zygotic Genome Activation in Vertebrates. *Dev. Cell.* **2017**, 42(4), 316-332.

[36] Hu H, Miao YR, Jia LH, Yu QY, Zhang Q, Guo AY. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019 Jan 8;47(D1):D33-D38. doi: 10.1093/nar/gky822. PMID: 30204897; PMCID: PMC6323978.

[37] Liu, Z.; Portero, E. P.; Jian, Y.; Zhao, Y.; Onjiko, R. M.; Zeng, C.; Nemes, P. Trace, Machine Learning of Signal Images for Trace-Sensitive Mass Spectrometry: A Case Study from Single-Cell Metabolomics. *Anal. Chem.* **2019**, 91(9), 5768-5776.

[38] Martin, C. C.; Tsang, C. H.; Beiko, R. G.; Krone, P. H. Expression and genomic organization of the zebrafish chaperonin gene complex. *Genome* **2002**, 45(5), 804-11.

[39] Lee, M. T.; Bonneau, A. R.; Takacs, C. M.; Bazzini, A. A.; DiVito, K. R.; Fleming, E. S.; Giraldez, A. J. Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. Nature 2013, 503(7476), 360-4.

[40] He, M.; Zhang, R.; Jiao, S.; Zhang, F.; Ye, D.; Wang, H.; Sun, Y. Nanog safeguards early embryogenesis against global activation of maternal β-catenin activity by interfering with TCF factors. *PLoS Biol.* **2020**, 18(7), e3000561.

[41] Schuff, M.; Siegel, D.; Philipp, M.; Bundschu, K.; Heymann, N.; Donow, C.; Knöchel, W. Characterization of Danio rerio Nanog and functional comparison to Xenopus Vents. *Stem Cells Dev.* **2012**, 21(8), 1225-38.

[42] Sun, J.; Yan, L.; Shen, W.; Meng, A. Maternal Ybx1 safeguards zebrafish oocyte maturation and maternal-to-zygotic transition by repressing global translation. *Development* **2018**, 145(19), dev166587.

[43] Wang, W.; Li, X.; Lee, M.; Jun, S.; Aziz, K. E.; Feng, L.; Tran, M. K.; Li, N.; McCrea, P. D.; Park, J. I.; Chen, J. FOXKs promote Wnt/β-catenin signaling by translocating DVL into the nucleus. *Dev. Cell* **2015**, 32(6), 707-18.

[44] Wang, B.; Zhang, X.; Wang, W.; Zhu, Z.; Tang, F.; Wang, D.; Liu, X.; Zhuang, H.; Yan, X. Forkhead box K2 inhibits the proliferation, migration, and invasion of human glioma cells and predicts a favorable prognosis. *Onco. Targets Ther.* **2018**, 11, 1067-1075.

[45] Du, F.; Qiao, C.; Li, X.; Chen, Z.; Liu, H.; Wu, S.; Hu, S.; Qiu, Z.; Qian, M.; Tian, D.; Wu, K.; Fan, D.; Nie, Y.; Xia, L. Forkhead box K2 promotes human colorectal cancer metastasis by upregulating ZEB1 and EGFR. *Theranostics*. **2019**, 9(13), 3879-3902.

[46] Essner, J. J.; Amack, J. D.; Nyholm, M. K.; Harris, E. B.; Yost, H. J. Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left-right development of the brain, heart and gut. *Development* **2005**, 132(6), 1247-60.

[47] Howard, T. L.; Stauffer, D. R.; Degnin, C. R.; Hollenberg, S. M. CHMP1 functions as a member of a newly defined family of vesicle trafficking proteins. *J. Cell Sci.* **2001**, 114(Pt 13), 2395-404.

[48] Clague, M. J.; Urbé, S. Multivesicular bodies. *Curr. Biol.* **2008**, 18(10), R402-R404.

[49] Stauffer, D. R.; Howard, T. L.; Nyun, T.; Hollenberg, S. M. CHMP1 is a novel nuclear matrix protein affecting chromatin structure and cell-cycle progression. *J. Cell Sci.* **2001**, 114(Pt 13), 2383-93.

[50] Zheng, X.; Dumitru, R.; Lackford, B. L.; Freudenberg, J. M.; Singh, A. P.; Archer, T. K.; Jothi, R.; Hu, G. Cnot1, Cnot2, and Cnot3 maintain mouse and human ESC identity and inhibit extraembryonic differentiation. *Stem Cells* **2012**, 30(5), 910-22.

[51] Ito, K.; Inoue, T.; Yokoyama, K.; Morita, M.; Suzuki, T.; Yamamoto, T. CNOT2 depletion disrupts and inhibits the CCR4-NOT deadenylase complex and induces apoptotic cell death. *Genes. Cells* **2011**, 16(4), 368-79.

[52] Vaquerizas, J. M.; Kummerfeld, S. K.; Teichmann, S. A.; Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev Genet.* **2009**, 10(4), 252-63.

[53] Boyadjiev, S. A.; Jabs, E. W. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin. Genet.* **2000**, 57(4), 253-66.

[54] Zhang, H. M.; Chen, H.; Liu, W.; Liu, H.; Gong, J.; Wang, H.; Guo, A. Y. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.* **2012**, 40(Database issue), D144-9.

[55] Tadros, W.; Lipshitz, H. D. The maternal-to-zygotic transition: a play in two acts. *Development* **2009**, 136(18), 3033-42.

[56] Stainier, D. Y. R.; Raz, E.; Lawson, N. D.; Ekker, S. C.; Burdine, R. D.; Eisen, J. S.; Ingham, P. W.; Schulte-Merker, S.; Yelon, D.; Weinstein, B. M.; Mullins, M. C.; Wilson, S. W.; Ramakrishnan, L.; Amacher, S. L.; Neuhauss, S. C. F.; Meng, A.; Mochizuki, N.; Panula, P.; Moens, C. B. Guidelines for morpholino use in zebrafish. PLoS Genet. **2017**, 13(10), e1007000.

[57] Bedell, V. M.; Westcot, S. E.; Ekker, S. C. Lessons from morpholino-based screening in zebrafish. *Brief Funct. Genomics.* **2011**, 10(4), 181-8.

[58] Xu, C.; Fan, Z. P.; Müller, P.; Fogley, R.; DiBiase, A.; Trompouki, E.; Unternaehrer, J.; Xiong, F.; Torregroza, I.; Evans, T.; Megason, S. G.; Daley, G. Q.; Schier, A. F.; Young, R. A.; Zon, L. I. Nanog-like regulates endoderm formation through the Mxtx2-Nodal pathway. *Dev. Cell* **2012**, 22(3), 625-38.

CHAPTER 5. Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics

Part of this chapter was adapted from Anal. Chem. 2020, 92(5), 3503-3507. with permission

5.1. Introduction

MS-based top-down proteomics as a strategy in proteomics directly measures intact proteins. Theoretically, top-down proteomics conserves all PTM combinations on proteins. Each unique combination between a protein and PTM(s) is called a proteoform. Top-down proteomics aims to delineate proteoforms in cells comprehensively with high confidence and throughput [1-5]. Proteoforms extracted from biological samples are typically separated by RPLC or CZE, followed by ESI-MS/MS. Database search is then performed for the ID of proteoform spectrum matches (PrSMs), proteoforms, and proteins through comparing experimental and theoretical masses of proteoforms and their fragments. To improve the confidence of proteoform ID, the target-decoy database search approach is typically employed [6,7], and the identified PrSMs and proteoforms were filtered by certain FDRs. Recently, the Kelleher group showed that the FDR estimation in top-down proteomics was complicated and the FDRs could be drastically under-reported [8]. High-confidence proteoform and protein IDs are vital. Therefore, after filtering the data with a specific FDR, we need to validate the data further using an alternative approach to the FDR.

The retention/migration time of proteoforms in LC/CZE can be useful information for improving the confidence of IDs. Some previous studies have deployed the

retention/migration time of proteins and peptides to facilitate their IDs [9-12]. We believe that accurate prediction of the retention/migration time of proteoforms will push the use of separation time for ID forward drastically. By comparing the experimentally observed and accurately predicted separation time of proteoforms, we could further boost the confidence of identified proteoforms, determine wrong proteoform IDs, and even provide useful information to correct proteoform IDs.

Some work has been done in predicting migration time (electrophoretic mobility, μ_{ef}) of peptides from CZE separations [13-21]. It has been demonstrated that CZE outperformed RPLC regarding the prediction of migration/retention time of peptides for bottom-up proteomics [21]. One major reason is that the size and charge of peptides for CZE can be calculated relatively easily, by contrast, the interaction between peptides and beads for RPLC is complicated [21]. Krokhin *et al.* achieved a linear correlation (R^2=0.995) between predicted and experimental μ_{ef} of peptides in CZE using a large peptide dataset and an optimized semi-empirical model [21], which was based on the model reported by Cifuentes *et al.* [19]. More recently, we also applied a similar model for predicting the μ_{ef} of phosphorylated peptides and achieved a high correction (R^2=0.99) between the predicted and experimental μ_{ef} for mono-phosphorylated peptides from the HCT116 cell line [22].

Great success has been achieved for predicting μ_{ef} of peptides, but much more effort need to be made on proteins/proteoforms. Some initial effort has been made using a handful of standard proteins [17,23,24]. However, there is no report about predicting μ_{ef} of proteins/proteoforms using large-scale proteoform datasets. There are two major reasons for that. First, large-scale top-down proteomics datasets from CZE-MS have been limited. Second, proteins/proteoforms are much larger than peptides, leading to potential difficulties in calculating their size and charge

accurately. In the last 5 years, CZE-MS has been recognized as an important approach for large-scale top-down proteomics due to the improvement in CE-MS interfaces, capillary coatings, and online sample stacking techniques [25-32]. For instance, we identified nearly 600 proteoforms from an *E. coli* cell lysate in a single-shot CZE-MS/MS analysis [27]. In that study, we employed a commercialized electro-kinetically pumped sheath-flow CE-MS interface [33,34], a 1-meter-long linear polyacrylamide (LPA)-coated capillary [35], and a dynamic pH junction-based proteoform stacking method [36] to boost the sample loading capacity, separation window, and overall sensitivity of the CZE-MS system. In another study, we used a 1.5-meter-long LPA-coated capillary for CZE-MS/MS analysis of zebrafish brains and identified thousands of proteoforms in a single analysis with consumption of nanograms of protein material [29]. These large-scale proteoform datasets provide us great opportunities to push forward the prediction of μ_{ef} of proteoforms, which will be useful for improving the confidence of proteoform IDs in top-down proteomics.

Here, we applied previously reported semi-empirical mobility models in the prediction of proteoforms' μ_{ef} and evaluated their performance using large proteoform datasets from *E. coli* cells and zebrafish brains under different CZE conditions. We achieved a linear correlation between experimental and predicted μ_{ef} of *E. coli* proteoforms ($R^2 = 0.98$) with a simple semiempirical model, which utilizes the number of charges and molecular mass of each proteoform as the parameters. Our modeling data suggest that the complete unfolding of proteoforms during CZE separation benefits the prediction of their μ_{ef}. Optimization of the prediction model on histone proteoforms highlights the influence of charge suppression on the performance of the model. Our results also indicate that N-terminal acetylation and phosphorylation both decrease the proteoforms' charge by roughly one charge unit.

5.2. Experimental

5.2.1. Material and reagents

Unless stated otherwise, all reagents were purchased from Sigma-Aldrich (St. Louis, MO). Acrylamide was purchased from Acros Organics (NJ, USA). Fused silica capillaries (50/360 µm i.d./o.d.) were bought from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail in EASYpacks was purchased from Roche (Indianapolis, IN). LC-MS grade water, ACN, methanol, FA, and AA were obtained from Fisher Scientific (Pittsburgh, PA). Lyophilized calf thymus histone extract was obtained from Sigma-Aldrich (St. Louis, MO).

5.2.2. Sample preparation

A detailed zebrafish brain experiment was described in reference [29]. *Escherichia coli* (*E. coli*, strain K-12 substrain MG1655) was cultured in Luria-Bertani (LB) medium at 37 °C with 225 rpm shaking. *E. coli* cells were harvested by centrifugation at 4000 rpm for 10 min when OD600 reached 0.7. The cells were washed using PBS for 3 times. Cell lysis was performed in lysis buffer containing 8 M urea, 100 mM Tris-HCl (pH 8.0), and protease inhibitors by sonication using a Branson Sonifier 250 (VWR Scientific, Batavia, IL). The supernatant was collected after centrifugation at 18000 g for 10 min. The BCA assay was used for the measurement of protein concentration. The extracted proteins were stored at -80 °C before use.

One milligram of *E. coli* sample in lysis buffer was denatured at 37 °C for 30 min. The reduction was performed by adding 1.7 µL of 1 M DTT in 100 mM NH_4HCO_3

followed by incubation at 37 °C for 30 min. Alkylation was performed by adding 4.0 µL of 1 M IAA in 100 mM NH_4HCO_3 solution followed by incubation in dark for 20 min. Then the proteins were desalted with a C4-trap column (Bio-C4, 3 µm, 300 Å, 4.0 mm i.d., 10 mm long) from Sepax Technologies, Inc. (Newark, DE) in an HPLC system (Agilent Technologies, 1260 Infinity II). The proteins were eluted with 80% ACN, followed by lyophilization with a vacuum concentrator (Thermo Fisher Scientific). The dried samples were dissolved in 50 mM of NH_4HCO_3 (pH 8.0) with a 2 mg/mL final concentration for the CZE-MS/MS analyses.

5.2.3. SEC fractionation

A 4.6 mm i.d. × 50 mm length SEC guard column (Bio SEC-3, guard, Agilent) and a 4.6 mm i.d. × 300 mm length SEC separation column (Bio SEC-3, Agilent) were coupled for SEC fractionation. An Agilent Infinity II HPLC system was employed for the experiment. Mobile phase was 0.1 % FA in water.

Roughly 500 µg histone extract (10 µg/µL) was dissolved in mobile phase and then injected into the SEC columns for separation. The flow rate was set to 0.1 mL/min. Total 13 fractions were collected from 17 to 43 min (1 fraction/ 2 min). Then a small proportion of each fraction was subjected to the BCA assay for the measurement of protein concentration. 20 µL of each fraction was lyophilized and re-dissolved in 30% (v/v) ACN in water to roughly 1 µg/µL. Fractions 1 and 2, fractions 12 and 13 were combined into two fractions due to low protein concentration. Eventually, 11 fractions were subjected to CZE-MS/MS analysis.

5.2.4. CZE-ESI-MS/MS analysis

An ECE-001 CE autosampler (CMP scientific, Brooklyn, NY) was used for sample injection and separation. The autosampler was coupled to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific) through a third-generation electrokinetically pumped sheath flow CE-MS interface (CMP Scientific, Brooklyn, NY). A fused silica capillary (50/360 µm i.d./o.d., 103 cm long) was coated with LPA based on reference [35]. Then one end of the capillary was etched with hydrofluoric acid. There were three kinds of BGEs used in this work: 5% (v/v) AA in water, 20% (v/v) AA in water, and 20% (v/v) AA in water containing 10% (v/v) isopropanol (IPA) and 15% (v/v) dimethylacetamide (DMA). Sheath buffer was 0.2% FA with 10% methanol. For *E. coli* sample, the pressure of 5 psi was employed for sample injection, approximate 400 nL of samples were injected for each run. For the histone sample, 5 psi was applied for 5 s for each sample. Based on Poiseuille's law, approximately 25 nL of each sample was injected. +30 kV was applied for separation, and +2 kV was applied in the sheath buffer reservoir to generate electrospray. The ESI emitter was pulled from borosilicate glass capillaries (0.75/1.00 mm i.d./o.d.,10 cm long) with a Sutter P-1000 flaming/brown micropipette puller. The opening size of the emitter was 30-40 µm.

For *E. coli* sample, a top 5 DDA mode was used on the Q Exactive HF mass spectrometer. For MS, a mass resolution of 240,000 at m/z 200 was used, microscans was set to 3, AGC target was set to 1E6, maximum injection time was 50 ms and the scan range was from 600 to 2500 m/z. For MS/MS, we used 120,000 mass resolution, 3 microscans, 1E5 intensity threshold, 200 ms maximum injection time, 4 m/z isolation window, 20% normalized collision energy. Only ions with charge

state higher than 2 were selected for HCD fragmentation. The dynamic exclusion was set to 30 s.

For histone sample, a Top 3 DDA method was used. For MS, the resolution, AGC target, and maximum injection time were set to 120,000, 1e6 and 50 ms, respectively. The scan ranges of MS and MS/MS were 400-1500 and 200-2000 m/z, respectively. For MS/MS, the resolution, AGC target, and maximum injection time were set to 60,000, 1e6 and 400 ms, respectively. The ions for MS/MS were isolated in the quadrupole with an isolation window of 2 m/z, followed by fragmentation employing a stepped HCD method with three-step normalized collision energy as 12%, 16% and 20%. Dynamic exclusion was set to 50 s. Ions with charge states lower than 7 were excluded for the fragmentation.

5.2.5. Data analysis

Raw files of the *E. coli* sample were analyzed by the TopPIC (TOP-down mass spectrometry based proteoform identification and characterization) suite for proteoform identification [37, 38]. Raw files were first converted to mzML files using Msconvert tool. The mzML files were then analyzed by TopFD, a spectral deconvolution and msalign file generating tool. The msalign files were then searched by TopPIC against an E. coli database (UP000000625) and a zebrafish database (AUP000000437) downloaded from UniProt. Cysteine carbamidomethylation was set as a fixed modification. Error tolerances of precursor and fragment ions were 15 ppm. The maximum number of mass shift was 2. The maximum mass shift of unknown modifications ranged from -500 to 500 Da. The FDRs were estimated by the target-decoy database search approach [6,7]. A 0.1% spectrum-level FDR and a 0.5% proteoform-level FDR were employed to filter the data.

Proteome Discoverer 2.4 sp software (Thermo Fisher Scientific) with the ProSight PD top-down nodes was used for database search of histone sample. The MS1 spectra were first averaged using the cRAWler algorithm in Proteome Discoverer. The precursor m/z tolerance was set to 0.2 m/z. For both precursor and fragmentation Xtract parameters, the signal-to-noise ratio threshold, the lowest and the highest m/z were set to 3, 200 and 2000, respectively. Then deconvolution was performed by the Xtract algorithm followed by database search against a Bos Taurus downloaded from http://proteinaceous.net/-database-warehouse-legacy/ in April 2018. A three-prone database search was performed: (1) a search was performed with a 10-ppm mass tolerance of absolute mass for both MS1 and MS2; (2) a search was performed with a 200-Da mass tolerance for MS1 and a 10-ppm mass tolerance for MS2 for matching unexpected PTMs; (3) a subsequent search was performed to find unreported truncated proteoforms with 10 ppm tolerance for both MS1 and MS2. The target-decoy strategy was employed for evaluating the FDRs. For a possible ID, FDR estimation was performed for each of three search strategies. Proteoform IDs were filtered first using single FDR threshold from each search, and then by the global FDR estimated by the best q-value in three searches. The identified proteoform-spectrum matches (PrSMs), proteoforms and proteins were filtered using a 1% FDR.

To ensure confident identification of histone proteoforms, we applied a C-score filter to all PrSMs and proteoforms (C-score >3). The C-Score (Characterization Score) is used to indicate how well proteoforms are characterized (*e.g.,* location of PTMs). A higher C-Score indicates better proteoform characterization. Proteoforms with C-Scores higher than 3 are confidently identified and partially characterized; Proteoforms with C-Scores higher than 40 are fully characterized. Because the

backbone cleavage coverage of identified histone proteoforms are limited in this work and it is challenging to accurately assign and localize PTMs on proteoforms solely based on the matched fragment ions, all the PTM assignments and localization in this work are tentative.

5.2.6. Calculation of μ_{ef}

The experimental μ_{ef} was calculated using Equation (1),

$$\text{Experimental } \mu_{ef} = \frac{L^2}{(30-2)\times t_m} \text{ (unit of cm}^2\text{ kV}^{-1}\text{s}^{-1}\text{)} \quad (1)$$

Where L is the capillary length in cm, t_M is the migration time in s. The 30 and 2 are separation voltage and electrospray voltage in kilovolts.

The predicted μ_{ef} of proteoforms was calculated using six classical semi-empirical models [14-16,18-20], **Table 5.1**. The number of charges (Q) of each proteoform equals the number of positively charged amino acid residues within their sequences (K, R, H, and N-terminus). The molecular mass (M) of each proteoform equals the adjusted mass reported by the TopPIC. The length (N) of each proteoform equals the number of amino acid residues within the sequence. Only proteoforms without post-translational modifications (PTMs) were used for the calculation of experimental μ_{ef} and predicted μ_{ef}.

The calculation of electrophoretic mobility relative difference (EMRD) was performed using the following equation,

$$EMRD = \frac{\text{experimental } \mu_{ef} - \text{predicted } \mu_{ef}\text{ (abs)}}{\text{predicted } \mu_{ef}\text{(abs)}} \quad (2)$$

Where the predicted μ_{ef} (abs) is the absolute value of predicted μ_{ef} which can be calculated using the following equation based on the correlation equation shown in **Figure 5.3B**,

$$predicted\ \mu_{ef}\ (abs) = \frac{\text{predicted } \mu_{ef} - 0.016}{0.11} \quad (3)$$

With the assumption that our prediction model can accurately predict the μ_{ef} of unmodified proteoforms, the absolute predicted μ_{ef} of each modified proteoform calculated solely based on the proteoform sequence equals the μ_{ef} of its unmodified counterpart. Therefore, the EMRD represents the relative difference in μ_{ef} between the modified and unmodified proteoforms with the same protein sequence.

5.3. Results and discussion

5.3.1. Performances of different semiempirical models in predicting μ_{ef} of large-scale proteoform IDs.

To evaluate the performances of six semiempirical models in predicting proteoforms' μ_{ef}, we generated a large-scale *E. coli* proteoform dataset using CZE-MS/MS for this project. Technical triplicates were performed for each of three different BGEs. About 500-1100 non-modified proteoforms were used for the μ_{ef}

calculations. The molecular mass of proteoforms ranged from 1.5 kDa to 30 kDa. We also assumed that the electroosmotic flow (EOF) in an LPA-coated capillary with an acidic BGE was extremely low [27].

The results of semiempirical models are shown in Table 5.1. For Cifuentes's model, we obtained the final equation (5) based on the original equation (4) in ref. [19] via omitting the prefactor 900.

$$\mu_{ef} = 900 \times \frac{\ln(1+0.35 \times Q)}{M^{0.411}} \quad (4)$$

$$\mu_{ef} = \frac{\ln(1+0.35 \times Q)}{M^{0.411}} \quad (5)$$

Where Q is the number of charges and M is the molecular mass of each proteoform.

Table 5.1. Summary of the linear correlations between experimental μef and predicted μef of E. coli proteoforms using different semi-empirical models and under various CZE conditions.*

Semi-empirical model		BGE					
		5% (v/v) AA		20% (v/v) AA		10% (v/v) IPA 15% (v/v) DMA 20% (v/v) AA	
		R^2	Slope	R^2	Slope	R^2	Slope
$\ln(1+0.35*Q)/M^{0.411}$	**Cifuentes and Poppe [19,21]**	**0.97**	**0.22**	**0.98**	**0.26**	**0.98**	**0.51**
$\ln(1+Q)/N^{0.435}$	Grossman *et al.* [18]	0.76	1.72	0.82	2.1	0.82	4.4
$Q/M^{2/3}$	Offord [14]	0.93	0.25	0.94	0.29	0.92	0.58
$Q/M^{0.56}$	Kim *et al.* [16]	0.90	0.65	0.89	0.74	0.82	1.4
$Q/M^{1/2}$	Tanford [15]	0.86	1.1	0.84	1.2	0.74	2.3
$Q/M^{1/3}$	Reynolds *et al.* [20]	0.72	4.6	0.69	5.2	0.52	9.0

* Only proteoforms without PTMs were used. The R^2 and slope values were from the mean of the triplicate CZE-MS/MS runs, and the standard deviations of the R^2 values from the triplicate analyses were about 0.01.

The Cifuentes's model produced the best linear correlation (R^2: 0.97-0.98) between the predicted and experimental μ_{ef} of proteoforms according to the R^2 values for the three CZE conditions, followed by the Offord's model (R^2: 0.92-0.94) and Kim's model (R^2: 0.82-0.90). The Reynolds's model generated the lowest correlation coefficient (R^2: 0.52-0.72). The Cifuentes's model obtained a drastically better linear correlation regarding the R^2 value than the Grossman's model (0.97 *vs.*

0.76 for the 5%AA BGE) and the two models have two differences, $M^{0.411}$ *vs.* $N^{0.435}$ and 0.35*Q *vs.* Q. After a more detailed study using the 5%AA BGE data, we figured out that the R^2 value of the Grossman's model could be boosted from 0.76 to 0.94 by simply changing the Q to 0.35*Q. Only a minor effect on the R^2 value was observed by changing $N^{0.435}$ to $M^{0.411}$. We note that the slopes of the linear correlation curves from the two best models (the Cifuentes's model and the Offord's model) are comparable for the different CZE conditions, *e.g.,* 0.22 *vs.* 0.25 for the 5%AA BGE, and are obviously smaller than that from other models, suggesting that the predicted μ_{ef} from these two models are much smaller than that from other four models and significantly smaller than the experimental μ_{ef}. We can add a CZE condition-dependent prefactor to the Cifuentes's model to match the predicted and experimental μ_{ef}.

The data here represents the first try of predicting μ_{ef} of proteoforms using large-scale top-down proteomics datasets. The great correlation between experimental μ_{ef} and predicted μ_{ef} from the simple Cifuentes's model further implies that the μ_{ef} of proteoforms in CZE can be predicted easily. The predicted μ_{ef} of proteoforms discussed in the following parts were obtained from the Cifuentes's model.

5.3.2. Evaluation of the influence of BGE to μ_{ef} prediction

We evaluated how the BGE of CZE influenced the μ_{ef} of proteoforms, **Figure 5.1**. When the AA concentration in BGE increased from 5% to 20% and when 10% (v/v) IPA and 15% (v/v) DMA were added into the BGE, the experimental μ_{ef} of proteoforms decreased. Two possible reasons exist for that phenomenon. First, the lower pH of 20% (v/v) AA and the organic solvents unfold the proteoforms more completely, enlarging the size of proteoforms and reducing their mobility. It has been

reported recently that in CZE protein size can increase significantly due to unfolding when the pH of BGE decreases [39]. Second, the lower pH of 20% (v/v) AA and the organic solvents further eliminate the residual EOF in the capillary. In addition, when 20% (v/v) AA with or without 10% (v/v) IPA and 15% (v/v) DMA was used as the BGE, a better linear correlation was observed compared to the 5% (v/v) AA (0.98 *vs.* 0.96). For the BGE containing 20% (v/v) AA, 10% (v/v) IPA, and 15% (v/v) DMA, the absolute value of predicted μ_{ef} is much closer to that of experimental μ_{ef} compared to the other two BGEs, indicated by the much larger slope of the linear correlation curve (0.51 *vs.* 0.20-0.25). The number of outliers from the BGE containing IPA and DMA is also much smaller compared to the other BGEs. The results suggest that adding some organic solvents to the BGE of CZE could benefit the prediction of μ_{ef} of proteoforms. There is also some evidence in the literature. For instance, in 2000, Katayama *et al.* demonstrated that the use of methanol in BGE could improve the correlation between predicted μ_{ef} and experimental μ_{ef} of peptides [40]. We speculate that the organic solvents (IPA and DMA) in the BGE facilitate the complete unfolding of proteoforms, leading to better prediction of their μ_{ef}. It has been reported that certain types of polar solvents such as dimethyl sulfoxide (DMSO), dimethylformamide (DMF), and formamide have the ability to unfold proteins [41,42].

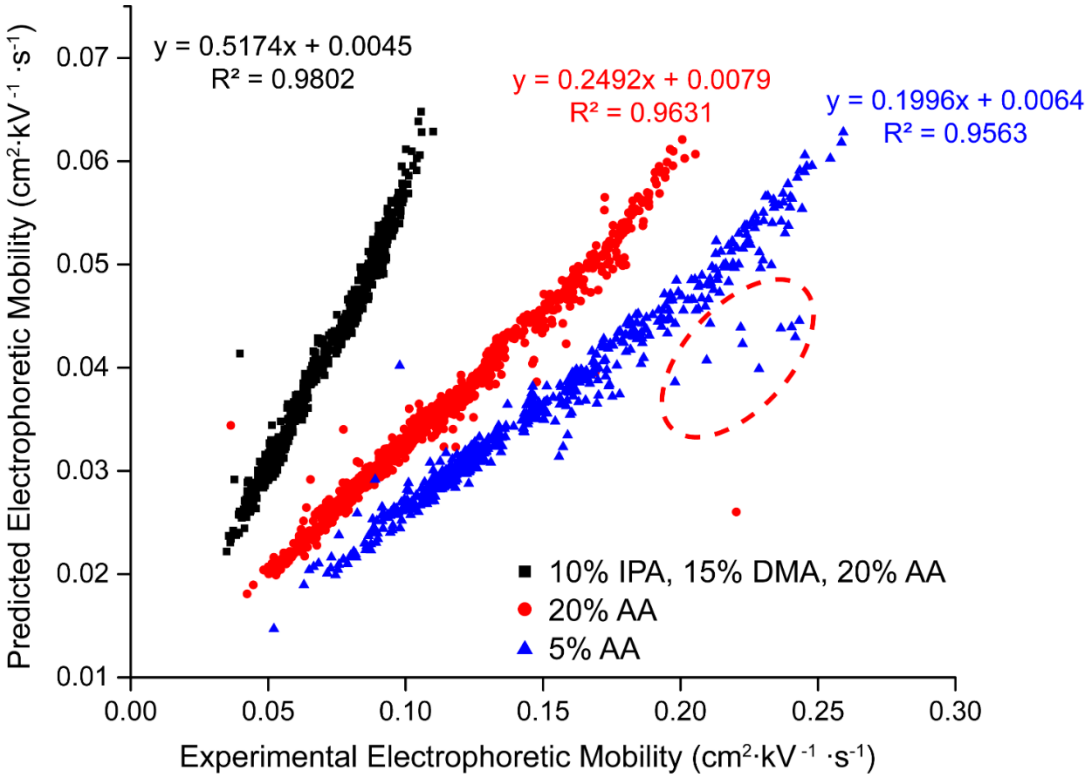

Figure 5.1. Linear correlations between predicted μ_{ef} and experimental μ_{ef} of proteoforms from *E. coli* cells under various CZE conditions. Only nonmodified proteoforms were used, and the data was from a single CZE-MS/MS run.

5.3.3. Performance of predicting μ_{ef} of proteoforms with certain PTMs

We then tested the Cifuentes's model on our published zebrafish brain (optic tectum (Teo)) data and evaluated the performance of the model for predicting μ_{ef} of proteoforms with certain PTMs (*i.e.,* N-terminal acetylation and phosphorylation) [29]. When we only used nonmodified proteoforms, the predicted μ_{ef} and experimental μ_{ef} showed reasonably good linear correlations (R^2=0.96). We then further included the proteoforms with N-terminal acetylation and/or phosphorylation in the analysis. The zebrafish Teo data from one CZE-MS/MS run was used, which included 1163 nonmodified proteoforms, 92 proteoforms with only N-terminal acetylation, 3 proteoforms with one phosphorylation site, and 2 proteoforms with both N-terminal acetylation and one phosphorylation site. N-terminal acetylation and phosphorylation can reduce the proteoforms' charge by one charge unit in theory. **Figure 5.2A** shows the linear correlation between the experimental and predicted μ_{ef} for these post-translationally modified proteoforms (97 in total) regardless of the PTMs. First, the linear correlation is poor (R^2=0.76). Second, it is clear that the addition of one acetylation modification or one phosphoryl group to a proteoform can decrease its mobility significantly. After considering the effect of these PTMs on the proteoforms' charge, we corrected the charge (Q) in the Cifuentes's model. We achieved a linear correlation for the 97 proteoforms with PTMs (R^2=0.92) after we adjusted the Q by -1, -1 and -2 for proteoforms with N-terminal acetylation, proteoforms with one phosphorylation site, and proteoforms with both N-terminal acetylation and phosphorylation, respectively, **Figure 5.2B**. The results show that the proteoforms' charge shifts are very close to the theoretical contributions of N-terminal acetylation and phosphorylation. Additionally, the results suggest that the μ_{ef} of proteoforms with N-terminal acetylation and phosphorylation could be predicted as accurately as

nonmodified proteoforms (R^2 0.92 *vs.* 0.96). We note that some outliers exist in **Figure 5.2B** due to two possible reasons. First, for these outliers, their experimental μ_{ef} values are larger than the predicted values, most likely due to the incomplete unfolding of these proteoforms in the BGE used in the experiment (10% (v/v) AA, pH 2.2). Second, since the proteoform IDs were filtered by a 0.5% FDR, some of the outliers could be simply the wrong proteoform IDs.

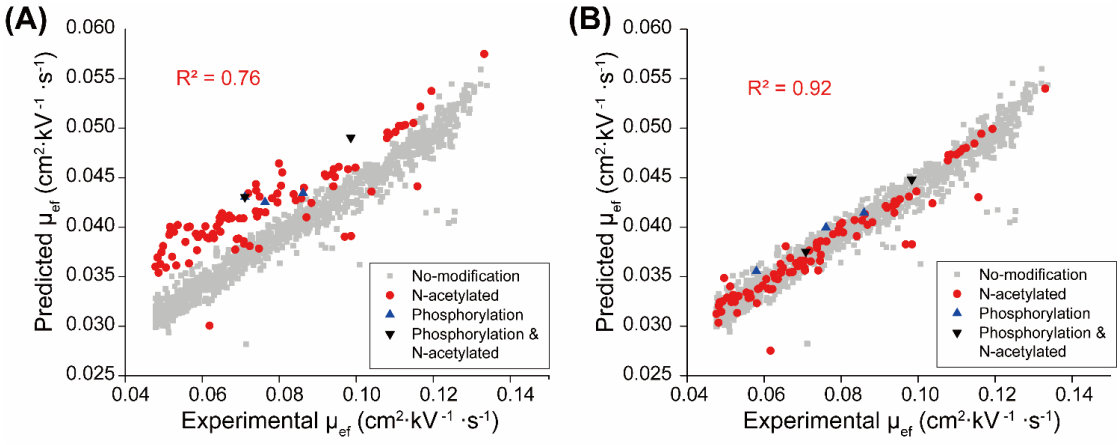

Figure 5.2. Linear correlations between predicted μ_{ef} and experimental μ_{ef} of proteoforms from zebrafish optic tectum (TEO). Nonmodified, N-terminal acetylated, and mono-phosphorylated proteoforms were employed. In (A), the charge of proteoforms in the BGE (Q) was calculated by counting the positively charged amino acid residues (K, R, H, and N-terminal) regardless of the PTMs. In (B), the charge of proteoforms (Q) was corrected based on their PTMs. For example, one charge reduction corresponded to one N-terminal acetylation or one phosphorylation.

5.3.4. Electrophoretic mobility prediction of histone proteoforms

To test the performance of the prediction model on highly charged proteins, we coupled SEC fractionation to CZE-MS/MS for analysis of the calf histone sample to boost the number of histone proteoform IDs. We fractionated the histone sample into 11 fractions using SEC. Then each fraction was analyzed by CZE-MS/MS. The SEC-CZE-MS/MS identified 48 proteins, 405 proteoforms, and 7832 PrSMs. The data were filtered with 1% FDRs and C-Score better than 3. Among the 405 proteoforms,

173 had C-score higher than 40 (43%), 390 were histone proteoforms and 15 were proteoforms of other proteins.

We evaluated the model in predicting μ_{ef} of histone proteoforms. Only proteoforms or PrSMs having C-score higher than 3, -log P-value higher than 11.5, and mass error less than 20 ppm were used for μ_{ef} calculation. As illustrated in **Figure 5.3A**, the correlation coefficient (R^2) between the predicted and experimental μ_{ef} of unmodified histone proteoforms was 0.96, which is lower than that in our previous report (0.98), **Table 5.1**. The prediction model, as shown in the eq (5), contains two terms: the charge, represented by the logarithmic value of the prefactored charge (Q) of a proteofrom, and the size, represented by the value of the proteoform's mass (M) to the power of 0.411. To improve the correlation between the predicted and experimental μ_{ef} of histone proteoforms, we optimized the prediction model via independent adjustments of the size and charge terms. Little improvement was observed by adjusting the size term. Interestingly, by reducing the prefactor of Q from 0.350 to 0.233, the correlation coefficient (R^2) was improved to 0.98, **Figure 5.3B**. The prefactor of Q was originally introduced to compensate the charge suppression, which resulted from mutual electrostatic interactions of the charged groups in the peptides/proteins [13]. The charge suppression becomes more significant in highly charged peptides/proteins because when the total charge of an analyte increases, the effect of any additional charge on its μ_{ef} decreases [18,43]. Histones are highly charged in acidic BGE, so that the charge suppression in the enriched histone sample ought to be strong, leading to the need for a smaller prefactor of Q for accurate prediction of μ_{ef}. Unless stated otherwise, the eq (6) is used for predicting the μ_{ef} of histone proteoforms in the work.

Predicted $\mu_{ef} = \frac{\ln(1+0.233\times Q)}{M^{0.411}}$ (6)

We then evaluated the influence of PTMs on the charge and μ_{ef} of histone proteoforms. Certain PTMs such as N-terminal acetylation and phosphorylation reduce the charge (Q) of peptides/proteins by roughly one charge unit as demonstrated in **Figure 3.8** and in **Figure 5.2**, thus decreasing their μ_{ef} significantly. To perform quantitative evaluation on the influence of PTMs to μ_{ef} of histone proteoforms, we defined the electrophoretic mobility relative difference (EMRD), a parameter that represents the influence of a PTM to the μ_{ef} of a proteoform. The detailed calculation of EMRD is in the experimental section (eq (2) and eq (3)). In theory, the EMRDs of N-terminal acetylated histone proteoforms tend to be less than 0 compared to the unmodified proteoforms due to the fact that the PTM reduces the number of positive charges of proteoforms in CZE. A dataset containing 89 histone proteoforms, of which 48 were unmodified and 41 had only N-terminal acetylation, was used. As shown in **Figure 5.3C**, the EMRDs of unmodified histone proteoforms centers around zero as expected. The median of EMRDs of N-terminal acetylated histone proteoforms is below 0, indicating that N-terminal acetylation indeed has negative effect on the μ_{ef} of histone proteoforms.

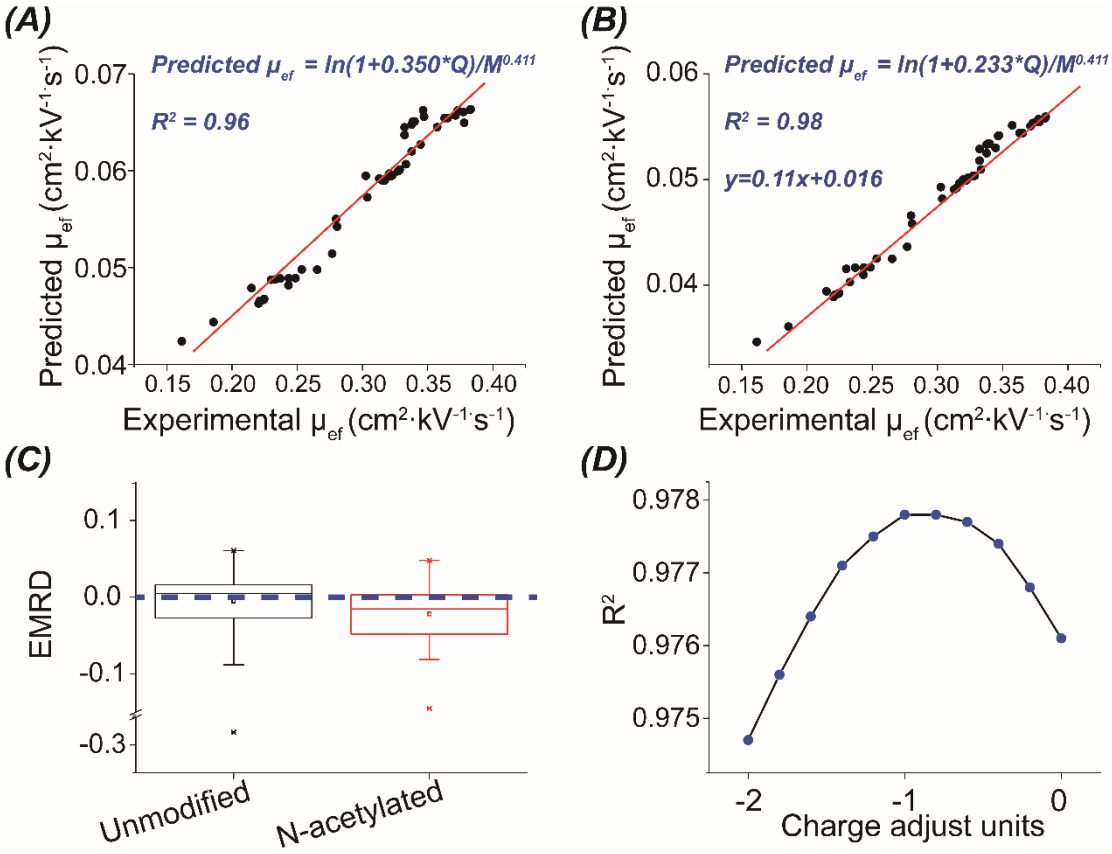

Figure 5.3. Correlations between predicted μ_{ef} and experimental μ_{ef} of unmodified proteoforms of calf histones before the model optimization using a pre-factor of 0.350 to Q (A) and after model optimization using a pre-factor of 0.233 to Q (B). (C) Box plots of EMRDs of unmodified and N-terminal acetylated proteoforms. (D) The R^2 values between predicted and observed μ_{ef} when different charge adjustment was made to N-terminal acetylated proteoforms.

We further evaluated the charge reducing effect of N-terminal acetylation on histone proteoforms. A high R^2 value of 0.976 was gained between the observed and the predicted μ_{ef} of all proteoforms in the dataset even without any charge adjustment (charge adjustment unit = 0) to the N-terminal acetylated proteoforms, suggesting that the μ_{ef} difference between a N-terminal acetylated histone proteoform and its unmodified counterpart is small due to histones' high charge-to-size ratios, **Figure 5.3D**. Because histones are highly charged in the 5% (v/v) AA (pH 2.4) BGE, the changes in charge (Q) caused by the N-terminal acetylation have limited overall impact on their charge-to-size ratios in our CZE conditions, leading to small effect on their μ_{ef}. Interestingly, when we reduced the positive charge (Q) of N-

terminal acetylated histone proteoforms stepwise, we observed the best R^2 value when the charge (Q) was reduced by 1 unit, **Figure 5.3D**. The data here highlight the challenges of achieving baseline separations of different proteoforms of one histone protein using our CZE condition. We need to note that the CZE conditions (*e.g.,* BGE composition) can modulate the μ_{ef} of histones substantially as demonstrated by the works from the Zhong group [44,45]. We expect that systematic optimizations of the CZE condition could amplify the effect of the PTMs (*e.g.,* N-terminal acetylation) on histone proteoforms' μ_{ef}, thus leading to better separations of modified and unmodified histone proteoforms.

5.4. Conclusions

In summary, in this work, for the first time, we evaluated various semi-empirical models for predicting proteoforms' μ_{ef} using large-scale top-down proteomics datasets. Using a simple semi-empirical model, we achieved a linear correlation between experimental μ_{ef} and predicted μ_{ef} of *E. coli* proteoforms and histone proteoforms (R^2=0.98). We note that some effort has been made on predicting retention time of proteins in RPLC using simple protein mixtures based on complicated models, producing reasonable correlations between predicted and experimental retention time (R^2=0.86-0.90) [11,46,47]. We also note that our current study still has some limitations. First, the proteoforms used in this study have masses lower than 30 kDa. Top-down proteomics datasets of large proteoforms using CZE-MS/MS are required to expand the model into a wider range of proteoforms in mass. Second, the number of proteoforms with PTMs (*i.e.,* acetylation and phosphorylation) used here is small, less than 100. Larger numbers of

proteoforms with PTMs are extremely important for improving the model for post-translationally modified proteoforms.

5.5. 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 the support from the National Science Foundation (CAREER Award, DBI-1846913) and the National Institutes of Health (R01GM125991).

REFERENCES

REFERENCES

[1] Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem*. **2016**, 9(1), 499-519.

[2] Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** 480(7376)*,* 254-8.

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

[4] Smith, L. M.; Kelleher, N. L.; Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013,** 10(3)*,* 186-7.

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

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

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

[8] LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Shams, D. P.; Thomas, P. M.; Kelleher, N. L. Accurate Estimation of Context-Dependent False Discovery Rates in Top-Down Proteomics. *Mol Cell Proteomics*. **2019**, 18(4), 796-805.

[9] Henneman, A. A.; Palmblad, M. Retention time prediction and protein identification. *Methods Mol. Biol*. **2013**, 1007, 101-18.

[10] Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G, 2nd.; Smith, R. D. Application of Peptide LC Retention Time Information in a Discriminant Function for Peptide Identification by Tandem Mass Spectrometry. *J. Proteome Res*. **2004**, 3(4), 760-9.

[11] Tarasova, I. A.; Masselon, C. D.; Gorshkov, A. V.; Gorshkov, M. V. Predictive chromatography of peptides and proteins as a complementary tool for proteomics. *Analyst* **2016,** 141(16), 4816-4832.

[12] Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. Proteomics **2002,** 2(5), 513-23.

[13] Mittermayr, S.; Olajos, M.; Chovan, T.; BonnA, G. K.; Guttman, A. Mobility modeling of peptides in capillary electrophoresis. *Trends Analyt. Chem*. **2008,** 27(5), 407-417.

[14] Offord, R. E. Electrophoretic mobilities of peptides on paper and their use in the determination of amide groups. *Nature* **1966,** 211(5049), 591-3.

[15] Tanford, C. Physical Chemistry of Macromolecules; Wiley: New York, U.S.A., **1961**.

[16] Kim, J.; Zand, R.; Lubman, D. M. Electrophoretic mobility for peptides with post-translational modifications in capillary electrophoresis. *Electrophoresis.* **2003**, 24, 782–793.

[17] Compton, B. J. Electrophoretic mobility modeling of proteins in free zone capillary electrophoresis and its application to monoclonal antibody microheterogeneity analysis. *Journal of Chromatography A*. **1991**, 559(1-2), 357-366.

[18] Grossman, P. D.; Colburn, J. C.; Lauer, H. H. A Semiempirical Model for the Electrophoretic Mobilities of Peptides in Free-Solution Capillary Electrophoresis. *Anal. Biochem*. **1989,** 179, 28–33.

[19] Cifuentes, A.; Poppe, H. Simulation and optimization of peptide separation by capillary electrophoresis. *J. Chromatogr. A* **1994**, 680(1), 321-40.

[20] Adamson, N. J.; Reynolds, E. C. Rules relating electrophoretic mobility, charge and molecular size of peptides and proteins. *J. Chromatogr., Biomed. Appl*. **1997**, 699, 133– 147.

[21] Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem*. **2017**, 89, 2000–2008.

[22] Chen, D.; Ludwig, K. R.; Krokhin, O. V.; Spicer, V.; Yang, Z.; Shen, X.; Hummon, A. B.; Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Large-Scale Phosphoproteomics with the Production of over 11,000 Phosphopeptides from the Colon Carcinoma HCT116 Cell Line. *Anal. Chem*. **2019,** 91(3), 2201-2208.

[23] Chae, K. S.; Lenhoff, A. M. Computation of the electrophoretic mobility of proteins. *Biophys. J*. **1995**, 68(3), 1120–1127.

[24] Rickard, E. C.; Strohl, M. M.; Nielsen, R. G. Correlation of Electrophoretic Mobilities from Capillary Electrophoresis with Physicochemical Properties of Proteins and Peptides. *Anal Biochem*. **1991**, 197(1), 197-207.

[25] Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *Trends Analyt. Chem*. **2019**, 120, pii: 115644.

[26] Gomes, F. P.; Yates, J. R., 3rd. Recent trends of capillary electrophoresis-mass spectrometry in proteomics research. *Mass Spectrom. Rev.* **2019**, 38(6), 445-460.

[27] 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.

[28] 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.

[29] Lubeckyj, R. A.; Basharat, A. R.; Shen, X.; Liu, X.; Sun, L. Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples. *J. Am. Soc. Mass Spectrom*. **2019,** 30(8), 1435-1445.

[30] 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(12), 2470-2479

[31] 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(17), 10095-10099.

[32] Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. Coupling Capillary Zone Electrophoresis to a Q Exactive HF Mass Spectrometer for Top-down Proteomics: 580 Proteoform Identifications from Yeast. *J. Proteome Res*. **2016,** 15(10), 3679-3685.

[33] 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(17), 2554-60.

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

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

[36] Britz-McKibbin, P.; Chen, D. D. Selective focusing of catecholamines and weakly acidic compounds by capillary electrophoresis using a dynamic pH junction. *Anal. Chem*. **2000,** 72(6), 1242-52.

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

[38] Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol. Cell. Proteomics* **2010**, 9(12)*,* 2772-82.

[39] Zhang, W.; Wu, H.; Zhang, R.; Fang, X.; Xu, W. Structure and effective charge characterization of proteins by a mobility capillary electrophoresis based method. *Chem. Sci*. **2019**, 10(33), 7779-7787.

[40] Katayama, H.; Ishihama, Y.; Oda, Y.; Asakawa, N. Electrophoretic mobility-assisted identification of proteins by nanoelectrospray capillary electrophoresis/mass spectrometry under methanolic conditions. *Rapid Commun. Mass Spectrom*. **2000**, 14(14), 1167-78.

[41] Knubovets, T.; Osterhout, J. J.; Klibanov, A. M. Structure of lysozyme dissolved in neat organic solvents as assessed by NMR and CD spectroscopies. *Biotechnol. Bioeng*. **1999**, 63(2), 242-8.

[42] Mattos, C.; Ringe, D. Proteins in organic solvents. *Curr. Opin. Struct. Biol*. **2001,** 11(6), 761-4.

[43] Cifuentes, A.; Poppe, H. Behavior of peptides in capillary electrophoresis: effect of peptide charge, mass and structure. *Electrophoresis* **1997**, *18*, 2362-76.

[44] Lee, J.; Perez, L.; Liu, Y.; Wang, H.; Hooley, R. J.; Zhong, W. Separation of Methylated Histone Peptides via Host-Assisted Capillary Electrophoresis. *Anal. Chem.* **2018**, 90,1881-1888.

[45] Lee, J.; Chen, J.; Sarkar, P.; Xue, M.; Hooley, R. J.; Zhong, W. Monitoring the crosstalk between methylation and phosphorylation on histone peptides with host-assisted capillary electrophoresis. *Anal. Bioanal. Chem.* **2020**, 412, 6189-6198.

[46] Champney, W. S. Reversed-phase chromatography of Escherichia coli ribosomal proteins. Correlation of retention time with chain length and hydrophobicity. *J. Chromatogr*. **1990**, 522, 163-70.

[47] Pridatchenko, M. L.; Perlova, T. Y.; Ben, Hamidane, H.; Goloborodko, A. A.; Tarasova, I. A.; Gorshkov, A. V.; Evreinov, V. V.; Tsybin, Y. O.; Gorshkov, M. V. On the utility of predictive chromatography to complement mass spectrometry based intact protein identification. *Anal. Bioanal. Chem*. **2012**, 402(8), 2521-9.

CHAPTER 6. Conclusion and Discussion

This dissertation discussed three improvements in proteomics. First, we improved the low loading capacity of CZE which boosted the performance of CZE-MS in BUP. After systematically optimizing dynamic pH junction, an online stacking method, we improved the loading capacity of CZE from nL scale to sub-μL scale which opened the door to large-scale BUP using CZE-MS [1]. We then coupled peptide fractionation with the optimized CZE-MS method to approach large-scale profiling of the mouse brain proteome digest and the human cancer cell line phosphoproteome digest [2,3]. We reconfirmed the complementarity in peptide identification and phosphopeptide characterization between the 2D-LC-MS/MS and SCX-RPLC-CZE-MS/MS. Second, using both 2D-LC and high-pH RPLC-CZE platforms, we built the largest quantitative proteome dataset to date, which depicted the zebrafish proteome dynamics during early embryogenesis. In the database, we found the wave-like expression patterns of TFs which may imply their functions during zebrafish early embryonic development. Third, we explored the possibility of using a semi-empirical model for predicting proteoforms' μ_{ef} as an alternative method for validating proteoform IDs [4,5]. The experimental and predicted μ_{ef} of the calf histone, *E. coli*, and zebrafish proteoforms were linearly correlated suggesting that accurate μef prediction can be achieved. The result also showed the great potential of using the model for further evaluation of proteoform ID confidence.

Future studies should consider using longer capillaries [6] and higher separation voltage [7] to further improve the separation window and peak capacity. Besides, we can couple the ion mobility techniques such as high field asymmetric waveform ion

mobility spectrometry (FAIMS) to CZE to further boost the sensitivity of the platform [8,9].

To better understand early embryogenesis, we expect the combination of loss-of-function studies and quantitative proteomics would provide significant insight into TF functions and their relationship to important embryonic events such as organogenesis. PTMs on histones play important regulatory roles in gene expression [10]. To study the specific function of histone PTMs in early embryogenesis, top-down proteomics analysis of zebrafish histones can be performed using CZE-MS/MS.

Although we provided a prediction model that has a linear correlation between the predicted and experimental μ_{ef} of proteoforms, we noted that only unmodified, single phosphorylated, and N-terminal acetylated proteoforms were studied using the semi-empirical model so far. To evaluate the influence of specific PTM on proteoforms' migration in CZE, one should consider enriching the low-abundant proteoforms with the specific PTMs (i.e., lactylated proteoforms) prior to CZE-MS analysis. Protein sequence specific factors should be explored and integrated into the model as well [11].

REFERENCES

REFERENCES

[1] Chen, D.; Shen, X.; Sun, L. Capillary zone electrophoresis-mass spectrometry with microliter-scale loading capacity, 140 min separation window and high peak capacity for bottom-up proteomics. *Analyst.* **2017**, *142*, 2118-2127.

[2] 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.

[3] Chen, D.; Ludwig, K. R.; Krokhin, O. V.; Spicer, V.; Yang, Z.; Shen, X.; Hummon, A. B.; Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Large-Scale Phosphoproteomics with the Production of over 11,000 Phosphopeptides from the Colon Carcinoma HCT116 Cell Line. *Anal. Chem*. **2019,** 91(3), 2201-2208.

[4] 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.

[5] 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(10), 4417-4424.

[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(8), 1435-1445.

[7] Henley, W. H.; Jorgenson, J. W. Ultra-high voltage capillary electrophoresis >300 kV: recent advances in instrumentation and analyte detection. *J. Chromatogr. A*. **2012**, 1261, 171-8.

[8] 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), M111.014985.

[9] 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.

[10] Su, Z.; Denu, J. M. Reading the Combinatorial Histone Language. *ACS Chem. Biol.* **2016**, *11*, 564– 74.

[11] Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem.* **2017**, 89(3), 2000-2008.