DEVELOPING ULTRASENSITIVE MS-BASED PROTEOMIC PLATFORMS FOR THE CHARACTERIZATION OF MASS-LIMITED SAMPLES

By

Zhichang Yang

A DISSERTATION

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

Chemistry - Doctor of Philosophy

2021

ABSTRACT

DEVELOPING ULTRASENSITIVE MS-BASED PROTEOMIC PLATFORMS FOR THE CHARACTERIZATION OF MASS-LIMITED SAMPLES

By

Zhichang Yang

Single cell analysis such as single cell sequencing has shed invaluable light on cellular heterogeneity and understanding molecular mechanisms such as cell differentiation. Modern single cell sequencing techniques have achieved throughput of analyzing thousands of single cells per day with sequencing depth of thousands of genes. Proteins play vital roles in almost all biological processes and have been crucial biomarkers for disease diagnosis and drug development. Unlike DNA and RNA molecules, protein molecules can't be amplified, making the large-scale characterization of proteins (proteomics) challenging for mass-limited samples such as single cells. Novel proteomics methodologies with extremely high sensitivity are vital for analysis of mass-limited samples. This work focuses on developing ultrasensitive Mass Spectrometry (MS)-based proteomics platforms to enable large-scale proteome profiling of mass-limited samples.

In Chapter 2, we applied nanoRPLC-CZE-MS/MS platform for large scale proteome profiling on 5 µg of a MCF7 cell digest. The digest was fractionated by the nanoRPLC, followed by dynamic pH junction based CZE-MS/MS. The nanoRPLC-CZE-MS/MS produced over 7500 protein IDs and nearly 60000 peptide IDs from the 5-µg MCF7 proteome digest. It reduced the required amount of complex proteome digests for LC-CZE-MS/MS-based deep bottom-up proteomics by two orders of magnitude.

In Chapter 3, we improved the sensitivity of the nanoRPLC-CZE-MS/MS system drastically. The improved system identified 6500 proteins from a MCF7 proteome digest starting with only 500-ng peptides using a Q-Exactive HF mass spectrometer. In addition, we coupled single spot solid phase sample preparation (SP3) method for sample preparation on 5000 HEK293T cells, resulting in 3689 protein IDs with the consumption of a peptide amount that corresponded to only roughly 1000 cells.

In Chapter 4, we developed a Nanoparticle-aided Nanoreactor for Nanoproteomics (Nano3) technique for processing few mammalian cells for bottom-up proteomics. The Nano3 technique employed nanoparticles packed in a capillary channel to form a nanoreactor (≤30 nL) for concentrating, cleaning, and digesting proteins followed by nanoRPLC-MS/MS analysis. The Nano3 method identified 40-times higher number of proteins from 2-ng mouse brain protein samples compared to the low volume SP3 method. The Nano3 method was further applied in processing 10-1000 HeLa cells for bottom-up proteomics, producing 1084 ± 287 (N=4) protein IDs from only 10 HeLa cells using a Q-Exactive HF mass spectrometer.

In Chapter 5, we developed a universal sample preparation method for denaturing top-down proteomics (dTDP), and the method combined the sodium dodecyl sulfate (SDS)-based protein extraction and the membrane ultrafiltration (MU)-based protein cleanup. The MU method outperformed CMP (chloroform-methanol precipitation) and SP3 methods, resulting in high and reproducible protein recovery from both *E. coli* cell (59±3%) and human HepG2 cell (86±5%) samples without a significant bias. The assay afforded identification of various post-translational modifications and protein containing transmembrane domains through top-down analysis.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Professor Liangliang Sun. It is still very vivid in my mind back in April 2017, when I was rejected by all the graduate schools I applied for PhD program, how stressful I was with all the expectation dashed while the heavy workload and pressure from work had not alleviated for a single minute. I still remember, with the thought of fighting for a losing battle, I sent an inquiry email to Dr. Sun asking for possibilities of joining his group, although the deadline of application had long passed. Very surprisingly, he replied my email right away with positive information. It was like sunray piercing through dark clouds to me at that moment when I was reading his message. I feel really really grateful for his acceptance. For the past four years of working with Dr. Sun, He really showed me what a great advisor he is. He generously supported my idea of research, kindly gave advice and favors to me when I was facing difficulties in research and in life and more importantly, he was always there whenever I had any questions, concerns and even disagreements. He always showed me his side of patience and encouraging. Not only Dr. Sun cares a lot about research and science, but he also cares a lot about us. I still remembered he often suggested us to do some outdoor activities in a good day, he bought a lot of face masks for us during the pandemic, he constantly asked about our health and if we needed any help from him. All these acts seem small, but they mean big to me.

I also would like to thank Dr. Spence, Dr. Blanchard, and Dr. Cibelli as my committee members. I really appreciate the advice they gave me along my pursuing of PhD degree. They taught me a lot on being serious with science and research. I really

enjoyed the discussion we had during my first committee meeting and oral comprehensive exam.

I also want to thank Dr. Xuefei Huang, Dr. Jian Hu, Dr. Heedeok Hong, Dr. Yuan Wang, Dr. David Lubman, Dr. Chen Chen. I asked a lot of favor from them and they are all super generous in helping me with my research.

I want to thank Billy Poulo and Zhaoran Zhang in helping me collecting samples for study. I have asked so many times of sample from them and they never expressed even a tiny reluctance with me. Without their help, my projects can never progress.

I want to thank my group members, Xiaojing, Daoyang, Rachele, Eli, Tian, Qianjie, Qianyi and Jorge. I enjoyed the time we have been working together, discussing science, chatting, group lunch in the office, our journey at ASMS. It will be treasure memory for me.

I also want to thank my friends, Fangchun, Yijin, Mengxia, Jiaqi. They are my families here in Michigan. I can never imagine what my life would be without them for the past four years. Together we have had so much fun. I will miss the most of our traditional Friday get-together dinner and movie. Their friendship supported me through all the loneliness during pandemic. I also want to thank my friend Henry. He is such a kind person. I really appreciate all the help and encouragement he gave me.

Lastly, I want to thank my parents and my brother. Their love is greatest tressure in my life. I feel really sorry that I was not around for the past four years. I wish I can have more time with them in the future.

TABLE OF CONTENTS

LIST OF FIGURES

KEY TO ABBREVIATIONS

2D	Two Dimensional
µRPLC	Microscale Reversed-phase LC
ACN	Acetonitrile
AFA	Adaptive Focused Acoustics
ATM	Accurate Time and Mass
AUC	Area Under Curve
BCA	Bicinchoninic Acid
BGE	Background Electrolyte
BSA	Bovine Serum Albumin
CID	Collision-induced-dissociation
CMP	Chloroform-methanol Precipitation
CV	Compensation Voltage
CZE	Capillary Zone Electrophoresis
DDA	Data-dependent Acquisition
DDM	N-Dodecyl B-D-Maltoside
DMEM	Dulbecco's Modified Eagle Medium
dTPD	Denature Top-down
DTT	Dithiothreitol
ECD	Electron Capture Dissociation
eFT	Enhanced Fourier Transform

EOF	Electroosmotic Flow
ESI	Electrospray Ionization
ETD	Electron Transfer Dissociation
FA	Formic Acid
FACS	Fluorescence Activated Cell Sorting
FAIMS	High Field Asymmetric Waveform Ion Mobility Spectrometry
FASP	Filter Aided Sample Reparation
FDR	False Discovery Rate
FTICR	Fourier Transform Ion Cyclotron Resonance
FTMS	Fourier Transform Mass Spectrometer
FWHM	Full Width at Half Maximum
GELFrEE	Gel-eluted Liquid Fraction Entrapment Electrophoresis
GO	Gene Ontology
GRAVY	Grand average of hydropathy
HCD	Higher Energy Collision-Induced Dissociation
HETP	Height Equivalent Theoretical Plate
HF	Hydrofluoric Acid
i.d.	Inner Diameter
IAA	Iodoacetamide
ID	Identification
IEF	Isoelectric Focusing
inDrops	Index Droplet
iPAD	Integrated Proteome Analysis Device

KE	Kinetic Energy
LC	Liquid Chromatography
LCM	Laser Capture Microdissected
LFQ	Label Free Quantification
LPA	Linear Polyacrylamide
LTQ	Linear Trapping Quadrupole
MBT	Mid-blastula Transition
MCP	Multichannel Plate
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MU	Membrane Ultrafiltration
MW	Molecular Weight
MWCO	Molecular Weight Cutoff
Nano3	Nanoparticle-assisted Nanoreactor for Nanoproteomics
NanoPOTs	Nanodroplet Processing in One Pot for Trace Samples
nanoRPLC	Nanoflow Rate Reverse Phase Liquid Chromatography
NOT	Narrow Open Tubular
OAD	Oil-air-droplet
PASEF	Parallel Accumulation - Serial Fragmentation
PBS	Phosphate Buffered Saline
PI	Isoelectric point
PLOT	Porous Layer Open Tube
PrSM	Proteoform Spectrum Match

PSM	Peptide-spectrum Matching
PTM	Post Translational Modification
RF	Radio Frequency
RSDs	Relative Standard Deviations
SCoPE-MS	Single Cell Proteomics By Mass Spectrometry
SDS	Sodium Dodecyl Sulfate
SEC	Size Exclusion Chromatography
SNP	Single Nucleotide Polymorphism
SP3	Single Spot Solid Phase Sample Preparation
SPME	Solid Phase Microextraction
TIMS	Trapped Ion Mobility Mass Spectrometry
tITP	Isotachophoresis
TMDs	Transmembrane Domains
TMT	Tandem Mass Tag
TOF	Time-of-flight
UMI	Unique Molecular Identifier

[1]CHAPTER 1. Introduction

1.1 Single cell analysis

Cell is the fundamental structural and functional units of living organisms. It is essential to study the composition of the biological molecules in cells, to have a comprehensive understanding of cell functions and intercellular interactions. The studies of collective characterization of the biological molecules are called -omics, including genomics, transcriptomics, proteomics and metabolomics etc., regarding the specific molecule type for characterization. Most -omics studies are implemented on either a mixture of heterogeneous cells or on sorted subpopulations, usually containing millions of cells, and are called bulk analysis. Bulk analysis is usually performed with easily accessible and well-established techniques with high accuracy and precision. However, bulk analysis can only provide average measuring of the compositional change of biological molecules between samples, i.e., with bulk analysis, the heterogeneity of individual cells is lost. Even in one cell type, various subtypes of cell populations exist, and they have distinct gene expression profiles.[1,2] Even under seemingly stable and identical environment, cells can display heterogeneous behaviors.[3] As shown in **Figure 1.1** , cellular response is monitored as a function of inducer concentration. When signal molecule (i.e., inducer) is below the average threshold level, no cellular response is observed through bulk analysis. However, by checking individual cells, a small fraction of cells are already turned on. The number of

[1] Part of this chapter was adapted with permission from: Yang, Z.; Sun, L., Analytical Method 2021, 13, 1214-1225.

cells respond to the inducer increases when the inducer concentration keeps increasing. When the concentration of the inducer is well above the threshold, the cellular response saturates, indicating fully turned-on of cells to stimulus. However, there is still a small portion of cells not responding, possibly due to mutation or cell death.[4] The cellular heterogeneity has essential biological significance. During the early embryogenesis, blastomeres in one embryo gradually differentiate from each other at the molecular level (e.g., protein), establishing the foundation for organogenesis.[5,6] Strong protein-level heterogeneity across cancer cells in one tumor makes drug development for cancer treatment challenging.[7-9]

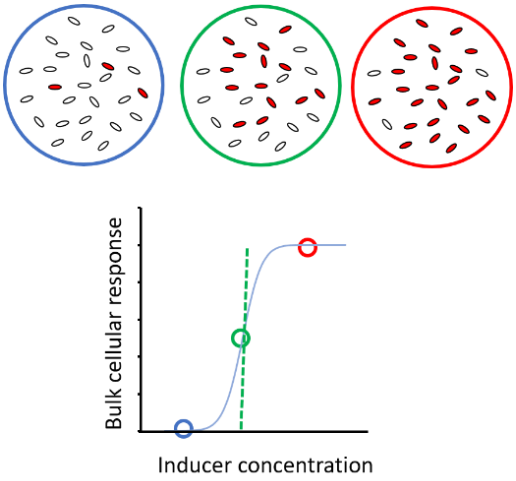

Figure 1. 1 Bulk measurement might not manifest the heterogenous response of cells to stimulus. Reproduced from ref. 4 with permission from Springer Nature, copyright (2003).

Single cell analysis has been applied in genomics and transcriptomics. In 1988, Li et. al. reported the very first analysis of DNA sequence in single diploid cell.[10] Two years later. Brady et al. reported a single cell transcripts analysis of hemopoietic cell.[11] In 2006, Kurimoto reported an innovative cDNA amplification strategy for single cell

microarray analysis, which improved gene expression profile accuracy and reproducibility.[12] A modified protocol of the cDNA amplification was reported by Tang et al. in 2009 and achieved full-length capture of cDNA with no bias. 1,753 previously unknown splice junctions were identified, and more gene expressions were detected at single-cell resolution than microarray assay with hundreds of cells.[13] In 2012, a mRNA sequencing strategy on single cell called Smart-Seq was developed.[14] This sequencing strategy resolved the issue of lacking technique-control to distinguish truly biological variation from technique variation, and the inability of generating whole read coverage of the transcripts. Circulating tumor cells from melanoma were analyzed by Smart-Seq. Distinct gene expression patterns indicating possible disease biomarker were identified.[14] Intrinsic characteristics of nucleotide-based molecules (i.e., DNA, RNA) enables the use of barcode to differentiate molecules from different samples so that pooling and parallel analysis of thousands of single cells are possible. As a result, the reproducibility and throughput of single cell analysis can be significantly improved.[15,16] Combining microfluidic strategy and barcoding, a technique called inDrops (index droplet) was developed by Klein et al. and achieved tens of thousands single cell transcriptomics analysis with great sequencing depth.[17] In inDrops, Single cells were uniquely barcoded with an oligonucleotide primer (one out of 147,456 synthesized primers) within an oil-formed-droplet through well-controlled microfluidic device. mRNA from the single cell was barcoded during synthesis of complementary DNA. After barcoding, droplets were broken and material from thousands of cells was pooled for sequencing (**Figure 1.2**). Informative insights of embryo development were revealed through the inDrops technique when it was applied to study single blastomeres from

zebrafish and Xenopus laevis embryos.[6,18]

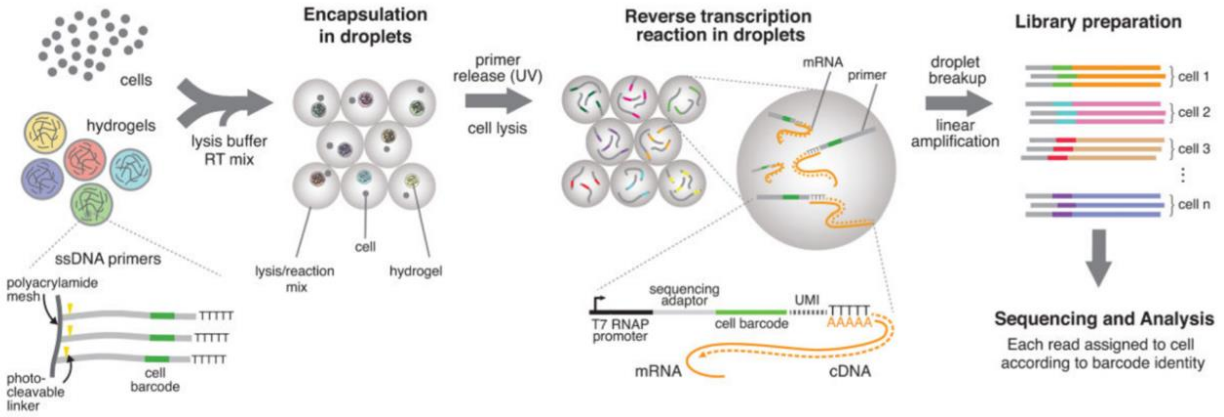

Figure 1. 2 A platform for DNA barcoding thousands of cells. Cells are encapsulated into droplets with lysis buffer, reverse-transcription mix, and hydrogel microspheres carrying barcoded primers. After encapsulation primers are released. cDNA in each droplet is tagged with a barcode during reverse transcription. Droplets are then broken and material from all cells is linearly amplified before sequencing. UMI = unique molecular identifier. Reproduced from ref. 17 with permission from Elsevier, copyright (2015).

1.2 Importance of protein analysis and proteomics

Proteins play central biological functions in cells and are involved in every biological process within cells including catalyzing metabolic reactions, regulating DNA replications, expression and repairing, providing structure to cells and tissues and transducing signals for cell communications etc. The central dogma describes a flow of information passing from DNA to RNA through transcription and from RNA to protein through translation. There are about 20,000 encoding genes in human, but millions of proteins with different forms due to alternative splicing, genetic variants and protein post-translational modifications (**Figure 1.3.**). These events create highly related but

chemically different protein variants performing distinct biological functions. Taken histone, a highly modified protein that packs DNA molecules into nucleosome, as an example. Histone plays central roles in gene regulation. Genes will be turned on or off based on the modification states of amino acids in histones. A variety of modifications such as methylation, acetylation and phosphorylation occur in the tail domain of the histone dynamically.[19] Enzymes transducing the modification to the amino acids are very specific to the particulate amino acid positions.[20,21] The combination of the modifications not only alters the structure of chromatin and the interaction between DNA and histones, but also forms a “code” that is read by other proteins. It has been found that the histone code regulates gene expression, DNA repairing and chromosomal condensations with complex mechanisms.[20] The fact that highly complex proteins with different forms performing distinct functions, represents an example of how genome information is not enough to explain phenotype differences.

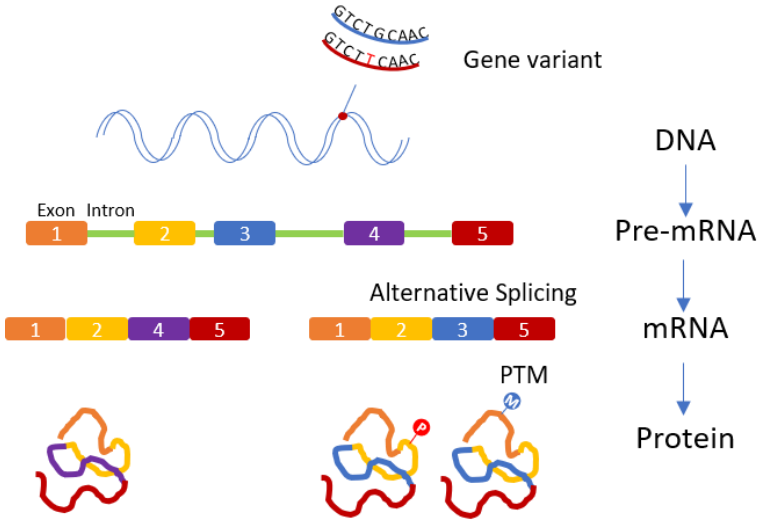

Figure 1. 3 Central dogma of molecular biology. The gene variants such as single nucleotide polymorphism (SNP) at DNA level, the alternative splicing at mRNA level, and the post translational modification (PTM) at protein level may all contribute to the difference of phenotype from same genotype.

Another reason of the significance of protein characterization, is that genome-wide correlation of mRNA and protein is poor, even though gene expression can be quantitatively analyzed at both mRNA level and protein level. Studies found that only 40% of variance in protein expression can be explained by changes at transcripts level.[22] Source of poor correlation may come from different mRNA and protein stability, different efficiency of translation, and alternative splicing products of transcripts. Proteins were also found being more stable, more abundant and spanning higher dynamic range than mRNA.[22-24]

Similar to the concept of genome and transcriptome, proteome is a representation of entire component of proteins in the biology system. Study of the proteome is called proteomics, includes but not limits to identification and quantification

of the entire set of proteins and PTMs, exploring protein interactions and protein high order structures.[25]

Mass spectrometry-based proteomics has become one of the major tools to measure protein molecules in a biological system at a global scale. Studies on protein identity, quantity, PTMs, localization, temporal and spatial dynamics, and even structural assembly have been implemented using MS-based proteomics. One of ultimate goals of proteomic research is comprehensive profiling of whole proteomes of cells. Normally, a comprehensive profiling study requires at least tens of micrograms of starting protein material. The requirement of large amount of protein material is mainly due to two reasons: first, protein abundance in cells can have very high dynamic range, from one copy per cell to millions of copies per cell, and second, sensitivity of modern MS-based platform is not high enough.[26-28] The protein mass in a single human somatic cell is on the order of sub-nanogram.[29,30] Unlike mRNAs and DNAs, protein molecules cannot be amplified. The application of oligonucleotide barcode is also not feasible for protein analysis, preventing pooling thousands of single cells to improve throughput and protein abundance. Hence, technique advancement for single cell proteomics is highly desired. With the advance of MS instrumentation, single ion detection is possible.[31,32] The state-of-the-art proteomics platforms have also achieved the sensitivity of 1-100 zmol for proteins.[29,33-38] Mammalian cells can express 12,000 to 15,000 different proteins spanning 7 orders of magnitude in concentration. With the modern proteomic techniques, over 4,000 proteins could be identified in theory from a single mammalian cell.[23] However, only few studies have reported the identification of hundreds of proteins from single human cells.[38,39]

Technical breakthrough in developing extremely sensitive proteomic methodologies will enable global characterization of proteins in a small number of cells and even single cells, leading to substantial impact on the understanding of various biological questions in cancer biology, developmental biology, neuroscience, etc.

1.3 Mass spectrometry-based proteomics

1.3.1 Electrospray ionization

Electrospray ionization mass spectrometry (ESI-MS) is one of the most important techniques for proteomics study. ESI can produce multiple charged ions with minimum or no fragmentation, allowing characterization of large biological molecules. Briefly, ESI utilizes electrical energy to transfer ions from liquid phase to gas phase for MS analyzing. For positive mode ESI-MS, chemicals such as acetic acid or formic acid are added into the solution to facilitate ionization and provide protons for analytes. With the high voltage applied, charged droplets are formed at the electrospray tip when liquid phase analytes exit the emitter. Solvent in the droplet evaporates very fast in condition of high voltage and high temperature, leading droplet size reducing and surface charge increasing. When the electric field strength within the droplet reaches to a critical point, namely Rayleigh limit, droplets fissure into smaller droplets until gaseous ions are generated. Charged gaseous ions are then introduced into MS inlet by electric field for analysis (**Figure 1.4**).[40-42] The no fragmentation nature of ESI allows intact mass to be characterized for large biological macromolecule such as proteins and peptides. The fact that multiple charges are deposited onto the large molecules increases the mass range MS can analyze.

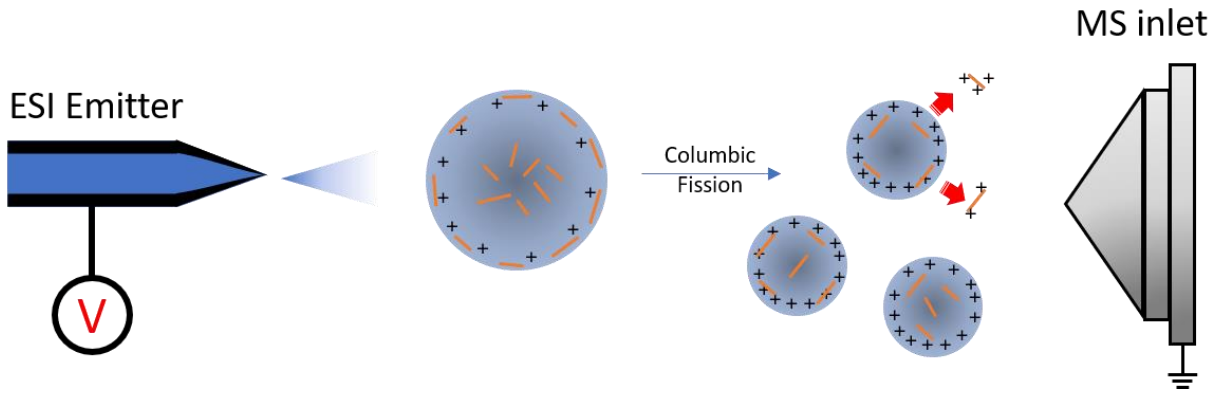

Figure 1. 4 Electrospray ionization process.

1.3.2 Top-Down and bottom-up proteomics

Depends on whether proteins are digested by proteolysis protease or not, proteomics study can be categorized into bottom-up proteomics and top-down proteomics. In top-down proteomics, proteins are directly analyzed by MS without enzyme cleavage. Without digestion, it is possible to achieve full sequence coverage characterization of proteins.

In bottom-up proteomics, proteins are digested into peptides by proteolysis protease such as trypsin because trypsin cuts the carboxyl side of amino acid lysine and arginine with high specificity.[43,44] In this way, all peptides contain at least two basic sites for protonation (N-terminus and C-terminus with lysine or arginine), which is beneficial for peptide ionization and fragmentation. The proper frequency occurrence of lysine and arginine in protein renders length of tryptic peptides in the range of 10-20 amino acids, which places peptide in ideal m/z range for MS analysis when multiply charged.

In top-down proteomics, the intact mass measurement and fragmentation-assisted sequencing afford preservation of PTM information, which could be very

challenging to localize among proteins with high sequence similarities in bottom-up proteomics (**Figure 1.5A**). However, the large size of proteins leads multiple charge states of a protein and consequently dilutes the MS signal into multiple charges states channels. Bottom-up proteomics on the other hand, does not suffer from the signal dilution since most tryptic peptides carry only 2 or 3 charges (**Figure 1.5B**). Thus, bottom-up proteomics usually has higher sensitivity[45,46] and is mainly used in trace proteomics and single cell proteomics study.

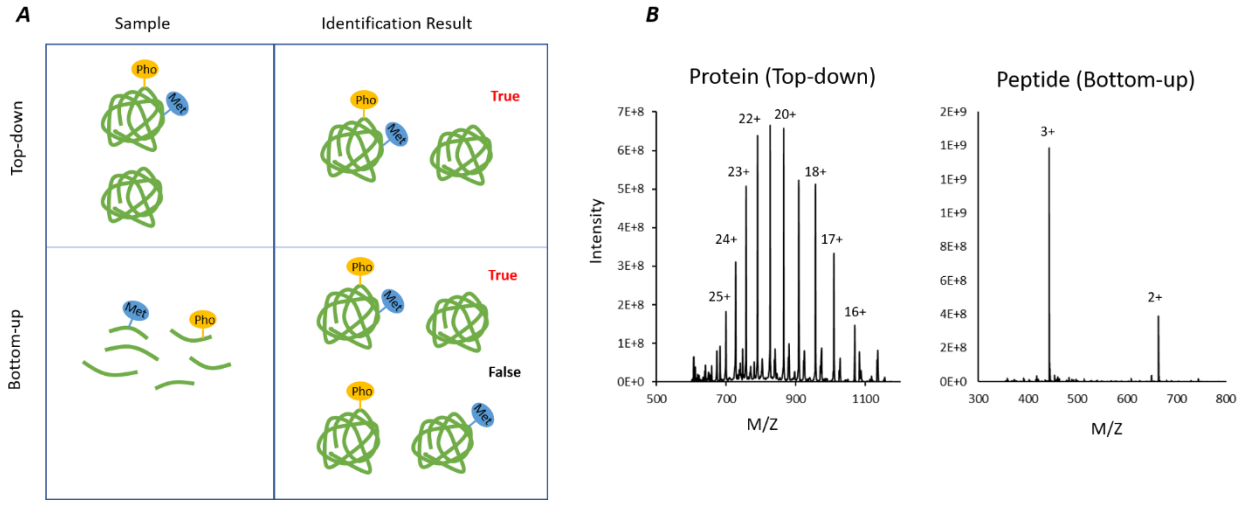

Figure 1. 5 Top-down and Bottom-up proteomics. (A) Determination of proteoforms and PTMs from top-down and bottom-up proteomics. It is challenging to allocate PTMs when sequence-similar proteins are digested into peptides in bottom-up proteomics. Pho: phosphorylation. Met: methylation. (B) Charge states distribution of protein and peptides.

1.3.3 Tandem mass spectrometry

Mass spectrometer measures molecular mass of gas-phase ions through the motions of ions in magnetic/electric field.[47] The interaction of ions with magnetic/electric field also depends on the elementary charge the ions carry (charge state). Mass

spectrometers acquire the ratio of molecular mass to charge states of ions (m/z) and determine the charge states through isotopic pattern of ions. With known m/z and charge states, the molecular mass can be calculated straightforwardly. The presence of isomeric and isobaric species in complex sample makes it challenging for peptide identification with intact molecular mass only (usually acquired through full MS can (MS1)). Fragmentation on peptides backbone is usually required and a tandem MS scan (MS^2) of fragment ions provides additional level of specificity. Tandem MS occurs when consecutive MS scans of fragment ions that are produced by dissociation of interrogated intact molecular ion. The use of intact ion mass and corresponding fragment ions mass series enables accurate identifications of peptides. For peptide Identification, the empirical tandem MS spectra is compared with database containing theoretical tandem MS spectra. In general, protein sequences derived from open reading frames of Genome are subjected to *in silico* digestion with designated cleavage specificity (depends on enzyme used for digestion). The consequential peptides are then fragmented with specific fragmentation patten based upon the applied fragmentation strategy to generate the database for comparison. In bottom-up proteomics, identified peptides were used for their origin proteins identification and quantification. This type of experiment is called “shotgun” proteomics.[48-50]

1.3.4 Fragmentation of peptides

In The most commonly used methods to generate fragment ions from peptides for tandem MS is collision-induced-dissociation (CID). In general, peptide ions are accelerated and collide with natural gas molecules such as helium or argon in the cell. The kinetic energy is imparted to the ions and when the energy is above the threshold

of energy needed to break a bond, fragmentation occurs.[51] The most prevalent fragment ions introduced by CID are b-type and y-type when fragmentation occurs at peptide bond (**Figure 1.6**). With known fragment pattern of peptides, assignment of tandem MS spectra to theoretical spectra is much easier.

Peptides with labile PTMs such as phosphorylation and glycosylation can go through PTM loss during CID fragmentation due to the low fragmentation energy threshold of the bond between PTM and peptides, making PTM mapping a challenge task if CID is applied to generate tandem MS.[31,52]

Electron capture/transfer dissociation(ETD/ECD) on the other hand, can be alternative fragmentation strategies for peptide sequencing.[53,54] In ETD, multiply charged ions receive electrons from radical anion produced from reagent such as fluoranthene and become radical cations, $[M+nH](n-1)+\bullet$. The radical cations introduce rapid and facile fragmentation at N-C bond and generate z and c type fragment ions(**Figure 1.6**).[54] ECD is very similar to ETD, with only electrons produced from emitter cathode are captured by multiply charged ions directly. The fact that the process of imparting internal energy to peptide ions is faster than the rate of energy randomization, leads fragmentation not necessarily on the weakest bond as that of CID.[31] As a result, ETD/ECD usually introduce fragmentation on the peptide backbone and preserve well the PTM information.

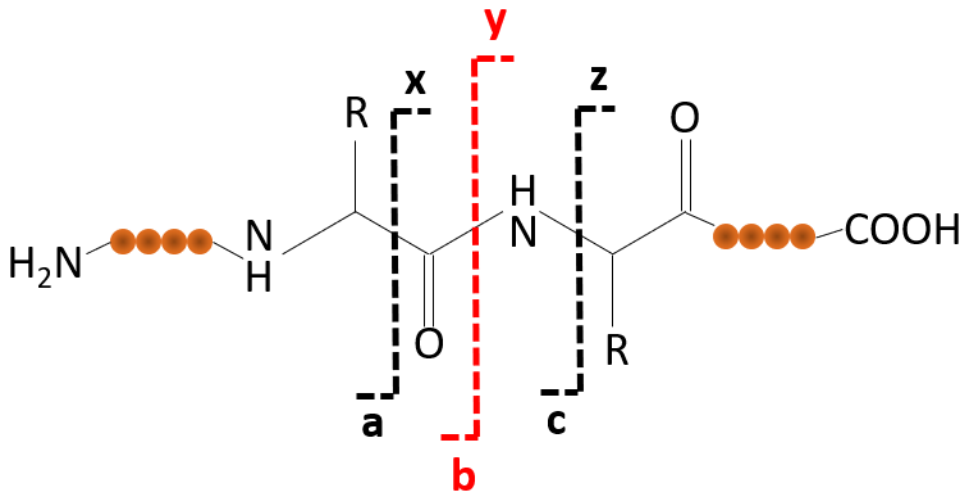

Figure 1. 6 Nomenclature of fragment ions of peptide.

1.3.5 Database searching

Spectra matching for peptide identification is usually performed through computational database searching.[55-57]

Briefly, the protein sequences of a proteome in the database are subjected to in-silico digestion following specific enzyme cleavage rules, e.g., tryptic peptides are produced from protein cleavage at the C-terminal of lysine and arginine. The resulting peptide list is serving as a master peptide list to which the empirical interrogated peptide mass compares with specified mass tolerance. With high resolution mass analyzer, the mass tolerance is usually set at ppm level. Subsequently, the MS/MS spectra are then compared against the theoretically possible fragment ions that are generated with specific fragmentation patterns from possible peptides. each comparison will be assigned a match score to estimate how close the empirical to the theoretical is.[58] Multiple factors could affect the score of peptide matching, such as number of matching fragment ions observed in empirical MS2 spectra, mass accuracy and intensity of

fragment ions. In addition to searching in target (i.e., normal) proteome database, A decoy database (usually a database with reverse sequence or randomized sequence of the target database) searching is usually performed[59] to estimate statistically the false discovery rate (FDR) of the matching. It has been proved that matching in the decoy database can be a good estimation of the false matching in the target database.[60] The FDR is usually set at 1% level for a shotgun proteomics database search. The identified peptides were then mapped back to the identification of proteins. A general workflow of bottom-up proteomics is shown in **Figure 1.7.**

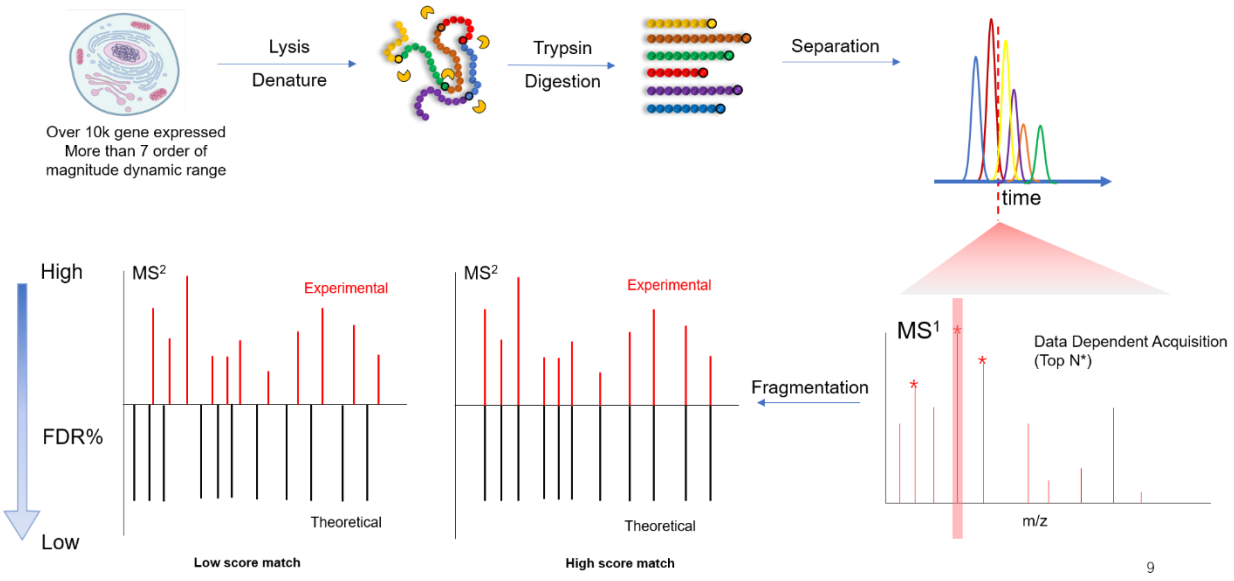

Figure 1. 7 General workflow of bottom-up proteomics.

1.3.6 Protein quantification in shotgun proteomics

Quantitative analysis of proteins provides insight of biological states of cells and tissues and has contributed significantly to understanding biological questions.[61] In shotgun proteomics, signal intensities from peptide precursor ions or fragment ions are usually used to estimate relative abundance change across samples.

To address technique variation from multiple-step sample handling and instrument measurement, isotopic labeling strategies are developed to achieve relative quantification[62-65] or absolute quantification of proteins from samples.[66] Tandem mass tag (TMT) is a representative of quantification techniques using isotopic labeling for relative quantification of proteins.[64] The tag contains an amine-reactive group that can react with amine group in peptides (**Figure 1.8**). Once fragmented, the reporter group is cleaved, and its relative intensity is used to indicate the relative abundance of labeled peptides across samples. The mass normalizer is used to balance the molecular mass of labeled peptides, so that same peptides from different sample source have exact same molecular mass after being labeled. Because same peptides from different sample source have same hydrophobicity and molecular mass, ensuring co-elution and co-isolation for fragmentation, the application of TMT labeling affords accumulation of MS1 signal from multiple channels and improves sensitivity of MS analysis. Up to 16 channels have been developed using TMT labeling.[67] The multiplexity significantly improves throughput of proteomics analysis and is crucial for quantifying proteins in cohort containing large number of samples.

Another popular protein quantification strategy is label free quantification (LFQ). In LFQ, isotopic labeling treatment is obviated so sample processing is simplified, and cost of labeling reagent is eliminated. It has also been reported that labeling reagent can have significant impact on peptide-spectrum matching (PSM) and consequentially influence number of identification of peptides and proteins.[68] Moreover, there is no limitation of sample numbers when quantification is achieved through LFQ. Instead of using intensities of fragment ions to achieve peptide abundance, peak area under curve

(AUC) of peptides from separation profile is used to estimate peptide intensities. In 2002, Chelius et al. demonstrated that, with proper normalization, AUC of peptide peaks can accurately reveal relative intensity of proteins across samples.[69] Sophisticated algorithms were also developed to improve quantification precision and accuracy.[70,71]

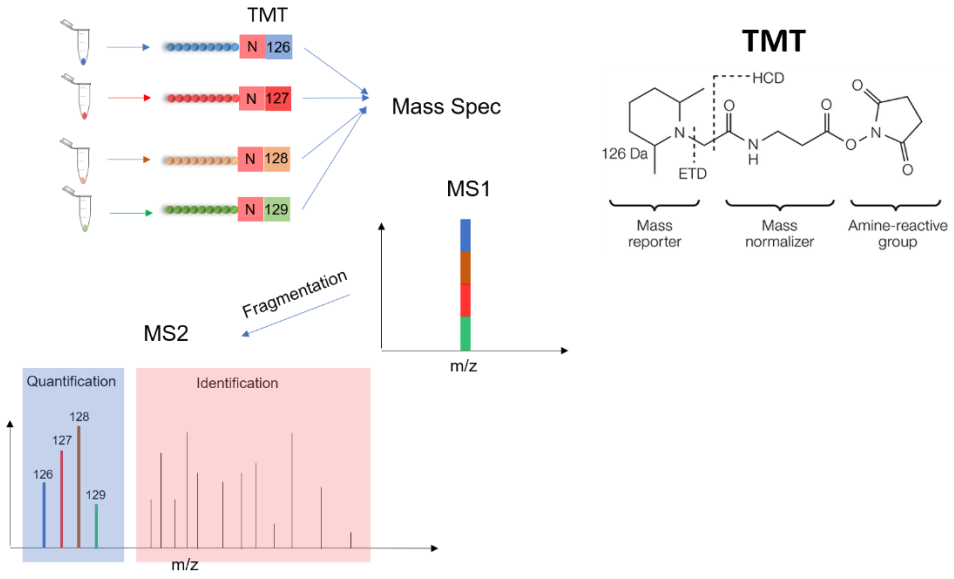

Figure 1. 8 Protein quantification with TMT labeling strategy.

1.4 MS-based proteomics analysis of mass-limited sample

1.4.1 Sample preparation in proteomics of mass-limited sample

A general sample preparation workflow for bottom-up proteomics study includes protein extraction and denaturing, reduction and alkylation,[72,73] digestion, and peptide clean-up.

Protein extraction includes disruption of the cell membrane and denaturing and solubilization of proteins. Mechanical disruption such as sonication and homogenization are commonly applied to facilitate protein extraction. During extraction, additives to lysis buffer such as chaotropic reagent (e.g., urea) or detergent (e.g., SDS) are usually applied to facilitate protein denaturing and solubilization. However, these additives

might affect the enzyme activities and following separation and MS detection. Trypsin tolerates up to 2 M urea concentration for enzyme activity while 8M urea concentration is normally applied for protein extraction.[74] Thus a 4-times dilution of urea is required before digestion, leading to increasing of processing volume. A follow-up desalting step is required to eliminate urea's effect on MS signal. SDS has been widely used in proteomics studies to facilitate protein extraction and protein solubilization. However, trace amount of SDS is detrimental to downstream processes such as enzyme digestion, LC separation and ions detection by MS.[75,76] Numerous techniques have been developed for SDS depletion in bottom-up proteomics studies.[77-80]

For mass-limited sample processing, protein extraction, digestion and clean-up need necessary adjustment to maximize sample recovery and minimize sample loss. One of the major challenges of comprehensive characterization of the mass-limited samples is the sample loss during the sample preparation and liquid-phase separation caused by hydrophobic adsorption of proteins on surfaces, such as processing containers, pipette tips, the stationary phase of LC columns, and sample loading valves in LC, leading to low sample recovery from sample preparation and limited sensitivity of LC-MS for proteomics.

Reducing sample loss due to nonspecific interactions between proteins/peptides and surfaces is the key point in processing trace proteome samples and can be achieved by minimizing sample processing volume, eliminating sample transferring, and blocking the non-specific interactions.

1.4.1.1 Minimizing sample processing volume

Controlling the sample processing in a very small volume (i.e., nanoliters) is one efficient approach for preparing mass-limited samples. In 2018, Zhu et. al. developed a strategy called NanoPOTs (nanodroplet processing in one pot for trace samples) platform for proteomics analyses of small numbers of human cells.[38] Nanowells were manufactured on a glass chip with hydrophilic surface and with diameters of 1 mm (**Figure 1.9A**). All steps of sample processing including protein extraction, reduction and alkylation and digestion took place in the nanowells with total volume of 200 nL. The NanoPOTs platform significantly reduced the reaction volume by 99.5% compared to regular sample processing in a 0.5-mL Eppendorf tube. In addition to the nanowell, the hydrophilic surface can reduce hydrophobic adsorption. Rapigest was used as the surfactant to facilitate cell lysis while did not affect trypsin digestion and MS signal.[81] The NanoPOTs has been applied to few cell and even single cell proteomics,[82] laser capture microdisected (LCM) tissues with a diameter low to 50 μm,[83] and circulating tumor cells.[84] Hundreds or over one thousand of proteins were identified from single HeLa cells with or without the match-between-runs algorithm.[71] The algorithm is integrated in the MaxQuant software.[85] The fact that the match-between-runs algorithm boosted the number of protein identifications drastically indicated the value of high mass accuracy and reproducible retention time of parent peptides on protein identification from trace proteome samples.

A nanoliter-scale oil-air-droplet (OAD) chip was introduced in 2018 by Li et. al. for sample preparation of low population of cells.[86] In a sandwiched device, a droplet containing cells was deposited on to a low retention millimeter sized chip and was

isolated from the outside world by an oil layer to prevent evaporation of liquid (**Figure 1.9B**). Surface tension prevented the direct contact of the oil and sample droplet. Followed by the deposition of the cell droplet, a series of reagents were added into the droplet to implement cell lysis, protein reduction and alkylation, and protein digestion. The total droplet volume was estimated around 550 nL. A C18 prepacked capillary was then inserted into the droplet. Peptides were directly loaded onto the capillary column from the droplet through pressurization and online desalting was performed. Using this approach, 51 to 1360 protein groups were identified from 1 to 100 HeLa cells. When comparing the OAD and NanoPOTs methods, we noted that the number of protein identifications from single HeLa cells using the OAD chip was much lower than that using the NanoPOTs platform (tens of proteins vs. hundreds of proteins). However, it doesn't necessarily indicate lower sensitivity or more sample loss of the OAD method compared to the NanoPOTs method since different LC-MS/MS systems were used for the two methods (50-i.d. LC column plus Orbitrap Elite vs. 30-i.d. column plus Orbitrap Fusion Lumos or Orbitrap Eclipse Tribrid).

Recently, another on-chip based microfluidic device was developed by the Vinh group.[87] The device integrated an ultrafiltration membrane in a micro-reaction chamber with 1.2 µL volume in total. The membrane divided the chamber in half with one side (0.6 µL) for protein extraction and digestion. The membrane played the role as filtration membrane applied in the filter aided sample reparation (FASP) technique[78] for protein clean up and allow clean peptides to be eluted for the downstream LC-MS/MS. A multi-reagent pump system and multi-way valve were integrated with the chipfilter device to deliver sample and reagent (**Figure 1.9C**). The chipfilter method achieved 10 times

higher sensitivity compared with traditional FASP strategy when the starting protein material was 1 µg. Although the chipfilter was not applied on low numbers of human cells as NanoPOTs and the OAD chip, the chipfilter has great potential on preparing low numbers of cells. First, it is directly connected to LC-MS/MS platform with no need of material transferring. Second, all the sample processing steps take place in one side of the reaction chamber upstream the membrane, which is only 600 nL, leading to comparable volume level as the NanoPOTs and OAD chip. Third, the good sealing of the chamber prevented liquid evaporation. Further sensitivity improvement is highly possible with a smaller chamber volume and special surface treatment of the device to reduce hydrophobic adsorption.

An integrated proteome analysis device (iPAD), **Figure 1.9F**, was developed by the Zhang group in 2015 for ultrasensitive proteome profiling of only 100 living cancer cells.[88] Cells were suspended in a cold solution (4 °C) containing salts (NH_4HCO_3, guanidine hydrochloride, and EDTA) and trypsin at a known cell concentration. An exact proportion of cell solution containing 100 cells was directly drawn into a fused silica capillary loop (100 µm i.d.× 40 cm, 3.2 µL in volume). One hour of heat treatment (50 °C) facilitated the cell lysis and digestion in the capillary loop. The digested peptides were then directly loaded onto a trap column for further LC-MS analysis. 813 proteins were identified from the 100 cells. In 2018, Shao et. al. developed an optimized version of iPAD technique, iPAD-1, for proteomics analysis of single HeLa cells.[89] In iPAD-1, a narrower capillary and simpler valve system were applied, **Figure 1.9G**. Single HeLa cell was aspirated into the capillary under microscope monitoring. The total volume of the processing capillary in iPAD-1 is only 20 nL (22 µm i.d. × 5 cm). In addition to heat

treatment, an ultrasonication probe was placed close to the capillary to facilitate cell lysis and protein digestion. The processing capillary was then connected to an LC column through a union with zero dead volume for LC-MS analysis. Over 180 proteins were confidently identified with MS/MS from one HeLa cell. The ultrasensitive performance of the iPAD-1 device can be attributed to the tiny sample processing volume (20 nL) and the direct connection between sample processing capillary with LC-MS without sample transfer.

All the methods we discussed above handle liquid at the nanoliter level and require special liquid handling systems. For example, NanoPOTs requires 70% humidity to reduce liquid evaporation so the entire sample processing is implemented in a closed humid chamber. The OAD chip method requires an installation of a self-alignment monolithic device for droplet deposition, pressurization and sample loading for LC-MS/MS. The chipfilter technique requires specialization of multi-reagent pump and multi-way valve for reagent delivering and good sealing reaction chamber for sample processing. The requirement of specific instrumentation limits the adaptability of the methods across different labs.

To overcome the instrumentation availability issue of above-mentioned methods, a micro-FASP technique was published by Zhang et al. recently.[90] The strategy adapted the idea of conventional FASP but reduced the surface area of the filtration membrane to 0.1 mm2. The membrane was integrated within a 20 µl pipette tip with bottom and top support for immobilization (**Figure 1.9D**). The sample loading, buffer washing, and elution volume were all controlled at microliter level. Peptides were directly eluted into sample vials for LC-MS/MS analysis to avoid additional liquid transferring. The micro-

FASP identified a comparable number of protein IDs and 20% more peptide IDs compared to the conventional FASP when 10 times lower of protein amount (1 µg vs 10 µg) was consumed. Over 3,000 proteins were identified starting with 1000 MCF-7 cells, indicating the great potential of the micro-FASP method for proteomic analysis on few human cells. More importantly, it does not require special instrumentation for sample preparation.

1.4.1.2 Eliminating sample transferring

Another crucial point on treating trace amount of protein materials is to reduce liquid transferring as much as possible during sample processing. Proteomic sample preparation normally employs detergents (e.g., SDS) or chaotropic reagent (e.g., urea) to facilitate cell lysis and protein extraction. Most of the detergents and chaotropic reagents are incompatible to downstream enzymatic protein digestion and LC-MS analysis. Detergents usually need to be removed through ultrafiltration or precipitation, and chaotropic reagents need to be removed through desalting to ensure the compatibility with follow-up LC-MS experiments. Those steps lead to limited sample loss when hundreds of micrograms of proteins are available but result in serious sample loss for trace protein material processing.

Researchers have been searching for ways to avoid further sample clean up to reduce sample loss. Budnik et. al. used water as lysis buffer, lysed single cells through mechanical sonication, and denatured proteins through high temperature.[27] Since chemicals were obviated, no further clean-up was applied. Detergent Rapigest was applied in preparation of mass-limited proteome samples.[38] Because Rapigest is compatible with enzymatic digestion and is degraded into non-interfering products under

an acidic condition. Organic solvent trifluoroethanol was also employed for preparation of mass-limited samples.[29,91] Trifluoroethanol can assist the cell lysis and protein denaturation. More importantly, it can be removed easily by lyophilization.

In 2014, Hughes et. al. introduced a single-pot solid-phase-enhanced sample preparation (SP3) method.[79] The method utilizes strong detergents (e.g., SDS) for protein extraction from cells. The cell lysates are incubated with carboxyl-coated paramagnetic nanoparticles under high concentration of acetonitrile (>70%). Proteins are captured on the beads through hydrophilic interaction and detergents can be removed efficiently via washing multiple times with organic solvents. Then the captured proteins are digested by enzymes on beads, followed by peptide elution from beads using an aqueous solution for LC-MS/MS analysis, **Figure 1.9E**. The fact that the paramagnetic nanoparticles have large surface area, ensuring strong interaction with proteins and supporting extensive washing. The SP3 method achieves sample preparation in a single Eppendorf tube, reduces sample loss during sample transferring, and enables preparation of mass-limited samples for proteomic analysis. Multiple studies have used SP3 method for processing mass limited sample and achieved good sensitivity on proteome profiling. Hughes et. al. processed single Drosophila embryos containing only 200 ng of proteins and identified almost 3000 proteins when using SP3 for sample processing.[79] Griesser applied SP3 to process proteins extracted from FFPE sample. From tissue containing about 3000 cells, over 5,600 protein IDs were confidently quantified.[92] Yang et. al. identified 3600 protein IDs from protein amount equal to 1000 HEK 293t cells using CE-MS/MS, when SP3 was used for sample processing.[93] We need to point out that the SP3 method is operated in Eppendorf tubes

and requires microliter-level solutions for sample preparation, resulting in limited performance for preparation of nanograms of proteome samples.

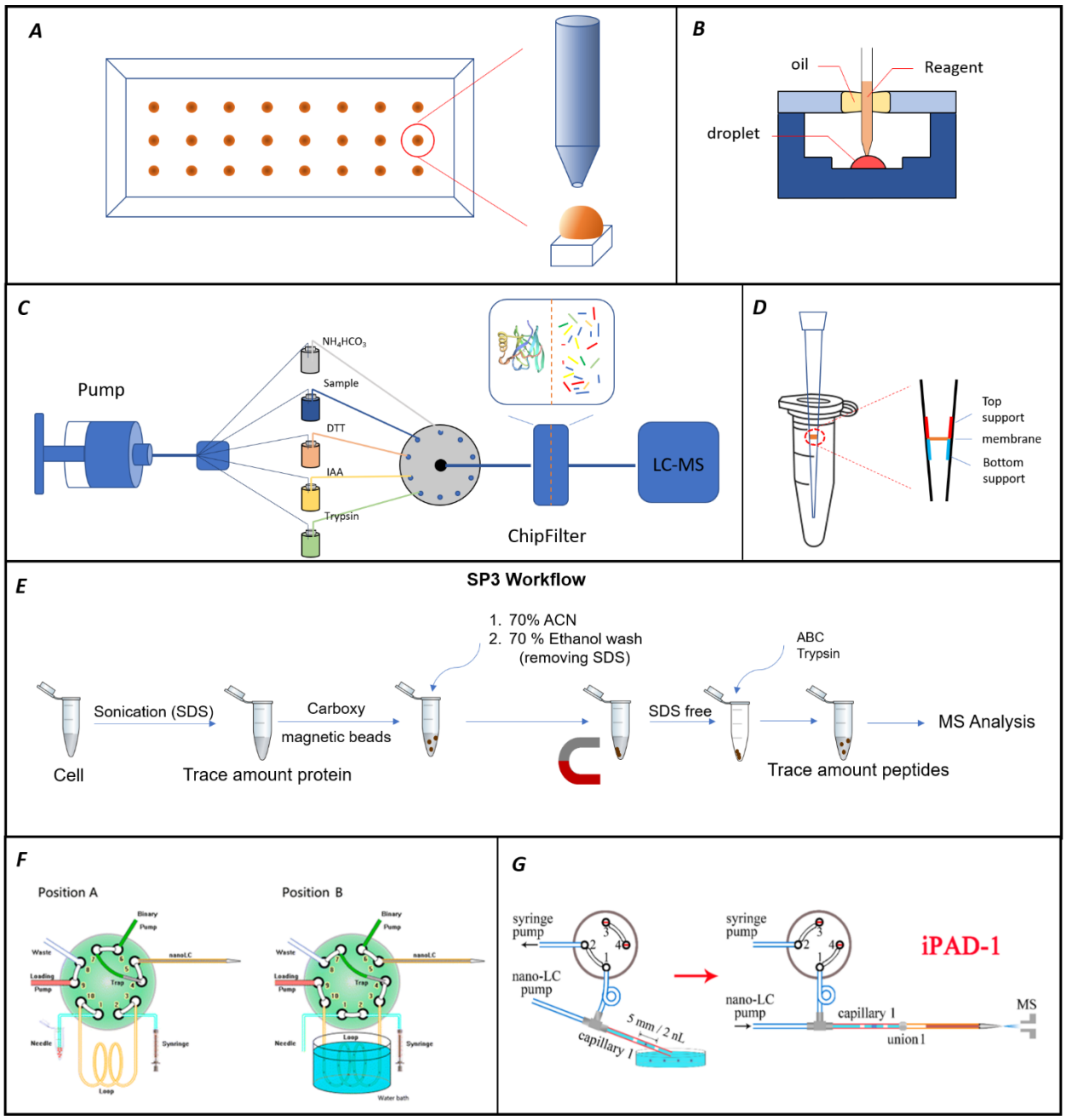

Figure 1. 9 Schematic diagrams of different sample preparation methods. (A) NanoPOTS, reproduced from ref. 38 with permission from Springer Nature, copyright (2018); (B) OAD, reproduced from ref. 86 with permission from American Chemical

Society, copyright (2018); (C) ChipFilter, reproduced from ref. 87 with permission from American Chemical Society, copyright (2020); (D) MicroFASP, reproduced from ref. 90 with permission from American Chemical Society, copyright (2020); (E) SP3; (F) iPAD, reproduced from ref. 88 with permission from American Chemical Society, copyright (2015); (G) iPAD-1, reproduced from ref. 89 with permission from American Chemical Society, copyright (2018).

1.4.1.3 Blocking the non-specific interactions

Blocking the non-specific hydrophobic adsorption of proteins/peptides on the surfaces of Eppendorf tubes, sample vials, and even beads packed in the LC columns has also been implemented in proteomic analyses of mass-limited samples. Yang et. al. proved that treatment of the sample vials with BSA improved the number of protein IDs by 40% and peptide intensity by 4-fold compared to that without treatment when only nanograms of peptides were in the sample vials for MS analysis.[93] BSA peptides were also used to block the hydrophobic adsorption of beads packed in reversed-phase LC columns before the columns were used for LC-MS/MS analysis of trace amounts of peptides.[94] Similar treatment of LC columns with an E. coli digest was also reported when mammalian proteome samples were analyzed.[86] Dou et al. applied n-Dodecyl β-D-maltoside (DDM) (0.01%) as collection buffer additive in the process of fraction collection and achieved significantly higher numbers of protein IDs compared to that without the DDM additive.[95] This outcome is because of the reduced interaction between peptides and the collection device surface in the presence of low-concentration DDM.[96] Another effective approach for reducing non-specific binding between peptides to potentially interacting surface is to introduce carrier peptides into the peptide

samples. The carrier peptides have much higher concentration than the peptides in the samples and are responsible for blocking the hydrophobic adsorption. One of the perfect examples is the Single Cell ProtEomics by Mass Spectrometry (SCoPE-MS) method developed by Budnik et. al.[27] In the SCoPE-MS method, one of the TMT channels is used to label the carrier sample and other TMT channels are used to label the target mass-limited samples (i.e., single mammalian cells). The carrier sample derive from the same source of protein materials as the single cell samples but with a significant higher peptide amount. The use of a high concentration of TMT-labeled carrier sample not only reduces the peptide loss of mass-limited samples but also boosts the peptide signal in mass spectra, facilitating the identification and quantification of peptides from the mass-limited samples. In addition, the involving of TMT labeling also improves the throughput of proteomic analysis of trace proteome samples, such as single cells.[97,98]

1.4.2 Separation in proteomics of mass-limited sample

Coupling liquid-phase separations (LC and CE) to MS and MS/MS plays vital roles in comprehensive and sensitive profiling of complex proteomes with high concentration dynamic ranges.[26] Millions of different peptide molecules exist in a typical proteome digest. Highly efficient separation of peptides before MS and MS/MS reduces ionization suppression and boosts the sensitivity of peptide measurement. The separation science provides an efficient approach to reduce the complexity of the proteome and affords more time for the mass spectrometer to analyze the proteome. The most widely applied separation strategies in bottom-up proteomics are RPLC,

which separates peptides based on hydrophobicity and CE, which separates peptides based on size to charge ratio.

1.4.2.1 Applying liquid chromatography in proteomics of mass-limited samples

In RPLC, a nonpolar stationary phase and a polar mobile phase are employed to facilitate peptides and proteins separation. Retention of peptides, which is determined by the distribution of peptides between stationary phase and mobile phase, highly correlates with the hydrophobicity of peptides. C18 is the most frequently used stationary phase for peptides separation. It is an octyldecylsilane that contains 18 carbons bound to the silica. For packed RPLC column, C18 material is covalently bound to porous amorphous silica microparticles. Aqueous mobile phase favors peptides trapping onto the stationary phase. The elution of the peptides is then determined by their hydrophobicity and the composition of organic and aqueous mobile phase. Using such binary combination of mobile phases affords flexibility to control elution strength and selectivity. Instead of using isocratic elution, where the separation is carried out through fixed proportion of mobile phases, a gradient elution with programmed changes of organic mobile phase is usually applied in bottom-up proteomics due to high sample complexity.[99-101] Gradient elution facilitates retention of poorly retained peptides and elution of strongly retained peptides with high selectivity and affords high resolution of peptides separation. Acetonitrile (ACN) is one of the most frequently used organic modifier as organic mobile phase due to its low viscosity that can help to reduce back pressure, high strength of elution, low reactivity and its dissolving capacity to peptides.[102]

In separation science, the separation efficiency can be evaluated through height equivalent theoretical plate (HETP). The HETP can be influenced by various factors described in Van Deemter equation.[103]

$$\boldsymbol{HETP = A + \frac{B}{u} + Cu} \qquad \textbf{Eq. 1.1}$$

In Van Deemter equation, A-term correlates with Eddy-diffusion parameters which is related to the multiple flow paths due difference of the packing particles and non-ideal packing in the column. B-term correlates with the longitudinal diffusion, resulting from dispersion of molecules along the column axis from high concentration to low concentration. The longitudinal diffusion is inverse proportional to the flow rate u. the C-term correlates with resisting to mass transfer, which is the major source of peak broadening.[104] Resisting to mass transfer describes the different extent of the analytes penetrate into the porous packing particles and difference of diffusion through "stagnant" mobile phase back to the surface of the particles. the C-term is proportional to flow rate, because slower the flow rate is, the difference of mass transfer of analytes becomes smaller. With all terms considered, separation efficiency of RPLC can be improved through application of universally distributed packing particles, a flow rate modest to the column size, and small packing particles that can reduce paths of analytes diffusion. High separation efficiency is critical to improve sensitivity of proteomics analysis.[105,106]

Conventional RPLC system applies separation column with i.d. at millimeter level and correspondingly requires optimum flow rate of hundreds of microliters to low milliliter per min. The large i.d. of separation column and high flow rate are detrimental for trace proteomics analysis for following reasons: first, the large inner diameter

requires higher amount of packing material and consequentially higher surface area the analyte will interact with. This could introduce more sample loss due to unavoidable hydrophobic adsorption. Second, the high flow rate causes sample dilution during transferring in the LC system. Third, a high flow rate causes low ionization efficiency due to large droplet size formed during ESI process. To improve sensitivity, a nanoLC is usually applied for proteomics analysis of mass-limited samples.[34,36, 106-112] Researchers have deployed RPLC columns with narrow inner diameter to accommodate lower flow rate of LC separation and to increase the sensitivity of LC-MS. It has been pointed out that the sensitivity of LC-MS can be boosted by decreasing the inner diameter (d) of the RPLC column from d1 to d2, and the improvement factor (f) is equal to $d1^2/d2^2$.[113] The i.d. of RPLC columns has been decreased from 75 µm as the most routinely used for RPLC-MS to 30 µm or smaller. [34,35,38,39,114,115] The flow rate is on the order of tens of nanoliter per min to even picolitre per min. Extremely high sensitivity has been achieved for bottom-up proteomics by applying such small i.d. RPLC columns. In 2004, a trace proteomics study was performed by coupling a 15-µm-i.d. RPLC column with an 11.4 T Fourier Transform Ion Cyclotron Resonance (FT-ICR) mass spectrometer. The flow rate of the system was controlled at 20 nL/min, and a peak capacity of 103 was achieved.[34] About 14% of the total proteome of Deinococcus radiodurans were confidently identified when only 2.5 ng of peptides were consumed. The sensitivity was proved to be at the low zmol level, corresponding to thousands of peptide molecules. Recently, 20-µm-i.d. RPLC columns have been applied to the characterization of peptides of single human cells processed by the NanoPOTS method, enabling the identification of over 300 proteins from single cells by MS/MS.[39] Porous layer open tube (PLOT) columns with

small i.d. have also been employed for proteomics of mass-limited samples.[35,114] In PLOT, a poly(styrene - divinylbenzene) in-situ polymerization was implemented in a 10-µm-i.d. capillary column with polymer thickness of 1-2 µm, **Figure 1.10**. The thin wall of the polymer as stationary phase significantly reduces the resistance to mass transfer and boosts the separation efficiency.[114] By applying a 10-µm i.d. PLOT column, 4,000 proteins were identified with the consumption of peptides corresponding to 100-200 cells.[35] Recently, Xiang et al. introduced an extremely narrow open tubular (NOT) column with only 2-µm i.d. for trace proteomics analysis.[115] The 2-µm-i.d. NOT column was operated at a flow rate lower than 1 nL/min and achieved over 1,000 proteins IDs when only 75 pg of tryptic peptides were loaded onto the system, demonstrating the tremendous potential for single-cell proteomics.

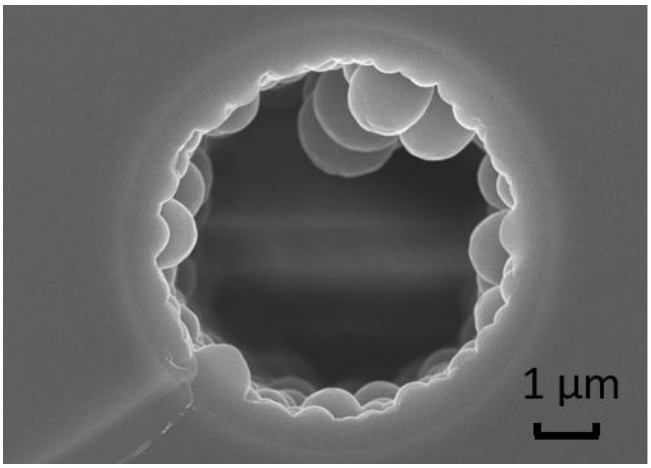

Figure 1. 10 SEM image of PLOT capillary cross section.

1.4.2.2 Applying capillary electrophoresis in proteomics of mass-limited samples

Capillary electrophoresis is another highly efficient separation strategy that is widely applied in proteomics.[116-120] The separation is performed within narrow-bore capillary (usually 10-75 µm) filled with electrolyte buffer (i.e., background electrolyte (BGE)). During separation, high electric voltage is applied across the capillary. At a

certain pH, zwitterion (i.e., peptides and proteins) carries either a net charge (positive/negative) or no charge based on the isoelectric point (i.e., PI) of the zwitterion. The voltage applied across the capillary drives positively charged ions towards cathode and negatively charged ions towards anode. Separation is based on different size to charge ratio of ions in capillary. Fused silica capillary is mostly used in CE separation. On the inner wall of the capillary, there are ionizable silanol group with pKa about 2-3 (borosilicate) or 4-5 (silica).[121-123] When the pH of BGE is higher than pKa, the negatively charged wall will attract positively charged ions from BGE in the capillary, creating an electrical double layer at the capillary wall interface. When high voltage is applied, the positive ions enriched at the wall interface will move towards the cathode end and carry the solvent alongside the movement (**Figure 1.11**). The resulting flow of bulk solvent is called electroosmotic flow (EOF). EOF is directly proportional to the charge density and the thickness of the double layer on the capillary inner wall. The migration time of ions through the length (L) of capillary is given by:

$$\boldsymbol{t = L^2/V\left(\mu_{electrophoresis} + \mu_{electroosmosis}\right)} \qquad \textbf{Eq. 1.2}$$

where V is the applied potential, $\mu_{electrophoresis}$ is the electrophoretic mobility of analytes and $\mu_{electroosmosis}$ is mobility caused by EOF. EOF can be manipulated through control the pH of the BGE and by applying neutral coating on the inner surface of the capillary. EOF can deliver rapid separation when un-coated capillary was applied. However, this also leads to short separation window and insufficient separation when analytes are not fully separated by their electrophoretic mobility. For complex proteome such as cell lysates, short separation window can be detrimental to number of identification of peptides and proteins. Neutral coating with hydrophilic polymer such as Linear

polyacrylamide (LPA) is a commonly used coating to effectively reduce EOF and minimize peptides interaction with capillary inner walls (**Figure 1.11)**.[124-126]

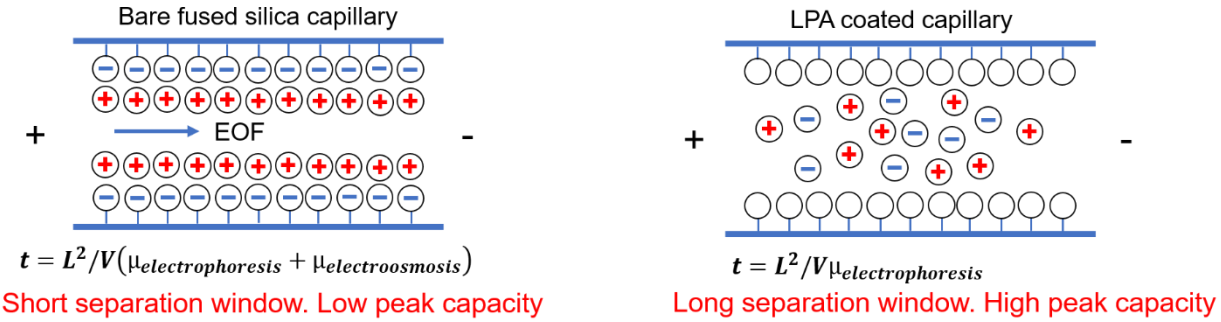

Figure 1. 11 Elimination of EOF through LPA coating.

CZE-MS has shown great potential for proteomic analysis of mass-limited samples. Multiple studies have compared the performance of CZE-MS and RPLC-MS for discovery proteomics in terms of sensitivity.[116-119,127] It has been found that CZE-MS outperforms RPLC-MS regarding the number of peptide and protein IDs when the sample size is smaller than 10 ng. The higher sensitivity of CZE-MS can attribute to several reasons. First, CZE can achieve very high separation efficiency for large biomolecules. In idealized systems, longitudinal diffusion is the only contribution for peak band broadening. The number of theoretical plates N is given by:

$$N = V\mu_{total}/2D \qquad \textbf{Eq. 1.3}$$

Where D is diffusion coefficient. V is applied potential. μ_{total} is the sum of $\mu_{electrophoresis}$ and $\mu_{electroosmosis}$. Applying high voltage leads high separation efficiency for diffusion limited molecules such as proteins and peptides. One million of theoretical plates have been achieved by CZE for separation of proteins.[128] Second, the flow rate in CZE separation is on the order of low nL/min when capillaries with neutral coatings are employed, ensuring the high electrospray ionization efficiency of peptides and proteins.

Third, the absence of stationary phase and direct sample injection from a sample vial without a valve and transferring tubing reduce sample loss. Low zmole peptide detection limits have been reported using CZE-MS.[33,129]

CZE-MS typically has excellent mass detection limit but poor concentration detection limit because of the low sample loading capacity of CZE (low nL). The low sample loading capacity had impeded CZE-MS/MS for large-scale proteomics of mass-limited samples. During the recent years, various online concentration approaches have been evaluated for boosting the sample loading capacity of CZE-MS for large-scale proteomics of nanograms of proteome samples. Sun et al. applied CZE-MS/MS for large-scale proteomic analysis of a HeLa cell proteome digest with a filed enhanced sample stacking method for increasing the sample loading volume (100 nL), identifying 10,000 peptides and 2,000 proteins from only 400 ng of Hela digest.[130] The principle of field enhanced sample stacking is simple. When the conductivity of the sample plug is lower than that of BGE, with same current across the capillary, the electric field is a lot higher in the sample plug. Ions in the sample plug as a result, will migrate faster and get concentrated at the boundary of sample plug and BGE. For field enhanced sample stacking, sample buffer usually contains organic solvent (e.g., 50% ACN) to reduce conductance.

Chen et al. optimized the dynamic pH junction-based sample stacking method for CZE-MS/MS according to the work reported by the Dovichi group[131] and reported a CZE-MS/MS system with a microliter-scale sample loading volume and an over 2-hours separation window, establishing the foundation of using CZE-MS/MS for large-scale proteomics.[132] When dynamic pH junction is applied for sample stacking, sample is

usually dissolved in basic buffer (e.g., 50 mM NH_4HCO_3, pH 8.0) and the BGE is usually acidic buffer (5% acetic acid, pH 2-3). During separation, analytes (i.e., peptides) are negatively charged and move towards positive electrode. Peptides are then neutralized and becomes positively charged when they contact the acidic BGE buffer at the positive electrode end. Positively charged peptides move towards negative electrode and back into the sample buffer and are neutralized again. The back-and-forth movement concentrate peptides at the moving pH boundary I (**Figure 1.12**).

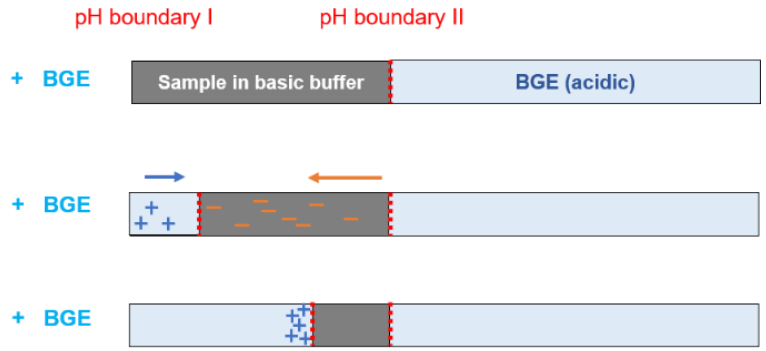

Figure 1. 12 Schematic illustration of the simplified mechanism of dynamic pH junction.

More recently, the dynamic pH junction-based CZE-MS/MS has shown great potential for large-scale proteomics of mass-limited samples.[93,133] The CZE-MS/MS system identified on average 100 proteins via consuming only 250 pg of a MCF7 proteome digest, corresponding to the protein content of roughly one MCF7 cell in mass.[133] Yang et al. reported the identification of over 6500 proteins from a MCF7 cell lysate starting with only 500-ng peptides via coupling the nanoRPLC fractionation with the dynamic pH junction-based CZE-MS/MS.[93] The well orthogonal separations of nanoRPLC and CZE for peptides guarantee the high peak capacity of the nanoRPLC-CZE-MS/MS platform for bottom-up proteomics, **Figure 1.13**. In the same work,

coupling the SP3 sample preparation method and nanoRPLC-CZE-MS/MS enabled the identification of nearly 4000 proteins from 5000 HEK293T cells with the consumption of a peptide amount that corresponded to only roughly 1000 cells.

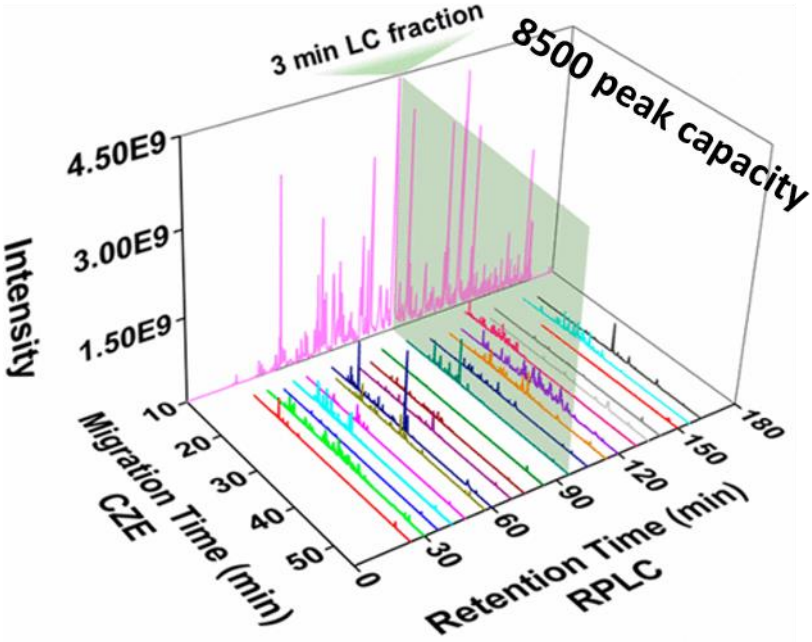

Figure 1. 13 High orthogonality affords high peak capacity of LC-CE-MS system. Reproduced from ref. 93 with permission from American Chemical Society, copyright (2019).

Besides online stacking, coupling SPME (solid phase microextraction) to CZE was also explored in multiple studies to improve the sample loading capacity. The SPME-CZE-MS theoretically can eliminate the limitation of sample loading volume of CZE as peptides can be preconcentrated by the SPME material before CZE-MS. In addition, the eluted sample plug from SPME can be further online concentrated with transient isotachophoresis (tITP)[119] and dynamic pH junction via carefully choosing the elution buffer and separation buffer.[134,135] The SPME-CZE-MS has shown over 3-folds more protein IDs than nanoRPLC-MS when only 5-ng peptides were used, clearly

demonstrating the advantage of SPME-CZE-MS for trace proteome samples.[119] Zhang et al. reported the identification of over 1000 proteins from 50-ng of Xenopus laevis proteome digest with the SPME-CZE-MS/MS.[135]

Interface that can couple CE separation with MS is crucial for high sensitivity and stability of CE-MS performance on mass-limited sample analysis. The interface completes the electrical circuit for CE separation and provides the voltage for ESI of MS. Multiple designs of interface have been reported.[126,136,137] Our lab uses the electro-kinetically pumped sheath flow interface with diagrams shown below. Different generations of the interface have been developed to further increase the sensitivity and robust operation (**Figure 1.14**).[138-140] A High voltage is applied in the sheath buffer reservoir and an EOF in the glass emitter is produced. Sheath liquid in the buffer reservoir is driven by the EOF at nL/min flow rates through the emitter for ESI. As shown in the design progress across the generations of the interface, the size of the emitter orifice was enlarged and the distance between the capillary end to the orifice was reduced. This third generation of the interface significantly increases the sensitivity by reducing sample dilution from the sheath buffer and improved the robustness of the interface.

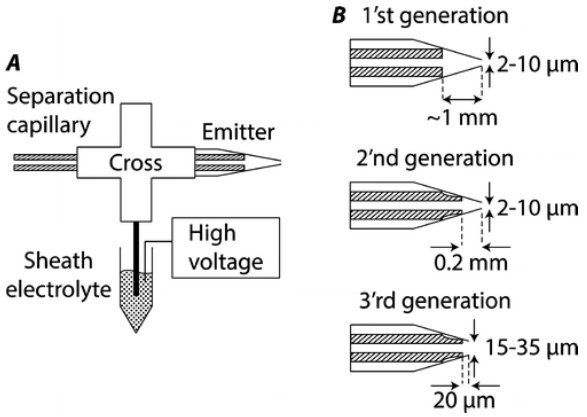

Figure 1. 14 Diagrams of the basic design of the electrokinetically pumped sheath flow CE-MS interface (A) and its three different generations (B). Reproduced from ref. 140 with permission from American Chemical Society, copyright (2015).

1.4.3 Mass spectrometer in proteomics of mass-limited sample

Mass spectrometry is one of the most important technology applied in proteomics study. Modern mass spectrometers for proteomics analysis compose multiple mass analyzers, usually with a high-resolution analyzer such as ion cyclotron resonance (ICR), orbitrap or time-of-flight (TOF) and a low-resolution analyzer such as linear trapping quadrupole (LTQ) or quadrupole. The high-resolution analyzers provide highly accurate (i.e., low ppm) and highly precise (i.e., ± 0.001 m/z) mass measurement of ions and enable accurate assignment of molecular identity and charge states determination. The low-resolution analyzers afford fast scan and efficient ion accumulation, isolation and transferring. The hybrid configuration of mass analyzer offers complementary advantages for fast peptide sequencing and deep proteome profiling.[141]

Various types of mass spectrometers have been explored on their performance on proteomics of mass-limited samples.

1.4.3.1 Fourier transform ion cyclotron resonance mass spectrometer (FTICR)

ICR is categorized as Fourier Transform Mass Spectrometer (FTMS). It measures the sinusoidal current that is produced by ion's motion. When ions packets with different m/z are transferred into an ion cyclotron resonance (ICR) cell applied with high magnetic field, ions' motion is produced by the Lorenz force exerted on ions.[142,143] Ions are then excited by applying radio frequency pulse (RF pulse) to opposing electrodes. When the frequency of ions cyclotron matches the RF pulse frequency, ions resonance occurs, and they are accelerated and excited to a larger cyclotron radius of motion so that the radial frequency is kept same. The length of RF pulse determines the final radius of ions cyclotron so that they are close enough to the electrodes for sensitive detection, while not strike to the electrodes. When multiple packets of ions with different m/z are analyzed, the RF pulse is performed through frequency sweep so that all ions are excited.[144] When the excitation amplifier is turned off, ions continue orbiting in the ICR cell with frequency directly related mass to charge ratio as given in formula:

$$\boldsymbol{f = \frac{qB}{2\pi m}} \qquad \textbf{Eq. 1.4}$$

where q is the charge of the ion, B is the magnetic field, m is the mass of the ion. The orbiting ions oscillate between opposing electrodes and induce oscillating charge image current and produce ICR signal (time-domain). The ICR signal is then Fourier transformed into frequency domain spectrum that can be further transformed into mass domain spectrum (**Figure 1.15**).[143] FTICR mass spectrometer is famous for its high resolution and high mass accuracy due to superior stability and uniformity of the magnetic field.[145] The high resolution makes it possible for the instrument to resolve target molecules and interference at detection limit level. Taking advantage of high

accuracy and precision of mass measurement of FTICR MS, it is possible to identify proteins by only measuring the intact mass of produced peptides without tandem MS/MS measurement. The idea is called accurate time and mass (AMT) tag.[146] It assumes that molecular masses of many peptides from enzyme digestion are unique enough among all possible peptides produced from an annotated genome for protein identification.[145] Briefly, a reference database with AMT tag information specific to a biological sample (e.g., human, mouse etc.,) is first established. The database searching only relies on highly accurate precursor mass (sub ppm level) and LC elution time for peptide matching. Avoidance of tandem MS/MS improved MS scan rate and dynamic range of peptide identification. The inherent advantages of FTICR MS makes it a promising instrument for highly sensitive analysis of biological sample. A single cell analysis with detection of hemoglobin at attomole level from a single erythrocyte using FTICR-MS was first reported in 1996.[147] Later in 2001, sensitive detection of peptide mixture at concentration of amol/μL from complex matrix was reported using FTICR.[148] Multiple studies have used LC-FTICR-MS analysis with AMT approach for highly sensitive analysis of proteome with only picograms of protein extract and achieved zeptomoles detection of proteins with high dynamic range.[34, 149]

1.4.3.2 Orbitrap-base mass spectrometer

Orbitrap is also a FTMS analyzer and was first introduced to public commercially in 2005 as a hybrid instrumentation, LTQ orbitrap.[150] The orbitrap includes an outer “cup” like electrode and a central “spindle” like electrode. With voltage applied on the outer and central electrode, a combination of a radial electric field and an axial electric field is created. Ions oscillate along the axial direction and orbit around the central

electrode at the same time. Outer electrodes receive the oscillating image current of the ions' axial oscillation. Similar to ICR, the image current is Fourier transformed into frequency domain spectrum and then mass domain spectrum (**Figure 1.15**).[32] The introduction of orbitrap to the proteomics community advanced the development of proteomics regarding comprehensive proteome profiling, high throughput, and high sensitivity. Single ion detection was proved on orbitrap instrumentation[151,152] indicating the potential of orbitrap MS in single cell proteomics.

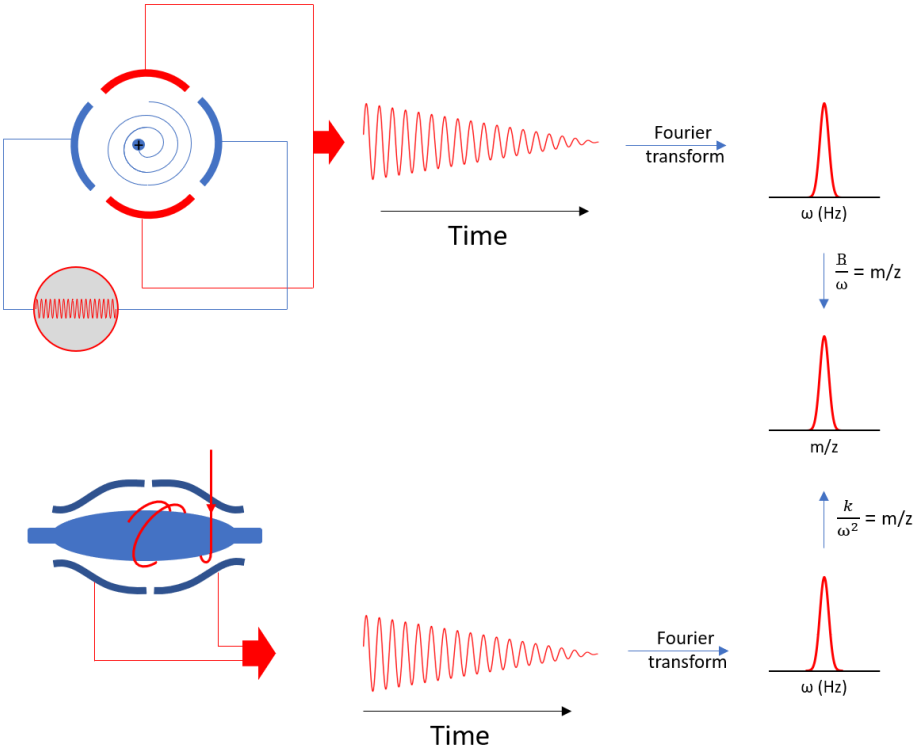

Figure 1. 15 FT-MS. Mass to charge ratio of ion in FT-MS is related to frequency of oscillation of the ion in the ion cyclone cell. The oscillation frequency is obtained from Fourier transform of the image current induced by ion motion. The relationship of frequency and m/z is indicated in the figure.

Orbitrap-based instruments have been used in multiple laboratories for proteome profiling of mass-limited sample. Since the first commercial instrument incorporated with

Orbitrap analyzer was introduced in 2005, successive generations of Orbitrap-based instruments have been developed with improved scan speed, sensitivity, resolution and duty cycle.[153] For example, the introduction of S-lens, a stacked ring ion guide[154] improved the ion transfer efficiency and boosted the sensitivity up to 10-fold compare to the original Orbitrap system coupled with a capillary-skimmer interface for ion introduction. The installation of higher energy collision-induced dissociation (HCD)[155] cell as a replacement of CID allowed beam-type collision dissociation of ions, which lowered the cutoff of fragment mass. Enhanced Fourier Transform (eFT) technique[156] was implemented on Orbitrap system to achieve high acquisition rate without compensation of resolution. The introduction of high field Orbitrap to the later model of Orbitrap system instrumentation, further increased the resolution and acquisition rate of the instrumentation.[157,158] Advanced precursor determination algorithms was incorporated in Orbitrap instrumentation so that overlapping isotopic envelop can be annotated more efficiently to determine charge states and assign monoisotopic m/z, which is critical to trigger MS/MS with optimal collision energy and ensures fully usage of TopN loop count.[159] With all the improvements, Up to 40 Hz acquisition rate (at resolution setting of 7500 at m/z 200) and half million resolution at m/z of 200 were achieved in latest model of orbitrap instrumentation. The superior resolving power, scan rate, and sensitivity make Orbitrap instrumentation the most widely used MS in proteomics analysis of trace material, especially single cell.[38,39,86,160]

Ion mobility mass spectrometry separates gas-phase ions by their mobility within gas media and provides additional dimension of separation of ions on tops of liquid-phase separation (i.e., LC/CE) and mass spectrometry. High field asymmetric waveform

ion mobility spectrometry (FAIMS) separates gas-phase ions based on their characteristic difference in mobility in high and low electric field.[161] For mass-limited yet complex sample such as proteome of single cell, FAIMS provides many advantages. First, FAIMS can filter out charge-one ions and affords MS spectrum with low background noise. Second, FAIMS provides a reproducible online fractionation of ions. For mass-limited sample such as single cell, online fractionation plays vital role to reduce sample complexity that overwhelms MS duty cycle and resolving power, especially when off-line fractionation is impractical.[162] FAIMS was not widely applied in MS instrumentation until recently because it attenuated ion signal up to one order of magnitude, and because the long ion transmission time of FAIMS resulted in delay switching of compensation voltage (CV). In 2018, Pfammatter et al. introduced a novel FAIMS device with shorter CV switch time and improved sensitivity and integrated it with the Orbitrap Tribrid mass spectrometer.[163] The novel FAIMS-Orbitrap platform achieved better detection of low-abundance peptides that were underrepresented in the platform without FAIMS and effectively reduced ratio compression effect in TMT quantification.[164] The fractionation ability of FAIMS was proved by Hebert et al.[165] In a 6 h single-shot LC-MS analysis with FAIMS on human cell line digest, 8151 proteins were identified, where 7776 proteins were identified from 4 LC-fractions analyzed with 1.5 h each. FAIMS-interfaced MS instrumentation has shown superior performance on proteomics analysis of trace material. More than 1000 proteins were identified from 5 ng of Hela digest with only 5 min gradient using FAIMS-MS.[28] More than 1000 protein IDs (by MS/MS only) were identified from a single Hela cell using FAIMS-MS, which is 2.3 times more identification than the platform without FAIMS.[160]

1.4.3.3 Time-of-flight mass spectrometer

Alternative to FTMS, time of flight mass spectrometer has also been used in bottom-up proteomics laboratories. TOF MS separates ions based on their velocity as they travel through a flight tube. Once introduced into MS, the charged ions are accelerated by electric field and are imparted with kinetic energy (KE). Ions then move into a field-free region where the only force driving the ion movement is the initial KE.[166] The time ions spend in the flight tube is determined by mass to charge ratio, as indicated in the equation:

$$\boldsymbol{t = D\sqrt{\frac{m}{2zeV}}} \qquad \textbf{Eq. 1.5}$$

Where D is the distance to the detector, m is the mass, z is the charge of ions, eV is the initial voltage applied. TOF MS measures the time ions reach detector and convert it into mass spectrum.

TOF MS has extremely high scan speed (> 1000 Hz) and is absence of space charge effect (a limitation of usable ions in trapping instrument), determining its potential of high sensitivity and wide dynamic range for proteomics study. Typical configuration of TOF-MS employed in proteomics laboratories includes a quadrupole for ion transferring and isolation, collision cell to generate fragment ions, an orthogonal acceleration unit to reflectron, and a multichannel plate (MCP) as detector.[167] Reflectron changes the trajectories of ions movement. Ions with higher KE will penetrate deeper in reflectron so that ion with same m/z but lower KE can catch up of flight path.[168] Flight time of ions with same m/z gets averaged and the peak broadening gets decreased. Thus, application of reflectron improves resolution of TOF-MS. Multiple improvement on TOF instrument including collision cell, drift region, reflectron and MCP detector improved the

ion transferring efficiency, reduced spatial distribution of ions, and improved data acquisition speed. Resolution of 40,000 at 1222m/z and mass accuracy of 1.45 ppm on average were achieved in TOF MS.[169] The extreme high scanning speed also makes possible to interface TOF instrument with trapped ion mobility mass spectrometry (TIMS). TIMS separates gas-phase ions based on collision cross section of ions. The separation is typically performed within 10s or 100s milliseconds.[170] The extremely high scan speed (scan/sub-millisecond) of TOF-MS makes it perfect to be coupled with TIMS since ions emerged from TIMS can be efficiently sampled by TOF analyzer. Bruker's timsTOF pro is a representative of interfacing TIMS with TOF-MS. By applying Parallel Accumulation - SErial Fragmentation (PASEF)[171] in timsTOF pro, peptides sequencing speed is significantly improved without loss of sensitivity. By using timsTOF pro, more than 2500 proteins were identified from 10 ng of Hela digest within 30 min of acquisition time.[172] Very recently, timsTOF pro was applied for single cell proteomics study. Over 800 proteins were identified from a single Hela cell and over 420 single Hela cells were analyzed, representing one of the largest data sets of single cell proteomics.[173]

1.4.4 Applications of the techniques to single-cell proteomics

The advance of techniques regarding sample preparation, separation and MS instrumentation enables researchers to explore proteome difference at single-cell resolution. Lombard-Banek et al. used a capillary microprobe to sample cellular proteins from single blastomeres in Xenopus early-stage embryos, followed by CZE-MS/MS-based label-free quantitative bottom-up proteomics.[174] Significant proteome differences were observed at the single-cell level between blastomeres collected from the animal pole and vegetal pole, **Figure 1.16A**. Brunner et al. processed FACS isolated single

cells in 384-well plate with very small processing volume (1-2 μL) and applied an improved LC-MS platform to study proteome difference in cells at different cell-cycle stages.[173] Cell-cycle of HeLa cells was arrested by drug treatment to produce four cell populations enriched in specific cell-cycle stages. Although all cells were HeLa cells, significant proteome differences were observed between cells at different cell-cycle stages, **Figure 1.16B**. This single-cell proteomics result also reflected different protein amounts in cells at different proliferation states by summarizing identified peptides intensity, **Figure 1.16C**. Significant cell-to-cell heterogeneity at the proteome level was also demonstrated in a single cell proteomics study of progenitor cell and descendant cell.[175] In this study, Zhu et al. processed individual hair cells and its progenitor, supporting cells, with NanoPOTS, and performed single-cell proteomics of these two kinds of cells with LC-MS equipped with an ultranarrow bore separation column (30 μm i.d.). By using FM1-43 as a labeling reagent (a membrane probe for identifying actively firing neurons), hair cells and supporting cells were distinguished through fluorescence-activated cell sorting, based on the fact that hair cells can be labeled more strongly by FM1-43 than supporting cells. The NanoPOTS-assisted single-cell proteomics study identified 60 proteins from a single hair cell and 600 proteins from a pool of 20 hair cells. By checking the identified proteins from a pool of 20 cells in each population (hair and supporting cells), different proteins were significantly enriched in different cell types, **Figure 1.16D**. Specht et al. applied SCoPE2, an optimized version of SCoPE method, for proteomics analyses of 1,490 single cells (monocytes and macrophage cells) and quantified over 3,000 proteins from those single cells.[176] Principle component analysis

of the large single-cell-proteomics data set clearly separated the two cell types and revealed cellular heterogeneity, **Figure 1.16E**.

Single-cell proteomics can shed light on the molecular mechanisms of cell differentiation. In the Lombard-Banek's work, cellular proteins were sampled from a blastomere called midline animal-dorsal cell (at 16-cell stage) and its descendant cells at different cell-division stages (at 32, 64, and 128-cell stages) prior to mid-blastula transition (MBT).[174] Before MBT, there is no de novo transcription. Since all descendant cells arise from the same founder cell, it is surprising to see the proteome change over embryo development as shown in the hierarchical cluster analysis result, **Figure 1.16F**. In Zhu's work, although less than 100 proteins were identified from a single cell due to the extremely small cell size, with sufficient sample size, developmental trajectories from various protein expression patterns at the single-cell level can be established, **Figure 1.16G**.[175] These single-cell proteomics studies revealed the protein expression dynamics during cell differentiation.

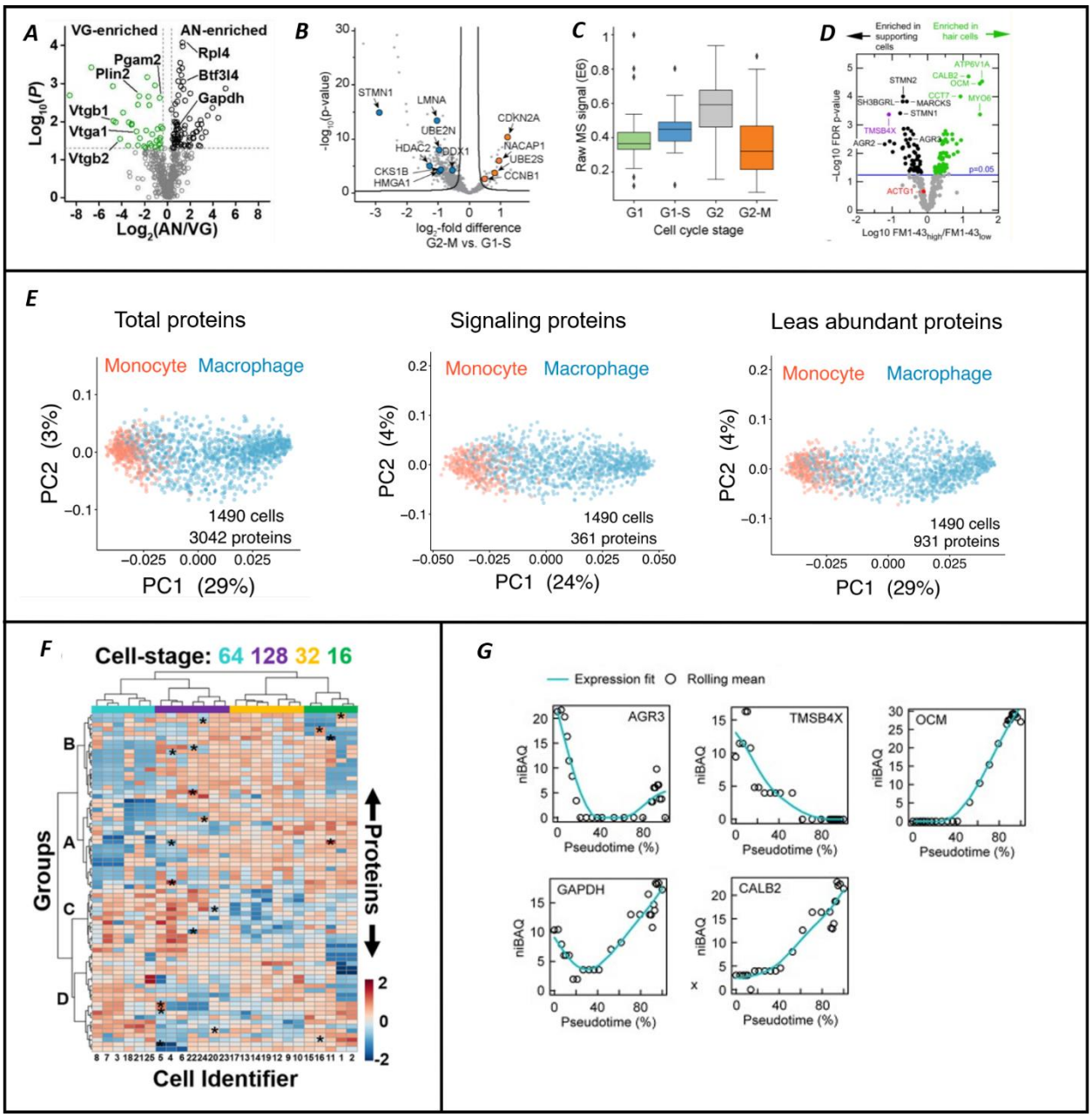

Figure 1. 16 Application of single cell proteomics. (A) Volcano plot revealing significant proteomic differences between blastomeres from animal and vegetal poles, reproduced from ref. 174 with permission from American Chemical Society, copyright (2019). (B) Volcano plot of the quantified proteins from single cells in the two drug arrested states, reproduced from ref. 173 with permission from bioRxiv, copyright (2020). (C) Violin plot of total protein signals of the analyzed single cells in the indicated

cell cycle stages as enriched by the drug treatments, reproduced from ref. 173 with permission from bioRxiv, copyright (2020). (D) Volcano plot showing significant proteomic differences between FM1-43high/FM1-43low cells (hair cell/supporting cell), reproduced from ref. 175 with permission from eLife, copyright (2019). (E) Weighted principal component analysis (PCA) of 1490 single cells using all 3,042 proteins, signaling proteins and least abundant proteins quantified across single cells. Cells are colored by cell type. The more spread-out of macrophage cells indicates significant cellular heterogeneity after differentiation from homogeneous monocytes cells, reproduced from ref. 176 with permission from BioMed Central, copyright (2021). (F) Hierarchical cluster analysis-heat map of quantified proteins from single Xenopus blastomeres isolated from various developmental stages, reproduced from ref. 174 with permission from American Chemical Society, copyright (2019). (G) Absolute expression dynamics (log2 niBAQ) of 5 proteins as a function of pseudotime, reproduced from ref. 175 with permission from eLife, copyright (2019).

1.5 Summary

This chapter introduced ultrasensitive proteomics analysis of mass-limited sample. Proteomics of mass-limited samples has been advanced aggressively in recent years because of significant technical progress in sample preparation, liquid-phase separation, and MS instrumentation. It took over 20 years from the detection of hemoglobin from single erythrocytes[177] to the identification of hundreds of proteins from a single HeLa cell[39,160,173,176] using MS. We expect that with further advancement of sample preparation methods and nanoRPLC/CE-MS platforms, the sensitivity and throughput of proteomics will be improved drastically, enabling routine proteomic

characterization of mass-limited samples (e.g., single cells) with high proteome coverage. The subsequent three chapters in the dissertation describe methodology development of ultrasensitive platforms for protein characterization of mass-limited samples. The last chapter focuses on evaluation of various sample preparation methods for top-down proteomics.

REFERENCES

REFERENCES

1. Kamme, F.; Salunga, R.; Yu, J.; Tran, D. T.; Zhu, J.; Luo, L.; Bittner, A.; Guo, H. Q.; Miller, N.; Wan, J.; Erlander, M., *The Journal of neuroscience : the official journal of the Society for Neuroscience* **2003,** *23* (9), 3607-15.

2. Georgiev, H.; Ravens, I.; Benarafa, C.; Forster, R.; Bernhardt, G., *Nature communications* **2016,** *7*, 13116.

3. Kalisky, T.; Quake, S. R., *Nature methods* **2011,** *8* (4), 311-4.

4. Lidstrom, M. E.; Meldrum, D. R., *Nature reviews. Microbiology* **2003,** *1* (2), 158-64.

5. Kimmel, C. B.; Ballard, W. W.; Kimmel, S. R.; Ullmann, B.; Schilling, T. F., *Developmental dynamics : an official publication of the American Association of Anatomists* **1995,** *203* (3), 253-310.

6. Wagner, D. E.; Weinreb, C.; Collins, Z. M.; Briggs, J. A.; Megason, S. G.; Klein, A. M., *Science* **2018,** *360* (6392), 981-987.

7. Cohen, A. A.; Geva-Zatorsky, N.; Eden, E.; Frenkel-Morgenstern, M.; Issaeva, I.; Sigal, A.; Milo, R.; Cohen-Saidon, C.; Liron, Y.; Kam, Z.; Cohen, L.; Danon, T.; Perzov, N.; Alon, U., *Science* **2008,** *322* (5907), 1511-6.

8. Marusyk, A.; Almendro, V.; Polyak, K., *Nature reviews. Cancer* **2012,** *12* (5), 323-34.

9. Dagogo-Jack, I.; Shaw, A. T., *Nature reviews. Clinical oncology* **2018,** *15* (2), 81-94.

10. Li, H. H.; Gyllensten, U. B.; Cui, X. F.; Saiki, R. K.; Erlich, H. A.; Arnheim, N., *Nature* **1988,** *335* (6189), 414-7.

11. Brady, G.; Barbara, M.; Iscove, N. In *Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies*, **1990**.

12. Kurimoto, K.; Yabuta, Y.; Ohinata, Y.; Ono, Y.; Uno, K. D.; Yamada, R. G.; Ueda, H. R.; Saitou, M., *Nucleic acids research* **2006,** *34* (5), e42.

13. Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B. B.; Siddiqui, A.; Lao, K.; Surani, M. A., *Nature methods* **2009,** *6* (5), 377-82.

14. Ramskold, D.; Luo, S.; Wang, Y. C.; Li, R.; Deng, Q.; Faridani, O. R.; Daniels, G. A.; Khrebtukova, I.; Loring, J. F.; Laurent, L. C.; Schroth, G. P.; Sandberg, R., *Nature biotechnology* **2012,** *30* (8), 777-82.

15. Islam, S.; Kjallquist, U.; Moliner, A.; Zajac, P.; Fan, J. B.; Lonnerberg, P.; Linnarsson, S., *Genome research* **2011,** *21* (7), 1160-7.

16. Jaitin, D. A.; Kenigsberg, E.; Keren-Shaul, H.; Elefant, N.; Paul, F.; Zaretsky, I.; Mildner, A.; Cohen, N.; Jung, S.; Tanay, A.; Amit, I., *Science* **2014,** *343* (6172), 776-9.

17. Klein, A. M.; Mazutis, L.; Akartuna, I.; Tallapragada, N.; Veres, A.; Li, V.; Peshkin, L.; Weitz, D. A.; Kirschner, M. W., *Cell* **2015,** *161* (5), 1187-1201.

18. Briggs, J. A.; Weinreb, C.; Wagner, D. E.; Megason, S.; Peshkin, L.; Kirschner, M. W.; Klein, A. M., *Science* **2018,** *360* (6392).

19. Jenuwein, T.; Allis, C. D., *Science* **2001,** *293* (5532), 1074-80.

20. Strahl, B. D.; Allis, C. D., *Nature* **2000,** *403* (6765), 41-5.

21. Turner, B. M., *BioEssays : news and reviews in molecular, cellular and developmental biology* **2000,** *22* (9), 836-45.

22. de Sousa Abreu, R.; Penalva, L. O.; Marcotte, E. M.; Vogel, C., *Molecular bioSystems* **2009,** *5* (12), 1512-26.

23. Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M., *Nature* **2011,** *473* (7347), 337-42.

24. Vogel, C.; Marcotte, E. M., *Nature reviews. Genetics* **2012,** *13* (4), 227-32.

25. Chandramouli, K.; Qian, P. Y., *Human genomics and proteomics : HGP* **2009,** *2009*.

26. Zubarev, R. A., *Proteomics* **2013,** *13* (5), 723-6.

27. Budnik, B.; Levy, E.; Harmange, G.; Slavov, N., *Genome biology* **2018,** *19* (1), 161.

28. Bekker-Jensen, D. B.; Martinez-Val, A.; Steigerwald, S.; Ruther, P.; Fort, K. L.; Arrey, T. N.; Harder, A.; Makarov, A.; Olsen, J. V., *Molecular & cellular proteomics : MCP* **2020,** *19* (4), 716-729.

29. Zhang, P.; Gaffrey, M. J.; Zhu, Y.; Chrisler, W. B.; Fillmore, T. L.; Yi, L.; Nicora, C. D.; Zhang, T.; Wu, H.; Jacobs, J.; Tang, K.; Kagan, J.; Srivastava, S.; Rodland, K. D.;

Qian, W. J.; Smith, R. D.; Liu, T.; Wiley, H. S.; Shi, T., *Analytical chemistry* **2019,** *91* (2), 1441-1451.

30. Wisniewski, J. R.; Hein, M. Y.; Cox, J.; Mann, M., *Molecular & cellular proteomics : MCP* **2014,** *13* (12), 3497-506.

31. Scigelova, M.; Hornshaw, M.; Giannakopulos, A.; Makarov, A., *Molecular & cellular proteomics : MCP* **2011,** *10* (7), M111 009431.

32. Zubarev, R. A.; Makarov, A., *Analytical chemistry* **2013,** *85* (11), 5288-96.

33. Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J., *Angewandte Chemie* **2013,** *52* (51), 13661-4.

34. Shen, Y.; Tolic, N.; Masselon, C.; Pasa-Tolic, L.; Camp, D. G., 2nd; Hixson, K. K.; Zhao, R.; Anderson, G. A.; Smith, R. D., *Analytical chemistry* **2004,** *76* (1), 144-54.

35. Li, S.; Plouffe, B. D.; Belov, A. M.; Ray, S.; Wang, X.; Murthy, S. K.; Karger, B. L.; Ivanov, A. R., *Molecular & cellular proteomics : MCP* **2015,** *14* (6), 1672-83.

36. Smith, R. D.; Shen, Y.; Tang, K., *Accounts of chemical research* **2004,** *37* (4), 269-78.

37. Sun, X.; Kelly, R. T.; Tang, K.; Smith, R. D., *The Analyst* **2010,** *135* (9), 2296-302.

38. Zhu, Y.; Piehowski, P. D.; Zhao, R.; Chen, J.; Shen, Y.; Moore, R. J.; Shukla, A. K.; Petyuk, V. A.; Campbell-Thompson, M.; Mathews, C. E.; Smith, R. D.; Qian, W. J.; Kelly, R. T., *Nature communications* **2018,** *9* (1), 882.

39. Cong, Y.; Liang, Y.; Motamedchaboki, K.; Huguet, R.; Truong, T.; Zhao, R.; Shen, Y.; Lopez-Ferrer, D.; Zhu, Y.; Kelly, R. T., *Analytical chemistry* **2020,** *92* (3), 2665-2671.

40. Ho, C. S.; Lam, C. W.; Chan, M. H.; Cheung, R. C.; Law, L. K.; Lit, L. C.; Ng, K. F.; Suen, M. W.; Tai, H. L., *The Clinical biochemist. Reviews* **2003,** *24* (1), 3-12.

41. Bruins, A. P., *Journal of Chromatography A* **1998,** *794* (1), 345-357.

42. Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S., *Analytical chemistry* **2013,** *85* (1), 2-9.

43. Bergmann, M.; Fruton, J. S.; Pollok, H., *Journal of Biological Chemistry* **1939,** *127* (3), 643-648.

44. Olsen, J. V.; Ong, S. E.; Mann, M., *Molecular & cellular proteomics : MCP* **2004,** *3* (6), 608-14.

45. Chait, B. T., *Science* **2006,** *314* (5796), 65-6.

46. Gregorich, Z. R.; Chang, Y. H.; Ge, Y., *Pflugers Archiv : European journal of physiology* **2014,** *466* (6), 1199-209.

47. Glish, G. L.; Vachet, R. W., *Nature reviews. Drug discovery* **2003,** *2* (2), 140-50.

48. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, *Nature biotechnology* **2001,** *19* (3), 242-7.

49. Wolters, D. A.; Washburn, M. P.; Yates, J. R., 3rd, *Analytical chemistry* **2001,** *73* (23), 5683-90.

50. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., 3rd, *Chemical reviews* **2013,** *113* (4), 2343-94.

51. Huang, Y.; Pasa-Tolic, L.; Guan, S.; Marshall, A. G., *Analytical chemistry* **1994,** *66* (24), 4385-9.

52. Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.; Hunt, D. F., *Biochimica et biophysica acta* **2006,** *1764* (12), 1811-22.

53. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., *Proceedings of the National Academy of Sciences of the United States of America* **2004,** *101* (26), 9528-33.

54. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W., *Journal of the American Chemical Society* **1998,** *120* (13), 3265-3266.

55. Craig, R.; Beavis, R. C., *Bioinformatics* **2004,** *20* (9), 1466-7.

56. Cox, J.; Mann, M., *Nature biotechnology* **2008,** *26* (12), 1367-72.

57. Eng, J. K.; McCormack, A. L.; Yates, J. R., *Journal of the American Society for Mass Spectrometry* **1994,** *5* (11), 976-89.

58. Hubbard, S. J., *Proteome bioinformatics*. Humana Press: Totowa, N.J., 2010.

59. Elias, J. E.; Gygi, S. P., *Methods in molecular biology* **2010,** *604*, 55-71.

60. Elias, J. E.; Gygi, S. P., *Nature methods* **2007,** *4* (3), 207-14.

61. Schubert, O. T.; Rost, H. L.; Collins, B. C.; Rosenberger, G.; Aebersold, R., *Nature protocols* **2017,** *12* (7), 1289-1294.

62. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., *Nature biotechnology* **1999,** *17* (10), 994-9.

63. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., *Molecular & cellular proteomics : MCP* **2002,** *1* (5), 376-86.

64. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., *Analytical chemistry* **2003,** *75* (8), 1895-904.

65. 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., *Molecular & cellular proteomics : MCP* **2004,** *3* (12), 1154-69.

66. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., *Proceedings of the National Academy of Sciences of the United States of America* **2003,** *100* (12), 6940-5.

67. Thompson, A.; Wolmer, N.; Koncarevic, S.; Selzer, S.; Bohm, G.; Legner, H.; Schmid, P.; Kienle, S.; Penning, P.; Hohle, C.; Berfelde, A.; Martinez-Pinna, R.; Farztdinov, V.; Jung, S.; Kuhn, K.; Pike, I., *Analytical chemistry* **2019,** *91* (24), 15941-15950.

68. Pichler, P.; Kocher, T.; Holzmann, J.; Mazanek, M.; Taus, T.; Ammerer, G.; Mechtler, K., *Analytical chemistry* **2010,** *82* (15), 6549-58.

69. Chelius, D.; Bondarenko, P. V., *Journal of proteome research* **2002,** *1* (4), 317-23.

70. Listgarten, J.; Emili, A., *Molecular & cellular proteomics : MCP* **2005,** *4* (4), 419-34.

71. Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M., *Molecular & cellular proteomics : MCP* **2014,** *13* (9), 2513-26.

72. Cleland, W. W., *Biochemistry* **1964,** *3*, 480-2.

73. Herbert, B.; Galvani, M.; Hamdan, M.; Olivieri, E.; MacCarthy, J.; Pedersen, S.; Righetti, P. G., *Electrophoresis* **2001,** *22* (10), 2046-57.

74. Carpenter, F. H., Treatment of trypsin with TPCK. In *Methods in Enzymology*, Academic Press: **1967**; Vol. 11, p 237.

75. Arribas, J.; Castano, J. G., *The Journal of biological chemistry* **1990,** *265* (23), 13969-73.

76. Botelho, D.; Wall, M. J.; Vieira, D. B.; Fitzsimmons, S.; Liu, F.; Doucette, A., *Journal of proteome research* **2010,** *9* (6), 2863-70.

77. Wessel, D.; Flugge, U. I., *Analytical biochemistry* **1984,** *138* (1), 141-3.

78. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., *Nature methods* **2009,** *6* (5), 359-62.

79. Hughes, C. S.; Foehr, S.; Garfield, D. A.; Furlong, E. E.; Steinmetz, L. M.; Krijgsveld, J., *Molecular systems biology* **2014,** *10*, 757.

80. Hughes, C. S.; Moggridge, S.; Muller, T.; Sorensen, P. H.; Morin, G. B.; Krijgsveld, J., *Nature protocols* **2019,** *14* (1), 68-85.

81. Pop, C.; Mogosan, C.; Loghin, F., *Clujul medical* **2014,** *87* (4), 258-62.

82. Zhu, Y.; Clair, G.; Chrisler, W. B.; Shen, Y.; Zhao, R.; Shukla, A. K.; Moore, R. J.; Misra, R. S.; Pryhuber, G. S.; Smith, R. D.; Ansong, C.; Kelly, R. T., *Angewandte Chemie* **2018,** *57* (38), 12370-12374.

83. Zhu, Y.; Dou, M.; Piehowski, P. D.; Liang, Y.; Wang, F.; Chu, R. K.; Chrisler, W. B.; Smith, J. N.; Schwarz, K. C.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Smith, R. D.; Qian, W. J.; Kelly, R. T., *Molecular & cellular proteomics : MCP* **2018,** *17* (9), 1864-1874.

84. Zhu, Y.; Podolak, J.; Zhao, R.; Shukla, A. K.; Moore, R. J.; Thomas, G. V.; Kelly, R. T., *Analytical chemistry* **2018,** *90* (20), 11756-11759.

85. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., *Journal of proteome research* **2011,** *10* (4), 1794-805.

86. Li, Z. Y.; Huang, M.; Wang, X. K.; Zhu, Y.; Li, J. S.; Wong, C. C. L.; Fang, Q., *Analytical chemistry* **2018,** *90* (8), 5430-5438.

87. Ndiaye, M. M.; Ta, H. P.; Chiappetta, G.; Vinh, J., *Journal of proteome research* **2020,** *19* (7), 2654-2663.

88. Chen, Q.; Yan, G.; Gao, M.; Zhang, X., *Analytical chemistry* **2015,** *87* (13), 6674-80.

89. Shao, X.; Wang, X.; Guan, S.; Lin, H.; Yan, G.; Gao, M.; Deng, C.; Zhang, X., *Analytical chemistry* **2018,** *90* (23), 14003-14010.

90. Zhang, Z.; Dubiak, K. M.; Huber, P. W.; Dovichi, N. J., *Analytical chemistry* **2020,** *92* (7), 5554-5560.

91. Wang, H.; Qian, W. J.; Mottaz, H. M.; Clauss, T. R.; Anderson, D. J.; Moore, R. J.; Camp, D. G., 2nd; Khan, A. H.; Sforza, D. M.; Pallavicini, M.; Smith, D. J.; Smith, R. D., *Journal of proteome research* **2005,** *4* (6), 2397-403.

92. Griesser, E.; Wyatt, H.; Ten Have, S.; Stierstorfer, B.; Lenter, M.; Lamond, A. I., *Molecular & cellular proteomics : MCP* **2020,** *19* (5), 839-851.

93. Yang, Z.; Shen, X.; Chen, D.; Sun, L., *Journal of proteome research* **2019,** *18* (11), 4046-4054.

94. Kulak, N. A.; Geyer, P. E.; Mann, M., *Molecular & cellular proteomics : MCP* **2017,** *16* (4), 694-705.

95. Dou, M.; Tsai, C. F.; Piehowski, P. D.; Wang, Y.; Fillmore, T. L.; Zhao, R.; Moore, R. J.; Zhang, P.; Qian, W. J.; Smith, R. D.; Liu, T.; Kelly, R. T.; Shi, T.; Zhu, Y., *Analytical chemistry* **2019,** *91* (15), 9707-9715.

96. Zhou, M.; Uwugiaren, N.; Williams, S. M.; Moore, R. J.; Zhao, R.; Goodlett, D.; Dapic, I.; Pasa-Tolic, L.; Zhu, Y., *Analytical chemistry* **2020,** *92* (10), 7087-7095.

97. Tsai, C. F.; Zhao, R.; Williams, S. M.; Moore, R. J.; Schultz, K.; Chrisler, W. B.; Pasa-Tolic, L.; Rodland, K. D.; Smith, R. D.; Shi, T.; Zhu, Y.; Liu, T., *Molecular & cellular proteomics : MCP* **2020,** *19* (5), 828-838.

98. Dou, M.; Clair, G.; Tsai, C. F.; Xu, K.; Chrisler, W. B.; Sontag, R. L.; Zhao, R.; Moore, R. J.; Liu, T.; Pasa-Tolic, L.; Smith, R. D.; Shi, T.; Adkins, J. N.; Qian, W. J.; Kelly, R. T.; Ansong, C.; Zhu, Y., *Analytical chemistry* **2019,** *91* (20), 13119-13127.

99. Jandera, P., *Journal of chromatography. A* **2006,** *1126* (1-2), 195-218.

100. Schellinger, A. P.; Carr, P. W., *Journal of chromatography. A* **2006,** *1109* (2), 253-66.

101. Snyder, L. R.; Dolan, J. W. High-performance gradient elution : the practical application of the linear-solvent-strength model. **2006**.

102. Rafferty, J. L.; Siepmann, J. I.; Schure, M. R., *Journal of Chromatography A* **2011,** *1218* (16), 2203-2213.

103. van Deemter, J. J.; Zuiderweg, F. J.; Klinkenberg, A., *Chemical Engineering Science* **1956,** *5* (6), 271-289.

104. Billen, J.; Desmet, G., *Journal of Chromatography A* **2007,** *1168* (1), 73-99.

105. Shen, Y.; Tolic, N.; Zhao, R.; Pasa-Tolic, L.; Li, L.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D., *Analytical chemistry* **2001,** *73* (13), 3011-21.

106. Shen, Y.; Moore, R. J.; Zhao, R.; Blonder, J.; Auberry, D. L.; Masselon, C.; Pasa-Tolic, L.; Hixson, K. K.; Auberry, K. J.; Smith, R. D., *Analytical chemistry* **2003,** *75* (14), 3596-3605.

107. Ivanov, A. R.; Zang, L.; Karger, B. L., *Analytical chemistry* **2003,** *75* (20), 5306-16.

108. Luo, Q.; Shen, Y.; Hixson, K. K.; Zhao, R.; Yang, F.; Moore, R. J.; Mottaz, H. M.; Smith, R. D., *Analytical chemistry* **2005,** *77* (15), 5028-35.

109. Cech, N. B.; Enke, C. G., *Mass spectrometry reviews* **2001,** *20* (6), 362-87.

110. Page, J. S.; Kelly, R. T.; Tang, K.; Smith, R. D., *Journal of the American Society for Mass Spectrometry* **2007,** *18* (9), 1582-90.

111. Wilm, M.; Mann, M., *Analytical chemistry* **1996,** *68* (1), 1-8.

112. Marginean, I.; Tang, K.; Smith, R. D.; Kelly, R. T., *Journal of the American Society for Mass Spectrometry* **2014,** *25* (1), 30-6.

113. Meiring, H. D.; van der Heeft, E.; ten Hove, G. J.; de Jong, A. P. J. M., *Journal of Separation Science* **2002,** *25* (9), 557-568.

114. Yue, G.; Luo, Q.; Zhang, J.; Wu, S. L.; Karger, B. L., *Analytical chemistry* **2007,** *79* (3), 938-46.

115. Xiang, P.; Zhu, Y.; Yang, Y.; Zhao, Z.; Williams, S. M.; Moore, R. J.; Kelly, R. T.; Smith, R. D.; Liu, S., *Analytical chemistry* **2020,** *92* (7), 4711-4715.

116. Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J., *Analytical chemistry* **2012,** *84* (3), 1617-22.

117. Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J., *Analytical chemistry* **2013,** *85* (5), 2569-73.

118. Faserl, K.; Sarg, B.; Kremser, L.; Lindner, H., *Analytical chemistry* **2011,** *83* (19), 7297-305.

119. Wang, Y.; Fonslow, B. R.; Wong, C. C.; Nakorchevsky, A.; Yates, J. R., 3rd, *Analytical chemistry* **2012,** *84* (20), 8505-13.

120. Ramautar, R.; Heemskerk, A. A.; Hensbergen, P. J.; Deelder, A. M.; Busnel, J. M.; Mayboroda, O. A., *Journal of proteomics* **2012,** *75* (13), 3814-28.

121. Blass, J.; Köhler, O.; Fingerle, M.; Müller, C.; Ziegler, C., *physica status solidi (a)* **2013,** *210* (5), 988-993.

122. Gusev, I.; Horvath, C., *Journal of chromatography. A* **2002,** *948* (1-2), 203-23.

123. Fan, H. F.; Li, F.; Zare, R. N.; Lin, K. C., *Analytical chemistry* **2007,** *79* (10), 3654-61.

124. Zhu, G.; Sun, L.; Dovichi, N. J., *Talanta* **2016,** *146*, 839-43.

125. Haselberg, R.; de Jong, G. J.; Somsen, G. W., *Analytical chemistry* **2013,** *85* (4), 2289-96.

126. Busnel, J. M.; Schoenmaker, B.; Ramautar, R.; Carrasco-Pancorbo, A.; Ratnayake, C.; Feitelson, J. S.; Chapman, J. D.; Deelder, A. M.; Mayboroda, O. A., *Analytical chemistry* **2010,** *82* (22), 9476-83.

127. Sun, L.; Li, Y.; Champion, M. M.; Zhu, G.; Wojcik, R.; Dovichi, N. J., *The Analyst* **2013,** *138* (11), 3181-8.

128. Lubeckyj, R. A.; Basharat, A. R.; Shen, X.; Liu, X.; Sun, L., *Journal of the American Society for Mass Spectrometry* **2019,** *30* (8), 1435-1445.

129. Amenson-Lamar, E. A.; Sun, L.; Zhang, Z.; Bohn, P. W.; Dovichi, N. J., *Talanta* **2019,** *204*, 70-73.

130. 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., *Angewandte Chemie* **2014,** *53* (50), 13931-3.

131. Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J., *Analytical chemistry* **2014,** *86* (13), 6331-6.

132. Chen, D.; Shen, X.; Sun, L., *The Analyst* **2017,** *142* (12), 2118-2127.

133. Yang, Z.; Shen, X.; Chen, D.; Sun, L., *Analytical chemistry* **2018,** *90* (17), 10479-10486.

134. Zhang, Z.; Yan, X.; Sun, L.; Zhu, G.; Dovichi, N. J., *Analytical chemistry* **2015,** *87* (8), 4572-7.

135. Zhang, Z.; Sun, L.; Zhu, G.; Cox, O. F.; Huber, P. W.; Dovichi, N. J., *Analytical chemistry* **2016,** *88* (1), 877-82.

136. Maxwell, E. J.; Zhong, X.; Zhang, H.; van Zeijl, N.; Chen, D. D. Y., *Electrophoresis* **2010,** *31* (7), 1130-1137.

137. Wang, C.; Lee, C. S.; Smith, R. D.; Tang, K., *Analytical chemistry* **2013,** *85* (15), 7308-15.

138. Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J., *Rapid Communications in Mass Spectrometry* **2010,** *24* (17), 2554-2560.

139. Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J., *Angewandte Chemie International Edition* **2013,** *52* (51), 13661-13664.

140. Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J., *Journal of proteome research* **2015,** *14* (5), 2312-2321.

141. Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S., *Molecular & cellular proteomics : MCP* **2011,** *10* (9), M111 011015.

142. Baldeschwieler, J. D., *Science* **1968,** *159* (3812), 263-73.

143. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S., *Mass spectrometry reviews* **1998,** *17* (1), 1-35.

144. Guan, S.; Marshall, A. G., *International Journal of Mass Spectrometry and Ion Processes* **1996,** *157-158*, 5-37.

145. Bogdanov, B.; Smith, R. D., *Mass spectrometry reviews* **2005,** *24* (2), 168-200.

146. Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R., *Proteomics* **2002,** *2* (5), 513-23.

147. Hofstadler, S. A.; Severs, J. C.; Smith, R. D.; Swanek, F. D.; Ewing, A. G., *Rapid communications in mass spectrometry : RCM* **1996,** *10* (8), 919-22.

148. Quenzer, T. L.; Emmett, M. R.; Hendrickson, C. L.; Kelly, P. H.; Marshall, A. G., *Analytical chemistry* **2001,** *73* (8), 1721-5.

149. Shen, Y.; Tolic, N.; Masselon, C.; Pasa-Tolic, L.; Camp, D. G., 2nd; Lipton, M. S.; Anderson, G. A.; Smith, R. D., *Analytical and bioanalytical chemistry* **2004,** *378* (4), 1037-45.

150. Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R., *Journal of mass spectrometry : JMS* **2005,** *40* (4), 430-43.

151. Makarov, A.; Denisov, E., *Journal of the American Society for Mass Spectrometry* **2009,** *20* (8), 1486-95.

152. Rose, R. J.; Damoc, E.; Denisov, E.; Makarov, A.; Heck, A. J., *Nature methods* **2012,** *9* (11), 1084-6.

153. Eliuk, S.; Makarov, A., *Annual review of analytical chemistry* **2015,** *8*, 61-80.

154. Second, T. P.; Blethrow, J. D.; Schwartz, J. C.; Merrihew, G. E.; MacCoss, M. J.; Swaney, D. L.; Russell, J. D.; Coon, J. J.; Zabrouskov, V., *Analytical chemistry* **2009,** *81* (18), 7757-65.

155. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., *Nature methods* **2007,** *4* (9), 709-12.

156. Lange, O.; Damoc, E.; Wieghaus, A.; Makarov, A., *International Journal of Mass Spectrometry* **2014,** *369*, 16-22.

157. Makarov, A.; Denisov, E.; Lange, O., *Journal of the American Society for Mass Spectrometry* **2009,** *20* (8), 1391-6.

158. Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M., *Molecular & cellular proteomics : MCP* **2014,** *13* (12), 3698-708.

159. Kelstrup, C. D.; Bekker-Jensen, D. B.; Arrey, T. N.; Hogrebe, A.; Harder, A.; Olsen, J. V., *Journal of proteome research* **2018,** *17* (1), 727-738.

160. Cong, Y.; Motamedchaboki, K.; Misal, S. A.; Liang, Y.; Guise, A. J.; Truong, T.; Huguet, R.; Plowey, E. D.; Zhu, Y.; Lopez-Ferrer, D.; Kelly, R. T., *Chemical Science* **2021,** *12* (3), 1001-1006.

161. Kolakowski, B. M.; Mester, Z., *The Analyst* **2007,** *132* (9), 842-64.

162. Swearingen, K. E.; Moritz, R. L., *Expert review of proteomics* **2012,** *9* (5), 505-17.

163. Pfammatter, S.; Bonneil, E.; McManus, F. P.; Prasad, S.; Bailey, D. J.; Belford, M.; Dunyach, J. J.; Thibault, P., *Molecular & cellular proteomics : MCP* **2018,** *17* (10), 2051-2067.

164. Pfammatter, S.; Bonneil, E.; Thibault, P., *Journal of proteome research* **2016,** *15* (12), 4653-4665.

165. 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., *Analytical chemistry* **2018,** *90* (15), 9529-9537.

166. Clark, A. E.; Kaleta, E. J.; Arora, A.; Wolk, D. M., *Clinical microbiology reviews* **2013,** *26* (3), 547-603.

167. Morris, H. R.; Paxton, T.; Dell, A.; Langhorne, J.; Berg, M.; Bordoli, R. S.; Hoyes, J.; Bateman, R. H., *Rapid communications in mass spectrometry : RCM* **1996,** *10* (8), 889-96.

168. Cotter, R. J.; American Chemical, S.; Meeting; Pittsburgh Conference on Analytical, I. In *Time-of-flight mass spectrometry*, Washington, DC, **1994**; American Chemical Society: Washington, DC.

169. Beck, S.; Michalski, A.; Raether, O.; Lubeck, M.; Kaspar, S.; Goedecke, N.; Baessmann, C.; Hornburg, D.; Meier, F.; Paron, I.; Kulak, N. A.; Cox, J.; Mann, M., *Molecular & cellular proteomics : MCP* **2015,** *14* (7), 2014-29.

170. Eiceman, G. A.; Karpas, Z.; Hill, H. H., *Ion mobility spectrometry*. CRC press: Boca Raton, Fla. [etc.], **2016.**

171. Meier, F.; Beck, S.; Grassl, N.; Lubeck, M.; Park, M. A.; Raether, O.; Mann, M., *Journal of proteome research* **2015,** *14* (12), 5378-5387.

172. Meier, F.; Brunner, A. D.; Koch, S.; Koch, H.; Lubeck, M.; Krause, M.; Goedecke, N.; Decker, J.; Kosinski, T.; Park, M. A.; Bache, N.; Hoerning, O.; Cox, J.; Rather, O.; Mann, M., *Molecular & cellular proteomics : MCP* **2018,** *17* (12), 2534-2545.

173. Brunner, A.-D.; Thielert, M.; Vasilopoulou, C. G.; Ammar, C.; Coscia, F.; Mund, A.; Hoerning, O. B.; Bache, N.; Apalategui, A.; Lubeck, M.; Richter, S.; Fischer, D. S.; Raether, O.; Park, M. A.; Meier, F.; Theis, F. J.; Mann, M., *bioRxiv* **2021**, 2020.12.22.423933.

174. Lombard-Banek, C.; Moody, S. A.; Manzini, M. C.; Nemes, P., *Analytical chemistry* **2019,** *91* (7), 4797-4805.

175. Zhu, Y.; Scheibinger, M.; Ellwanger, D. C.; Krey, J. F.; Choi, D.; Kelly, R. T.; Heller, S.; Barr-Gillespie, P. G., *eLife* **2019,** *8*.

176. Specht, H.; Emmott, E.; Petelski, A. A.; Huffman, R. G.; Perlman, D. H.; Serra, M.; Kharchenko, P.; Koller, A.; Slavov, N., *Genome biology* **2021,** *22* (1), 50.

177. Hofstadler, S. A.; Swanek, F. D.; Gale, D. C.; Ewing, A. G.; Smith, R. D., *Analytical chemistry* **1995,** *67* (8), 1477-80.

[2]CHPATER 2. 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

2.1 Introduction

Capillary zone electrophoresis-tandem mass spectrometry (CZE-MS/MS) has a long history in bottom-up proteomics. Yates group reported bottom-up proteomics of yeast ribosome using CZE-MS/MS in 1999.[1] Since then, there have been many reports on exploring CZE-MS/MS for bottom-up proteomics of samples with varying complexity.[2-14] CZE-MS/MS has some unique features compared to widely used reversed-phase liquid chromatography (RPLC)-MS/MS for bottom-up proteomics. First, it outperforms RPLC-MS/MS for analyses of low nanograms of complex proteome digests in terms of the number of protein identifications (IDs).[3,6,9,11] This feature makes CZE-MS/MS very useful for mass-limited sample analyses, *e.g.,* characterization of single cells.[11] Second, migration time of peptides in CZE can be predicted easily with high accuracy while prediction of peptides' retention time in RPLC is much more difficult due to various interactions between peptides and beads.[15] This valuable feature of CZE provides the proteomics community with a new approach to validate the confidence of peptide ID from a database search. Third, CZE can reach much better separation of phosphorylated and unphosphorylated forms of peptides than RPLC,[9] which makes CZE-MS/MS very useful

[2] Part of this chapter was adapted with permission from: Yang, Z.; Shen, X.; Chen, D.; Sun, L., Analytical chemistry 2018, 90 (17), 10479-10486.

for phosphoproteomics. Fourth, CZE-MS/MS has no bias against the ID of very hydrophilic peptides that usually have no or very weak retention on RPLC columns.[3,8]

Recently, CZE-MS/MS has achieved great progress in deep bottom-up proteomics due to improvement in CE-MS interface,[16-19] mass spectrometer,[20,21] and online sample stacking method.[6, 22-26] Yan *et al.* identified over 4000 proteins from *Xenopus* embryos via coupling RPLC prefractionation to CZE-MS/MS.[25] Very recently, our group reported over 8000 protein IDs from a mouse brain proteome digest using two-dimensional (2D) LC-CZE-MS/MS.[27] LC-CZE-MS/MS is becoming a useful tool for deep bottom-up proteomics. LC-CZE-MS/MS and the state-of-the-art 2D-LC-MS/MS are complementary in peptide and protein IDs from complex proteomes and combining those two techniques can boost the proteome coverage from bottom-up proteomics.[27]

Hundreds of micrograms of proteome digests were typically required for the LC-CZE-MS/MS-based deep bottom-up proteomics studies.[5,25-27] There are at least two reasons for the requirement of large amounts of proteome digests. First, typically LC fractionation employs a LC column with a large inner diameter (2.1 or 4.6 mm). The total surface area of beads packed in the column and the optimum flow rate of mobile phase for separation are high, leading to unavoidable sample loss on the beads and on the inner wall of Eppendorf tubes used for fraction collection.[28] Second, the sample loading volume of CZE is low, typically less than 100 nL.[5,25,26] A reasonable amount of peptide material is required for identification of low abundant proteins in a complex proteome sample with MS/MS. Therefore, a relatively high peptide concentration in the sample is essential for CZE-MS/MS in order to reach large-scale proteomics. The significant sample loss during LC fractionation and the requirement of high peptide concentration in the sample for CZE-

MS/MS lead to the need of a large amount of starting peptide material for deep bottom-up proteomics using the LC-CZE-MS/MS. The requirement of large amount of peptides impedes the application of CZE-MS/MS for deep bottom-up proteomics of mass-limited samples.

Microscale RPLC (μRPLC)-based peptide fractionation and CZE-MS/MS with a much higher sample loading volume are vital for dramatically reducing the required amount of initial peptide material. Several groups have tested different μRPLC methods for fractionation of complex proteome digests prior to CZE-MS/MS or nanoRPLC-MS/MS.[6,28-31] RP-based solid phase microextraction (RP-SPME) and strong cation exchange monolith based SPME have been online coupled with CZE-MS/MS for highly sensitive bottom-up proteomics using nanograms of proteome digests.[6,29,30] The SPME-CZE-MS/MS platforms will be very useful for large-scale proteomics of mass-limited proteome samples, *e.g.,* single cells. C18 ZipTip has low-μg loading capacity and it is a simple and efficient μRPLC method for peptide fractionation. Choi *et al.* have employed the C18 ZipTip-CZE-MS/MS recently to enhance the protein ID from mammalian neuron proteome digests that correspond to the protein content of a few mammalian neurons in mass.[31] SPME and C18 ZipTip can be considered as low-resolution μRPLC. Kulak *et al.* recently showed that coupling high-resolution and high-pH μRPLC fractionation to low-pH nanoflow RPLC-MS/MS enabled deep bottom-up proteomics of human cell lines using low-μg of proteome digests.[28] The data in that work clearly demonstrated that high-resolution μRPLC could reach highly efficient peptide fractionation with high peptide recovery. In order to reach a high sample loading volume in CZE, efficient online sample concentration is essential. The dynamic pH junction method[32,33] has been evaluated for

the online concentration of peptides and intact proteins in CZE.[14,24,34,35] Our group demonstrated that the dynamic pH junction based CZE-MS/MS could approach a microliter scale sample loading volume and a 2-hour separation window simultaneously.[14,34] We speculate that coupling the low-resolution C18 ZipTip or the high-resolution μRPLC prefractionation to the dynamic pH junction based CZE-MS/MS will enable deep bottom-up proteomics starting with low-μg complex proteome digests.

In this work, we first performed a calibration curve experiment using the dynamic pH junction based CZE-MS/MS for bottom-up proteomics of the MCF7 proteome with 0.1-100 ng of protein digests as the starting material. Then we investigated C18 ZipTip-dynamic pH junction based CZE-MS/MS for deep bottom-up proteomics of the MCF7 proteome starting with 5-μg protein digest. Finally, we coupled high-resolution nanoflow RPLC (nanoRPLC) to the dynamic pH junction based CZE-MS/MS to deepen the coverage of the MCF7 proteome with 5-μg protein digest.

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, methanol, acetonitrile (ACN), HPLC grade acetic acid (AA), formic acid (FA), and hydrofluoric acid (HF) were purchased from Fisher Scientific (Pittsburgh, PA). Urea was purchased from Alfa Aesar (Haverhill, MA). Acrylamide was ordered from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ).

2.2.2 Preparation of the MCF7 breast cancer cell proteome digest

The human breast cancer cells (MCF-7) were kindly provided by Dr. Xuefei Huang's group at the Department of Chemistry, Michigan State University. MCF7 cells were cultured at 37°C under a 5% CO_2 in ATCC-formulated Eagle's Minimum Essential Medium, supplemented with 0.01 mg/ml human recombinant insulin and 10% fetal bovine serum. The cells were washed with PBS and lysed in a lysis buffer containing 8 M urea, 100 mM ammonia bicarbonate (NH_4HCO_3, pH 8.0) and protease inhibitors, with the assistance of ultrasonication (Branson Sonifier 250, VWR Scientific, Batavia, IL) on ice for 10 minutes. After centrifugation at 10000 g for 10 min, the supernatant was collected and the protein concentration was measured by the BCA assay.

250 µg of MCF-7 proteins were precipitated with cold acetone and stored at -20 ^oC overnight. After centrifugation at 14000 g for 10 min, the protein pellet was washed with cold acetone once and was air-dried for a couple of minutes at room temperature. The protein pellet was then redissolved in 100 µL of 8 M urea containing 100 mM NH_4HCO_3 (pH 8.0). The proteins were reduced by dithiothreitol (DTT) at 37^oC for 30 min and alkylated by iodoacetamide (IAA) at room temperature for 20 min in the dark. DTT was added to quench extra IAA. The protein sample was then diluted to 500 µL with 100 mM NH_4HCO_3 (pH 8.0), followed by tryptic digestion at 37°C for overnight with 8 µg of trypsin (Bovine pancreas TPCK-treated). The digestion process was terminated by adding 2 µL of formic acid. Peptides were then desalted using Sep-Pak C18 Cartridge (Waters) and lyophilized with a vacuum concentrator (Thermo Fisher Scientific). The peptide sample was stored at -80°C before use.

2.2.3 Calibration curve experiment

We prepared a series of MCF7 peptide samples in 20 mM ammonium bicarbonate (NH_4HCO_3, pH 8.0) with various concentrations: 50 ng/μL, 5 ng/μL, 0.5 ng/μL, and 0.05 ng/μL. 2 μL of each sample was put into a sample vial and a 500-nL aliquot of each sample was injected for CZE-MS/MS. The mass of peptides in the sample vial was 100 ng, 10 ng, 1 ng, and 0.1 ng corresponding to different peptide concentrations from 50 ng/μL to 0.05 ng/μL. Each sample was analyzed by one CZE-MS/MS run. We repeated the experiment three times to obtain triplicate analyses of each sample.

2.2.4 C18 ZipTip fractionation

A C18 ZipTip (Millipore, ZTC18S096) was first conditioned with 100 μL of 80% (v/v) acetonitrile (ACN) containing 20 mM NH_4HCO_3 (pH 8.0) and equilibrated with 100 μL of 3% (v/v) ACN containing 20 mM NH_4HCO_3 (pH 8.0). 5-μg MCF7 proteome digest dissolved in 20 mM NH_4HCO_3 (pH 8.0) was loaded onto the ZipTip by pipetting up and down for 20 times. The flow-through from the loading was kept for further analysis. The peptides retained on the ZipTip were separated by step-wise elution using 6%, 15%, 20%, 25%, 40%, and 80% (v/v) ACN containing 20 mM NH_4HCO_3 (pH 8.0). Each elution was processed by centrifugation at 1000 rpm for 2 min and 4000 rpm for 30 seconds using 20-μL elution buffer. The flow-through from loading and the eluate from 80% (v/v) ACN were combined. Each eluate was lyophilized and redissolved in 5 μL of 20 mM NH_4HCO_3 (pH 8.0) for CZE-MS/MS analysis.

2.2.5 NanoRPLC fractionation

An EASY nanoLC-1200 (Thermo Fisher Scientific) was used for the fractionation. An analytical column (75 µm i.d. x 50 cm, C18, 2 µm, 100 Å, Thermo Fisher Scientific) was employed for RPLC. Buffer A containing 0.1% (v/v) formic acid (FA) and buffer B containing 80% (v/v) ACN and 0.1% (v/v) FA were used to generate gradient separation. 5-µg MCF7 proteome digest was loaded onto the RPLC column with buffer A at 800-bar pressure. Then the peptides retained on the column were separated by a linear gradient. The flow rate was 200 nL/min.

For the first RPLC fraction experiment, the gradient used for peptide separation was as follows: from 8 to 30% (v/v) B in 50 min, from 30% to 50% (v/v) B in 25 min, from 50% to 100% (v/v) B in 5 min and maintain at 100% (v/v) B for 10 min. The fraction collection started from the sample loading. The flow-through during sample loading and the peptides eluted during the first 14 min of the gradient were collected in one Eppendorf tube as the first fraction. After that, we collected one fraction every 7 min. The last fraction was collected from 70 min to 90 min. In total, 10 fractions were collected. The eluate for each fraction was directly transferred to the bottom of one 0.6-mL Eppendorf tube during fraction collection. After lyophilization, each fraction was redissolved in 3 µL of 20 mM NH_4HCO_3 (pH 8.0) by pipetting up and down for CZE-MS/MS.

For the second RPLC fractionation experiment, the gradient for RPLC separation was as follows: from 5 to 20% (v/v) B in 100 min, from 20% to 40% (v/v) B in 50 min, from 40% to 100% (v/v) B in 15 min and maintain at 100% (v/v) B for 15 min. The fraction collection started from the sample loading. The flow-through during sample

loading and the peptides eluted during the first 15 min of the gradient were collected as the first fraction. Then, each fraction was collected every 7 min. The last fraction was collected starting from 141 min until the end of the gradient. We collected 20 fractions in total into 0.6-mL Eppendorf tubes. All fractions were lyophilized and each fraction was redissolved in 1.5 µL of 20 mM NH_4HCO_3 (pH 8.0) for CZE-MS/MS.

2.2.6 CZE-MS/MS analysis

A one-meter linear polyacrylamide (LPA) coated fused silica capillary (50 µm i.d./360 µm o.d) was used for CZE. The LPA coating was prepared based on references 14 and 36. One end of the capillary was etched with hydrofluoric acid (HF) to reduce its outer diameter to ~70 µm. The length of the etched part was about 5 mm. The LPA coated capillary was stored at room temperature before use.

The commercialized electro-kinetically pumped sheath flow CE-MS interface (CMP scientific, Brooklyn, NY) was used to couple CZE to MS.[17-19] The automated CZE operations were carried out with an ECE-001 autosampler (CMP scientific) or a CESI 8000 plus CE system (Sciex). The background electrolyte (BGE) for CZE was 5% (v/v) acetic acid (pH 2.4). The sheath buffer was 10% (v/v) methanol and 0.2% (v/v) FA. The glass emitter for electrospray was pulled from a borosilicate glass capillary (0.75 mm i.d., 1 mm o.d.) by a Sutter P-1000 flaming/brown micropipette puller. The orifice of the electrospray emitter was 20-40 µm. The distance between the etched end of the separation capillary and the emitter orifice was less than 300 µm. The distance between the emitter orifice and the MS entrance was 2 mm.

Sample injection was performed by applying 5-psi pressure for 90 seconds. About 500 nL of the sample was loaded into the capillary for CZE separation. After

sample loading, the injection end of the capillary was moved into the BGE vial. 30 kV voltage was applied at the injection end for CZE separation and 2 kV was applied in the sheath buffer vial for electrospray. The CZE separation was finished in 90 min or 100 min. After the CZE separation, 15-psi pressure and 30 kV were applied at the injection end to flush the capillary for 5 min. In order to inject a 500-nL sample plug for CZE-MS/MS from only 5 µL or lower volume of sample in a sample vial, we pushed the injection end of the capillary to the bottom of the sample vial for injection. For the experiments on the CESI 8000 plus CE system, we used the nanoVials (Part number 5043467, SCIEX) for sample injection.

A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used in all the experiments. For all experiments, the resolution for full MS was 60000, the AGC was 1E6, the maximum injection time was 50 ms and the scan range was 300-1800 *m/z*. A Top10 data-dependent acquisition (DDA) method was applied. Only ions with a charge of 2 or higher were isolated in the quadrupole and fragmented in the high energy collision dissociation (HCD) cell with normalized collision energy (NCE) as 28%. Other parameters were varied for different experiments.

For the calibration-curve experiment, we changed some parameters for different amounts of MCF7 peptides. For CZE-MS/MS analysis of 0.1-ng to 10-ng MCF7 peptides, the resolution for MS/MS was 60000, the AGC target for MS/MS was 1E5, and the maximum injection time for MS/MS was 250 ms. The isolation window was 4 *m/z* and the intensity threshold for fragmentation was 2.2E4. Dynamic exclusion was 10 s. For CZE-MS/MS analysis of 100-ng sample, all the parameters were the same as that used for 0.1-ng to 10-ng samples except the MS/MS maximum injection time as

110 ms, the isolation window as 2 *m/z*, the intensity threshold for fragmentation as 5E4, and the dynamic exclusion as 30 s. The parameter settings were referred to reference 37. It has been demonstrated that a high injection time (≥250 ms) and a wide isolation window (4 *m/z*) can benefit the analysis of mass-limited samples.[37] For the 100-ng sample, a 110-ms maximum injection time and a 2-*m/z* isolation window were used to acquire more MS/MS spectra and reduce the co-isolation of peptides for fragmentation.

For CZE-MS/MS analysis of the fractions from C18 ZipTip and nanoflow RPLC fractionation, the resolution for MS/MS was 30000, the maximum injection time for MS/MS was 50 ms, the AGC target for MS/MS was 1E5, the isolation window was 2.0 m/z, the intensity threshold for fragmentation was 5E4, and the dynamic exclusion was 30 s.

2.2.7 Nano 2D-RPLC-MS/MS analysis

A C18 RPLC column (75 µm i.d. x 50 cm, C18, 2 µm, 100 Å, Thermo Fisher Scientific) connected to an EASY nanoLC-1200 system (Thermo Fisher Scientific) was used for the first dimensional high-pH RPLC separation. Buffer A containing 5 mM ammonium bicarbonate (pH 9.0) and buffer B containing 80% (v/v) ACN and 5 mM ammonium bicarbonate (pH 9.0) were used to generate gradient separation. 5 µg of the MCF7 protein digest was loaded onto the RPLC column with buffer A at 800-bar pressure. Then the peptides retained on the column were separated by a linear gradient. The flow rate was 200 nL/min. The gradient for RPLC separation was as follows: from 5 to 20% (v/v) B in 100 min, from 20% to 40% (v/v) B in 50 min, from 40% to 100% (v/v) B in 15 min and maintain at 100% (v/v) B for 15 min. The fraction collection started from the sample loading. The flow-through during sample loading and

the peptides eluted during the first 15 min of the gradient were collected as the first fraction. Then, each fraction was collected every 5 min. The last fraction was collected starting from 155 min until the end of the gradient. We collected 30 fractions in total and combined fractions into 15 fractions. All fractions were lyophilized and each fraction was redissolved in 5 µL of 0.1 % (v/v) FA for low pH nanoRPLC-MS/MS.

80% of the peptides in each fraction were loaded on the analytical column (75 µm i.d. x 50 cm, C18, 2 µm, 100 Å, Thermo Fisher Scientific) for low pH RPLC separation. An EASY nanoLC-1200 (Thermo Fisher Scientific) was used. Buffer A containing 0.1% (v/v) FA and buffer B containing 80% (v/v) ACN and 0.1% (v/v) FA were used to generate gradient separation. Sample was loaded onto the RPLC column with buffer A at 800-bar pressure. Then the peptides retained on the column were separated by a linear gradient. The flow rate was 200 nL/min. The gradient for RPLC separation was as follows: from 5 to 20% (v/v) B in 60 min, from 20% to 40% (v/v) B in 40 min, from 40% to 100% (v/v) B in 10 min and maintain at 100% (v/v) B for 10 min.

The parameter settings of each nanoLC-MS/MS analysis were all consistent with CZE-MS/MS analysis except 120 min was set for data acquisition.

2.2.8 Data Analysis

Proteome Discoverer 2.2 (Thermo Fisher Scientific) was used for the database search. SEQUEST HT searching engine was used. Precursor mass tolerance was 20 ppm while fragment mass tolerance was 0.05 Da. UniProt human proteome database (UP000005640) was used for the database searching. Trypsin was set as the enzyme with two maximum missed cleavages. The false discovery rate (FDR) was evaluated by the target and decoy approach.[38,39] Methionine oxidation, Protein N-terminal acetylation,

and Asparagine or Glutamine deamination were set as variable modifications. Carbamidomethylation of cysteine was set as the fixed modification. The peptides were filtered with confidence as high, corresponding to a 1% peptide-level FDR. Protein grouping was enabled, and the strict parsimony principle was applied. All the numbers of protein and peptide IDs reported in this work were from Proteome Discoverer 2.2.

We also performed database search using MaxQuant software[40] (version 1.5.5.1) to obtain label-free quantification (LFQ) information. All the parameters were kept as default settings. The LFQ and match-between-runs function were enabled.[41] The UniProt human proteome database (UP000005640) was used for the database searching. Perseus software (version 1.6.0.7) was used for further analysis of the MaxQuant results.[42] The MaxQuant data was only used for evaluation of the quantitative reproducibility of the C18-ZipTip-CZE-MS/MS based on the LFQ information.

Grand average of hydropathy (GRAVY) values of peptides were calculated through GRAVY Calculator (http://www.gravy-calculator.de/). Positive GRAVY values indicate hydrophobic and negative values show hydrophilic.

2.3 Results and Discussion

2.3.1 Calibration curve data

CZE-MS typically can approach excellent mass detection limit and relatively unsatisfied concentration detection limit due to the low sample loading volume of CZE. Hundreds to thousands of protein IDs have been reported using CZE-MS/MS for analysis of low ng even pg of complex proteome digests.[11,18,31,43] However, due to the low nL loading volume of CZE, the concentration of proteome digests in sample vials for injection need to be high for identification of low abundant proteins in complex

proteomes, leading to the requirement of at least 1-µg protein digest for the CZE-MS/MS studies. Recently, we boosted the sample loading volume of CZE to microliter scale based on the dynamic pH junction online sample stacking method. [14,34] The large sample loading volume offers us the opportunity to fully use the available peptide material in the sample vial, dramatically reducing the required mass of peptides for CZE-MS/MS. Here we investigated the dynamic pH junction based CZE-MS/MS for analysis of a MCF7 proteome digest using 0.1-100 ng of peptides as the starting material. Each CZE-MS/MS run took 90 min.

Two-µL aliquots of MCF7 peptide samples with a various concentration in 20 mM NH_4HCO_3 (pH 8.0) were put into sample vials for CZE-MS/MS. The mass of peptides in the sample vials ranged from 0.1 ng to 100 ng. Our CZE-MS/MS system enabled a 500-nL sample injection from the 2-µL available sample in the vial, using 25% of the available sample for analysis. A diagram of the dynamic pH junction based CZE-MS/MS is shown in **Figure 2.1A**. **Figure 2.1B** shows the number of peptide and protein IDs as a function of the amount of available peptide material and injected peptide material. On average 1200 proteins and 5300 peptides were identified by the CZE-MS/MS when only 100-ng peptide was available in the vial. The data clearly indicate that our dynamic pH junction based CZE-MS/MS is able to reduce the required mass of complex proteome digests by one order of magnitude compared with the literature data (100 ng *vs.* 1 µg).[11] The CZE-MS/MS was able to identify on average 100 proteins and 15 proteins from the MCF7 proteome starting with only 1-ng and 0.1-ng peptide, which correspond to the protein content of only ten and one MCF7 cells in mass.[44] The data highlight the potential of the CZE-MS/MS system for single cell proteomics.

As shown in **Figure 2.1B**, the number of protein and peptide IDs steadily increase as the amount of available peptide material grows. The CZE-MS/MS produced reproducible protein and peptide IDs from triplicate analyses with the relative standard deviations (RSDs) lower than 10% when 10 ng and 100 ng of peptides were available. We extracted the peaks of 12 peptides across all the raw files of 0.1-100 ng samples. The 12 peptides were randomly chosen from that identified from the 0.1-ng sample. We plotted the peptide intensity as a function of the available peptide amount, **Figure 2.1C.** The peptide intensity grows linearly as the peptide amount increases. The data suggest high confidence of those identified peptides from the 0.1-ng sample.

The calibration curve data indicate high sensitivity of our dynamic pH junction based CZE-MS/MS system and inspired us to think about deep proteomics of the MCF7 cells with only low-µg peptides using µRPLC fractionation-dynamic pH junction CZE-MS/MS.

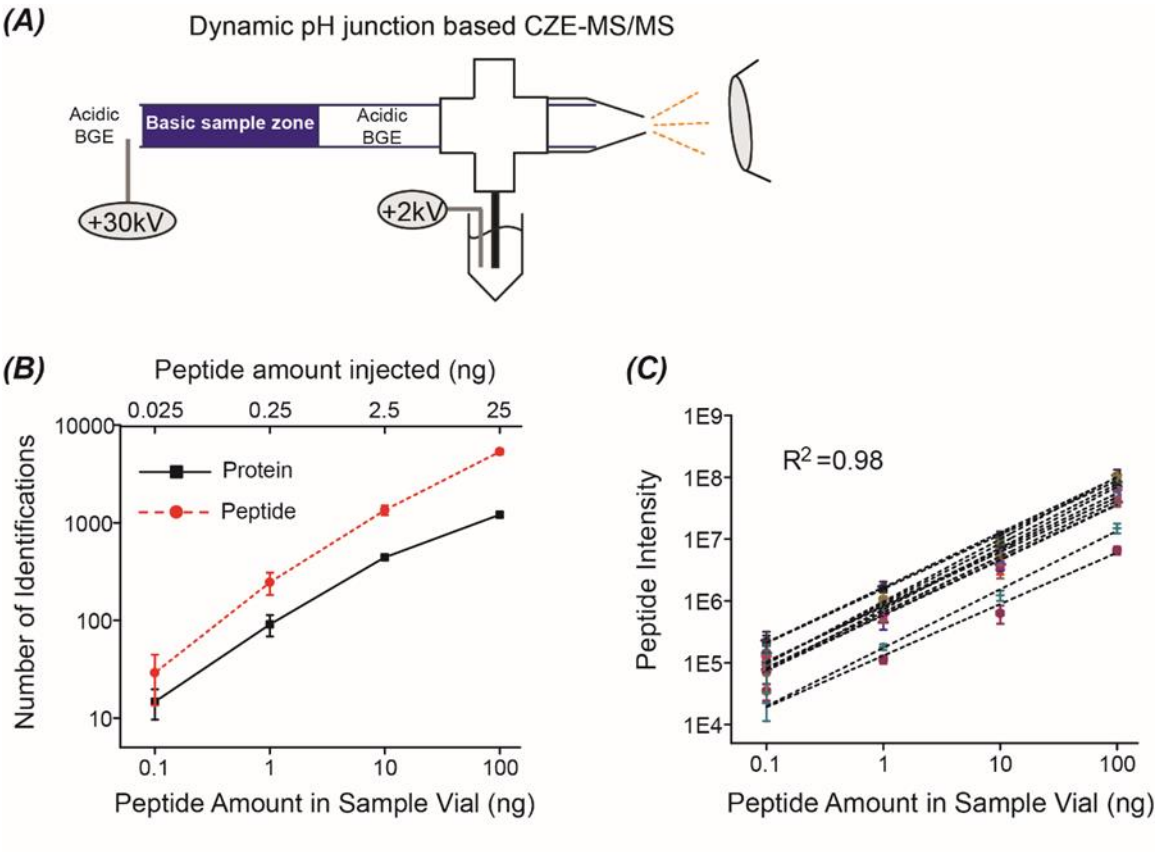

Figure 2. 1 The calibration curve data. (A) Diagram of dynamic pH junction based CZE-MS/MS. The sample was dissolved in 20 mM NH_4HCO_3 (pH 8.0) and the BGE was 5% (v/v) acetic acid (pH 2.4). Roughly 25% of the separation capillary was filled with the sample for CZE-MS/MS. (B) The number of protein and peptide identifications as a function of the peptide amount in the sample vial (log-log plot). The peptide amount injected for CZE-MS/MS analysis was labeled on the top. The error bars represent the standard deviations of the number of identifications from CZE-MS/MS in triplicate. (C) The peptide intensity as a function of the peptide amount in the sample vial (log-log plot). Twelve peptides were randomly chosen from that identified with the 0.1-ng sample for this figure. The peptide intensity was obtained based on the extracted peptide peaks

with 20-ppm mass tolerance and Gaussian smoothing (5 points). The error bars represent the standard deviations of the peptide intensity from CZE-MS/MS in triplicate.

2.3.2 C18 ZipTip fractionation-CZE-MS/MS

We investigated the C18 ZipTip fractionation-dynamic pH junction CZE-MS/MS for deep bottom-up proteomics of the MCF7 cells starting with only 5-µg proteome digest. The C18 ZipTip fractionated the 5-µg peptides into 6 fractions based on their hydrophobicity. Each eluate was lyophilized and redissolved in 5 µL of 20 mM NH_4HCO_3 (pH 8.0) for CZE-MS/MS analysis. Each fraction was analyzed by one CZE-MS/MS run in 105 min with a 500-nL sample loading volume. One C18 ZipTip-CZE-MS/MS experiment took 630 min of instrument time. The whole experiment was performed three times.

The platform produced 4453±304 (n=3) protein IDs, 26928±2489 (n=3) peptide IDs, 93354±11209 (n=3) peptide-spectrum matches (PSMs) and 229148±10524 (n=3) MS/MS spectra, **Figure 2.2A**. The data indicate the good reproducibility of the platform. More importantly, for the first time, CZE-MS/MS was able to approach deep bottom-up proteomics of a human cell line proteome reproducibly using only low-µg of starting peptide material. The C18 ZipTip fractionation is simple and straightforward. It can be completed within 20 min. We believe the C18-ZipTip fractionation-dynamic pH junction CZE-MS/MS can be a useful tool for deep bottom-up proteomics.

The C18 ZipTip-CZE-MS/MS was performed in triplicate and the protein-level overlap between replicates is reasonable (>70%), **Figure 2.2B**. The platform also shows great quantitative reproducibility based on the LFQ intensity of quantified proteins ($R^2 \geq 0.97$), **Figure 2.2C**. We further investigated the performance of C18 ZipTip

for peptide fractionation, **Figure 2.2D**. We used the first replicate data and obtained the peptide-level overlap between adjacent C18 ZipTip fractions. CZE-MS/MS generated 4000-9000 peptide IDs from one C18 ZipTip fraction. The adjacent fractions share a significant portion of the identified peptides, and the number of shared peptides ranges from 2000 to 4000. The obvious peptide-level overlap between adjacent C18 ZipTip fractions is most likely due to the low resolution of C18 ZipTip for peptide separation. We speculated that coupling high-resolution µRPLC fractionation to the dynamic pH junction CZE-MS/MS could further deepen the coverage of the MCF7 proteome.

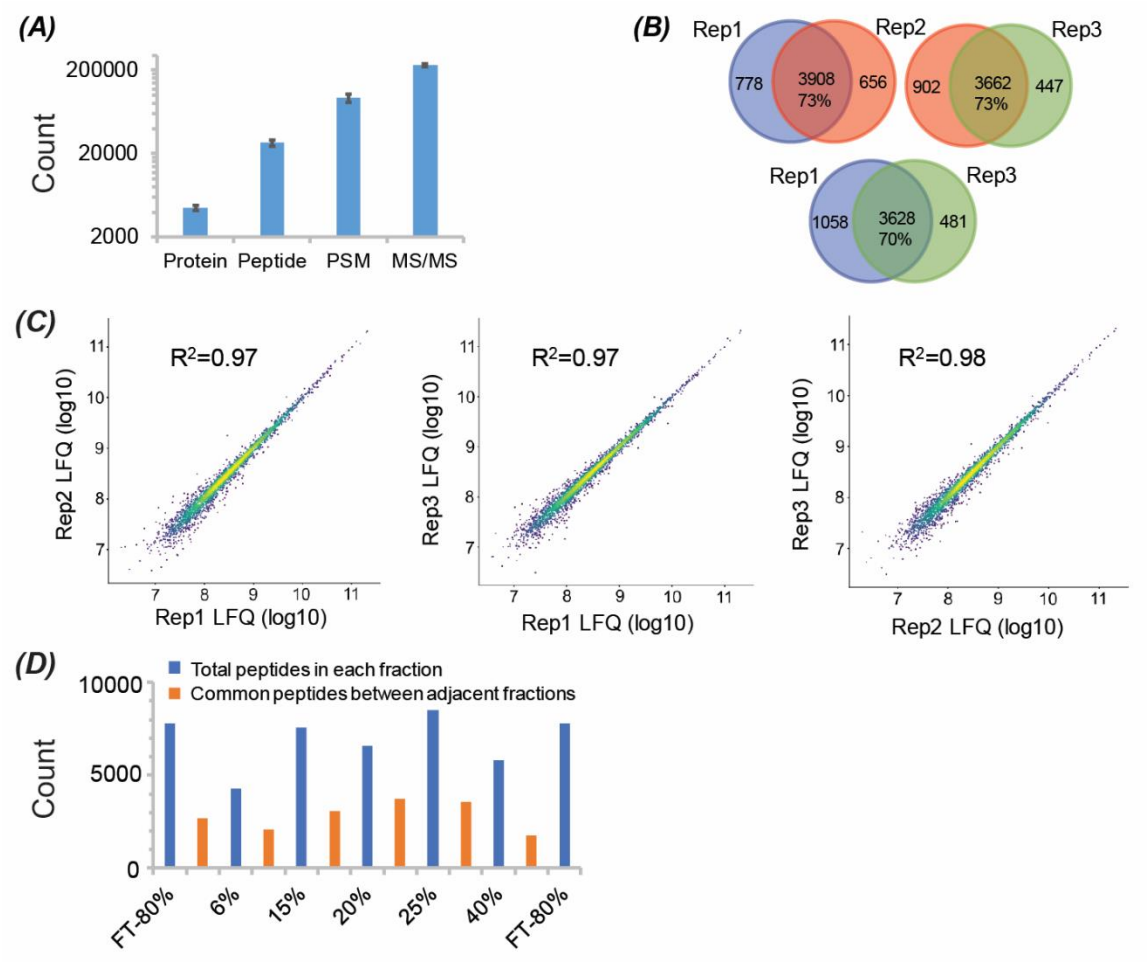

Figure 2. 2 The C18 ZipTip-CZE-MS/MS data. (A) Protein IDs, Peptide IDs, peptide-spectrum matches (PSMs), and the acquired MS/MS spectra. The error bars represent

the standard deviations of the IDs from the platform in triplicate. (B) The protein-level overlap between the C18 ZipTip-CZE-MS/MS replicates. (C) The correlations of protein LFQ intensity between the C18 ZipTip-CZE-MS/MS replicates. (D) The number of peptide IDs in each C18 ZipTip fraction and the number of common peptides between adjacent fractions. “FT” represents the flow through during the sample loading. The 6%, 15%, 20%, 25%, 40%, and 80% represent the concentration of ACN for peptide elution from the ZipTip.

2.3.3 NanoRPLC fractionation-CZE-MS/MS

We further coupled high-resolution nanoRPLC fractionation to the dynamic pH junction based CZE-MS/MS for deep bottom-up proteomics of the MCF7 cell proteome starting with only 5-µg proteome digest. One RPLC column (75-µm i.d., 50-cm length) packed with 2-µm C18 beads was used for the high-resolution fractionation. The flow rate was 200 nL/min.

Two experiments were performed. In the first experiment, we fractionated the 5-µg proteome digest with the RPLC column into 10 fractions, lyophilized the fractions, and redissolved each fraction in 3 µL of 20 mM NH_4HCO_3 (pH 8.0) for CZE-MS/MS with a 500-nL sample loading volume. The CZE-MS/MS identified 5769 proteins and 35575 peptides from the 10 RPLC fractions. The experiment took 1100-min instrument time. In the second experiment, we fractionated another 5-µg MCF7 proteome digest sample with the same RPLC column and obtained 20 fractions. Because the average amount of peptides in each fraction was lower than that in the first experiment due to more fractions (20 *vs.* 10), we dissolved the peptides in each fraction in only 1.5 µL of 20 mM NH_4HCO_3 (pH 8.0) for CZE-MS/MS. Over 30% of the peptide material in each fraction

was used for the analysis (500 nL *vs.* 1.5 µL) in order to identify lower abundant proteins. CZE-MS/MS analyses of the 20 RPLC fractions in 34 hours generated 7512 protein IDs, 59403 peptide IDs, 152086 PSMs, and 543622 MS/MS spectra. The number of peptide and protein IDs was improved by 67% and 30%, respectively, compared to that from the first experiment. Compared with recent deep bottom-up proteomics work using LC-CZE-MS/MS,[5,25-27] this work reduced the required amount of complex proteome digests by two orders of magnitude (5 µg *vs.* ≥500 µg). We noted that different mass spectrometers used in those studies could also affect the overall sensitivity of LC-CZE-MS/MS. We also noted that CZE-MS/MS analyses of only 8 RPLC fractions in less than 14 hours could yield 6000 protein IDs, **Figure 2.3**. The nanoRPLC-CZE-MS/MS platform will become a powerful tool for highly sensitive and deep proteomics. The mass spectrometry proteomics raw files have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD009991.[45,46]

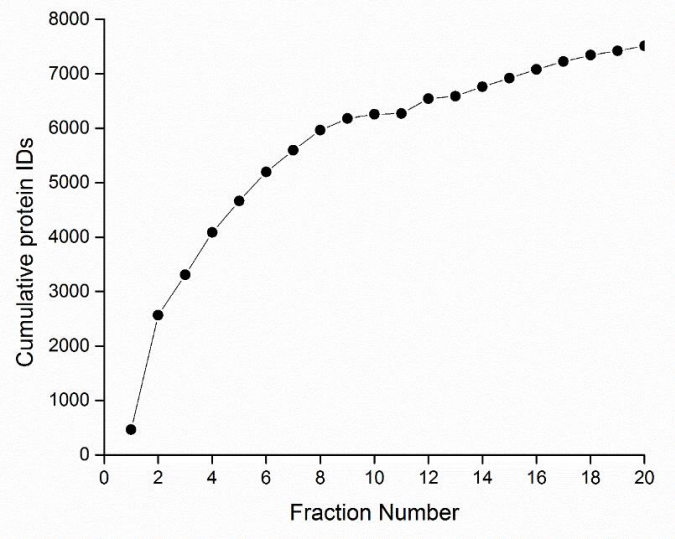

Figure 2. 3 Cumulative protein IDs vs. number of RPLC fractions. The data from the nanoRPLC-CZE-MS/MS experiment (20 RPLC fractions) were used for the figure.

We attribute the drastically high sensitivity of our nanoflow RPLC-CZE-MS/MS system to three major reasons. First, the nanoRPLC system significantly reduced the sample loss due to the much lower total surface area of beads packed in the column and much lower mobile phase flow rate compared with the RPLC systems used in the literature (75-μm-i.d. column *vs.* ≥2.1-mm-i.d. column; 200 nL/min *vs.* ≥200 μL/min).[5, 25-27] The data reported recently by the Mann's group provided strong evidence for this point.[28] Second, the nanoRPLC produced high-resolution separation of peptides, leading to negligible peptide-level overlap between adjacent RPLC fractions, **Figure 2.4A**. The high-resolution RPLC fractionation reduced the sample complexity and ion suppression during the electrospray ionization. Third, the dynamic pH junction based CZE-MS/MS was able to use over 30% of the peptide material in each fraction for analysis. In previous LC-CZE-MS/MS studies,[5,25,26] the CZE-MS/MS system was only able to inject less than 1% of the peptide material in each LC fraction for analysis.

We noted that the nanoRPLC and dynamic pH junction based CZE yielded highly orthogonal separation of peptides. RPLC separates peptides based on their hydrophobicity. As shown in **Figure 2.4B**, the peptides in later RPLC fractions tend to be more hydrophobic, which agrees well with the separation mechanism of RPLC. CZE separates peptides based on their size-to-charge ratios.[15] The retention time of peptides in RPLC and their migration time in CZE have no significant correlation, demonstrating that RPLC and CZE are highly orthogonal for peptide separation, **Figure 2.4C**. The peptides in every 7-min eluate from the nanoRPLC were further separated by CZE in a 60-min window.

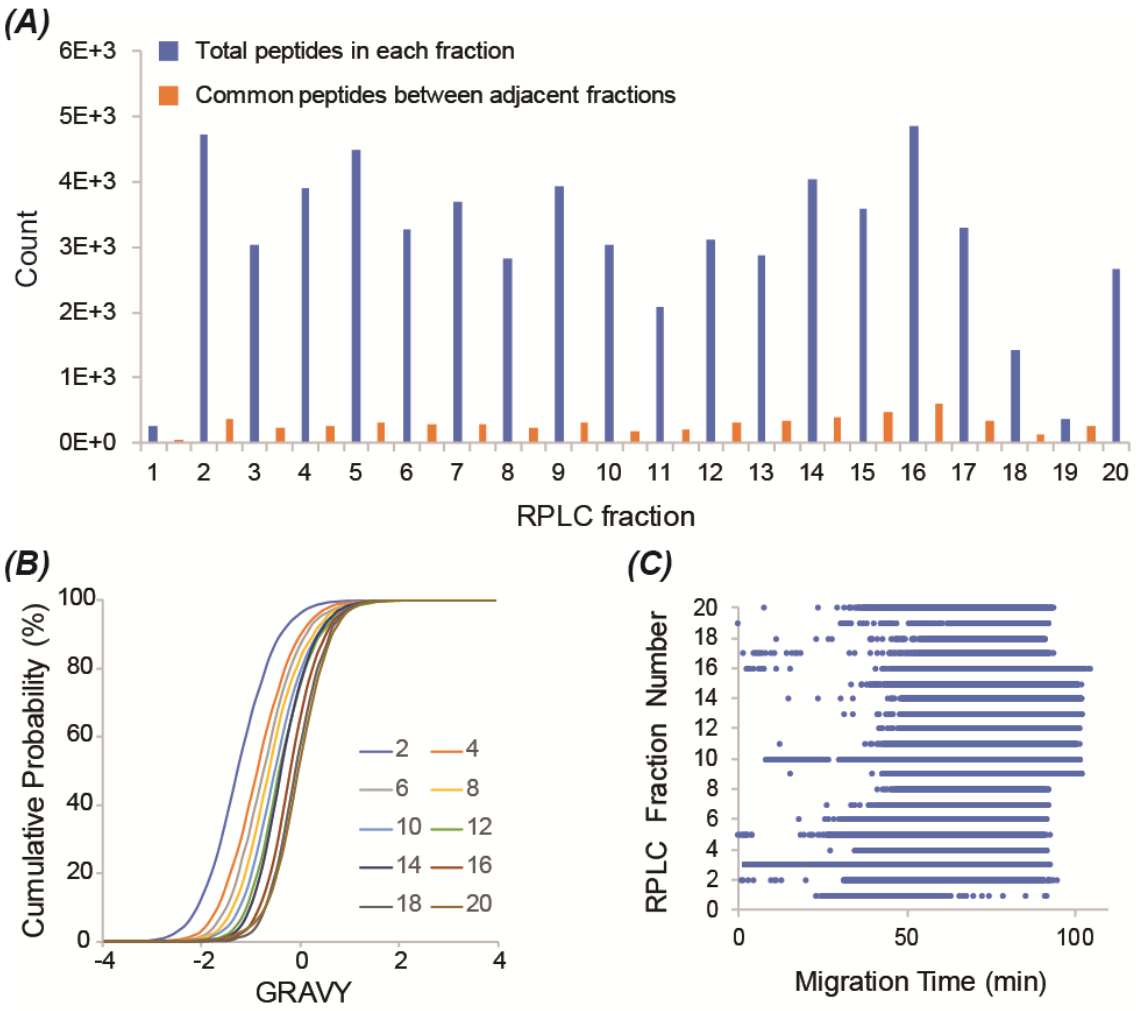

Figure 2. 4 The nanoRPLC fractionation-CZE-MS/MS data. The data was from the 20 RPLC fractions. The bigger RPLC fraction number indicates longer retention time in RPLC. (A) The number of peptide IDs in each nanoRPLC fraction and the number of common peptides between adjacent fractions. (B) The cumulative distribution of the GRAVY of the identified peptides in 10 out of the 20 nanoRPLC fractions. The fraction number was labeled in the figure. Positive GRAVY values indicate hydrophobic and negative values show hydrophilic. (C) The correlation between the RPLC fraction number and the migration time of peptides in CZE. All the identified peptide-spectrum

matches (PSMs) were used to generate this figure. The migration time corresponding to each PSM was obtained directly from the database search result.

Recently, the Mann's group developed a highly sensitive 2D-RPLC-MS/MS system via coupling high-pH nanoRPLC fractionation to low-pH nanoRPLC-MS/MS for deep bottom-up proteomics.[28] In this work, we also applied the nano-2D-RPLC-MS/MS for deep bottom-up proteomics of the MCF7 proteome using 5-µg of the proteome digest as the starting material. We fractionated the digest into 30 fractions with high-pH nanoRPLC and combined the 30 fractions into 15 fractions. Then each fraction was analyzed by low-pH nanoRPLC-MS/MS with a 2-hour gradient. The total MS analyses including the sample loading and column equilibrium took roughly 40 hours. We identified 8605 proteins and 91546 peptides after the database search. Our nanoRPLC-CZE-MS/MS generated 13% lower number of protein IDs (7512 *vs.* 8605 proteins) and 35% lower number of peptide IDs (59403 *vs.* 91546 peptides) compared to nano-2D-RPLC-MS/MS with 15% shorter MS time (34 *vs.* 40 hours). Those two platforms have good complementarity in peptide ID and moderate complementarity in protein ID, **Figure 2.5.** Only 70% of the peptides identified using nanoRPLC-CZE-MS/MS were covered by those identified using nano-2D-RPLC-MS/MS. 677 proteins were only identified by the nanoRPLC-CZE-MS/MS.

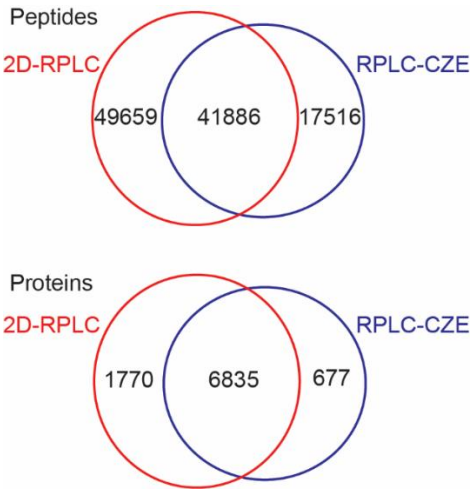

Figure 2. 5 Comparisons between nano-2D-RPLC-MS/MS and nanoRPLC-CZE-MS/MS in terms of the protein-level and peptide-level overlaps.

2.4 Conclusion

We developed two µRPLC-CZE-MS/MS platforms for highly sensitive and deep bottom-up proteomics. The C18 ZipTip-CZE-MS/MS identified 4453 proteins from 5-µg MCF7 proteome digest with good qualitative and quantitative reproducibility. The nanoRPLC-CZE-MS/MS platform reached over 7500 protein IDs and nearly 60000 peptide IDs with only 5-µg MCF7 proteome digest and improved the overall sensitivity of LC-CZE-MS/MS for deep bottom-up proteomics drastically. The nanoRPLC-CZE-MS/MS is complementary with the nano-2D-RPLC-MS/MS in terms of the protein and peptide IDs from the MCF7 proteome. This work provides the proteomics community with a powerful tool for deep proteomics of mass-limited samples.

2.5 Acknowledgments

We thank Prof. Xuefei Huang's group at Department of Chemistry, Michigan State University for kindly providing the MCF7 cells for our research. 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.

REFERENCES

REFERENCES

1. Tong, W.; Link, A.; Eng, J. K.; Yates, J. R. 3rd., *Analytical chemistry* **1999,** *71*, 2270–2278.

2. Garza, S.; Moini, M., *Analytical chemistry* **2006,** *78*, 7309-16.

3. Faserl, K.; Sarg, B.; Kremser, L.; Lindner, H., *Analytical chemistry* **2011,** *83*, 7297-305.

4. Sarg, B.; Faserl, K.; Kremser, L.; Halfinger, B.; Sebastiano, R.; Lindner, H. H., *Molecular & cellular proteomics : MCP* **2013,** *12*, 2640-56.

5. Faserl, K.; Kremser, L.; Müller, M.; Teis, D.; Lindner, H. H., *Analytical chemistry* **2015,** *87*, 4633-40.

6. Wang, Y.; Fonslow, B. R.; Wong, C. C.; Nakorchevsky, A.; Yates, J. R. 3rd., *Analytical chemistry* **2012,** *84*, 8505-13.

7. Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J., *Analytical chemistry* **2012,** *84*, 1617-22.

8. Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J., *Analytical chemistry* **2013,** *85*, 2569-73.

9. Ludwig, K. R.; Sun, L.; Zhu, G.; Dovichi, N. J.; Hummon, A. B., *Analytical chemistry* **2015,** *87*, 9532-7.

10. 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., *Angewandte Chemie* **2014,** *53,* 13931-3.

11. Lombard-Banek, C.; Moody, S. A.; Nemes, P., *Angewandte Chemie* **2016,** *55*, 2454-8.

12. Guo, X.; Fillmore, T. L.; Gao, Y.; Tang, K., *Analytical chemistry* **2016,** *88*, 4418-25.

13. Shen, X.; Sun, L., *Proteomics* **2018,** *18*, e1700432.

14. Chen, D.; Shen, X.; Sun, L., *Analyst* **2017,** *142*, 2118-27.

15. Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J., *Analytical chemistry* **2017,** *89*, 2000-2008.

16. Moini, M., *Analytical chemistry* **2007,** *79*, 4241-6.

17. Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J., *Rapid Commun. Mass Spectrom.* **2010,** *24*, 2554-60.

18. Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J., *Angewandte Chemie* **2013,** *52*, 13661-4.

19. Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J., *J. Proteome. Res*. **2015,** *14*, 2312-21.

20. Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M., *Molecular & cellular proteomics : MCP* **2014,** *13*, 3698-708.

21. Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J., *Molecular & cellular proteomics : MCP* **2014,** *13*, 339-47.

22. Busnel, J. M.; Schoenmaker, B.; Ramautar, R.; Carrasco-Pancorbo, A.; Ratnayake, C.; Feitelson, J. S.; Chapman, J. D.; Deelder, A. M.; Mayboroda, O. A., *Analytical chemistry* **2010,** *82*, 9476-83.

23. Zhang, Z.; Sun, L.; Zhu, G.; Yan, X.; Dovichi, N. J., *Talanta* **2015,** *138*, 117-122.

24. Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J., *Analytical chemistry* **2014,** *86,* 6331-6.

25. Yan, X.; Sun, L.; Zhu, G.; Cox, O. F.; Dovichi, N. J., *Proteomics* **2016,** *16*, 2945-2952.

26. Faserl, K.; Sarg, B.; Sola, L.; Lindner, H. H., *Proteomics* **2017,** *17*, doi: 10.1002/pmic.201700310.

27. Chen, D.; Shen, X.; Sun, L., *Anal. Chim. Acta* **2018,** *1012*, 1-9.

28. Kulak, N. A.; Geyer, P. E.; Mann, M., *Molecular & cellular proteomics : MCP* **2017,** *16*, 694-705.

29. Zhang, Z.; Yan, X.; Sun, L.; Zhu, G.; Dovichi, N. J., *Analytical chemistry* **2015,** *87*, 4572-7.

30. Zhang, Z.; Sun, L.; Zhu, G.; Cox, O. F.; Huber, P. W.; Dovichi, N. J., *Analytical chemistry* **2016,** *88*, 877-82.

31. Choi, S. B.; Lombard-Banek, C.; Muñoz-LLancao, P.; Manzini, M. C.; Nemes, P., *J. Am. Soc. Mass. Spectrom*. **2018,** *29*, 913-922.

32. Aebersold, R.; Morrison, H. D., *J. Chromatogr.* **1990,** *516*, 79-88.

33. Britz-McKibbin, P.; Chen, D. D., *Analytical chemistry* **2000,** *72*, 1242-52.

34. Lubeckyj, R. A.; McCool, E. N.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., *Analytical chemistry* **2017,** *89*, 12059-12067.

35. McCool, E. N.; Lubeckyj, R. A.; Shen, X.; Chen, D.; Kou, Q.; Liu, X.; Sun, L., *Analytical chemistry* **2018,** *90*, 5529-5533.

36. Zhu, G.; Sun, L.; Dovichi, N. J., *Talanta* **2016,** *146*, 839-43.

37. Sun, B.; Kovatch, JR.; Badiong, A.; Merbouh, N., *J. Proteome. Res.* **2017,** *16,* 3711-3721.

38. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., *Analytical chemistry* **2002,** *74*, 5383-92.

39. Elias, J. E.; Gygi, S. P., *Nat. Methods* **2007,** *4*, 207-14.

40. Cox, J.; Mann, M. *Nat. Biotechnol*., **2008,** *26*, 1367-1372.

41. Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M., *Mol. Cell Proteomics* **2014,** *13*, 2513-26.

42. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., *Nat. Methods* **2016,** *13*, 731-740.

43. Zhang, Z.; Peuchen, E. H.; Dovichi, N. J., *Analytical chemistry* **2017,** *89*, 6774-6780.

44. Cohen, D.; Dickerson, J. A.; Whitmore, C. D.; Turner, E. H.; Palcic, M. M.; Hindsgaul, O.; Dovichi, N. J., *Annu. Rev. Analytical chemistry* **2008,** *1,* 165-90.

45. Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q. W.; Wang, R.; Hermjakob. H., *Nucleic. Acids. Res.* **2016,** *44,* 447-456.

46. Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; Moritz, R. L.; Carver, J. J.; Wang, M.; Ishihama, Y.; Bandeira, N.; Hermjakob, H.; Vizcaíno, J. A., *Nucleic. Acids. Res.* **2017,** *45*, 1100-1106.

[3]CHPATER 3. An improved nanoflow RPLC-CZE-MS/MS system with high peak capacity and sensitivity for nanogram bottom-up proteomics

3.1 Introduction

Multi-dimensional separation before electrospray ionization-mass spectrometry (ESI-MS) is indispensable for bottom-up proteomics of complex proteomes.[1-3] Reversed-phase liquid chromatography (RPLC) and capillary zone electrophoresis (CZE) are both capable for high-resolution separation of peptides and are orthogonal.[4,5] Coupling RPLC to CZE-MS for complex sample analysis has been a very active research area in the literature.[6] Both offline and online RPLC-CZE-MS systems have been developed and applied in bottom-up proteomics. In general, the online RPLC-CZE-MS should produce better sensitivity and higher throughput than the offline version due to the reduced sample loss and labor; the offline RPLC-CZE-MS should have much simpler system setup and provide better flexibility for the two-dimensional (2D) separations compared to the online version.[2] It was well demonstrated that CZE-MS outperformed RPLC-MS for analysis of mass-limited proteome samples regarding the number of protein identifications (IDs), suggesting the better sensitivity of CZE-MS compared to RPLC-MS.[7-10] Therefore, CZE is typically online coupled to MS and RPLC is employed as the first dimension for peptide fractionation.

Several research groups have developed online RPLC-CZE-MS systems for proteomics.[11-17] The Ramsey group carried out the connection of RPLC and CZE-MS using a microfluidic device.[15,16] They applied the online RPLC-CZE-MS systems in

[3] Part of this chapter was adapted with permission from: Yang, Z.; Shen, X.; Chen, D.; Sun, L., Journal of proteome research 2019, 18 (11), 4046-4054.

bottom-up MS characterization of standard protein digests, an antibody, and an *E. coli* proteome sample. Under the optimal condition, a peak capacity of over 1000 in one hour was produced by the online system. Recently, the Neusüß group built an online RPLC-CZE-MS system using a 4-port valve as the interface for characterization of intact proteins.[17] An RPLC eluate containing proteins was selectively transferred into a sample loop integrated into the 4-port valve, followed by CZE-MS analysis. Although some successful examples were published, some challenges remain for the online RPLC-CZE-MS. First, the separation power of RPLC and CZE cannot be fully used because we need to balance the two dimensions for online operation. Second, usually only a small fraction of the RPLC eluate can be analyzed by CZE-MS. For instance, in the design of the Ramsey group,[16] only small fractions of RPLC eluates flowed into the microfluidic device and most of them went to waste through splitting to ensure the compatibility of RPLC and CZE.

Offline RPLC-CZE-MS is a good solution for fully using the separation power of both RPLC and CZE and has been applied for large-scale bottom-up proteomics.[5,18-25] Yan *et al*. coupled offline RPLC fractionation to CZE-MS/MS for bottom-up proteomics of *Xenopus* embryos and identified 4100 proteins.[19] Chen *et al*. developed a SCX-RPLC-CZE-MS/MS platform and achieved 8200 protein IDs in 70 h from a mouse brain proteome.[24] Both studies started with hundreds of micrograms of proteome digests. Choi *et al*. coupled a microscale RPLC fractionation using a C18 zip-tip to CZE-MS/MS for analysis of 1 to 20-μg neuron proteome digests.[23] Nearly 800 proteins were identified with consumption of only 1-ng proteome digest, demonstrating the power of microscale RPLC-CZE-MS/MS for bottom-up proteomics of mass-limited samples. Although the

offline RPLC-CZE-MS/MS showed its power for large-scale bottom-up proteomics, only a small fraction of peptides in each RPLC fraction (*i.e.,* less than 5%) was analyzed by CZE-MS/MS in the typical workflow because of the low sample loading capacity of CZE. The offline RPLC-CZE-MS/MS suffered from low overall sensitivity also because the serious sample loss on the analytical RPLC column (2.1-4.6 mm i.d.) used for peptide fractionation and in the Eppendorf tubes used for fraction collection.

The Mann group demonstrated that the RPLC fractionation using a capillary column (*i.e.,* 250-μm i.d.) produced drastically better sensitivity compared to that using an analytical column (*i.e.,* 2.1-mm i.d.) for bottom-up proteomic analyses of low micrograms of human cell proteins, due to the much lower sample loss from the obviously reduced surface area.[26] Recently, we coupled offline nanoRPLC fractionation using a 75-μm-i.d. column at 200 nL/min flow rate to high-capacity CZE-MS/MS for highly sensitive and deep bottom-up proteomics of MCF7 cancer cells.[27] We employed a dynamic pH junction-based sample stacking method[28] to improve the sample loading capacity of CZE and enabled a 500-nL sample injection from a 1.5-μL peptide solution, using up to 33% of the available peptide material for a CZE-MS/MS analysis. The combination of nanoRPLC and dynamic pH junction-based CZE-MS/MS identified 7500 proteins and nearly 60000 peptides starting from only 5-μg of a MCF7 proteome digest. The nanoRPLC-CZE-MS/MS can fully use the separation power of RPLC and CZE and has great overall sensitivity because the RPLC eluates can be efficiently used for CZE-MS/MS analysis.

Built upon our preliminary work, here we present an improved nanoRPLC-CZE-MS/MS with much better sensitivity than our previous system. The improved system

enabled deep bottom-up proteomic analysis of nanograms of MCF7 proteome digests with the production of 6500 protein IDs starting with only 500-ng MCF7 peptides using a Q-Exactive HF mass spectrometer. Only roughly 100 ng of peptides were actually consumed. We further applied the improved nanoRPLC-CZE-MS/MS in bottom-up proteomics of 5000 HEK293T cells. The single spot solid phase sample preparation (SP3) method[29,30] was employed for preparing the 5000-cell sample. We identified about 3700 proteins with the consumption of protein content of roughly 1000 cells. The optimized nanoRPLC-CZE-MS/MS system showed high sensitivity for bottom-up proteomics because of several novelties.

First, we applied a short capillary (*i.e.,* 70-cm long) in CZE separation and the sample loading volume was 500 nL, corresponding to 36% of the total capillary volume. One CZE-MS/MS run was completed in an hour with an average peak capacity of nearly 200. The peptides migrated drastically faster compared to our previous work with a 1-meter-long capillary,[27] resulting in sharper peaks and higher peptide intensity, which are important for analysis of mass-limited samples.

Second, we pretreated the inner wall of sample vials of CZE with bovine serum albumin (BSA) to reduce the nonspecific peptide adsorption. The BSA treatment improved the peptide intensity by nearly 200%.

Third, lyophilization and redissolution of nanoRPLC eluates before CZE-MS/MS analysis were avoided to reduce sample loss, to decrease labor, and to improve throughput. The nanoRPLC eluates were directly collected into 0.6-mL low-retention Eppendorf tubes containing 1.4-2.4 μL of an ammonium bicarbonate (NH_4HCO_3) buffer

(pH 8.5). Then the samples were transferred into the BSA-treated sample vials of CZE for the dynamic pH junction-based CZE-MS/MS without any sample preparation steps.

Fourth, we coupled the SP3-based sample preparation method with nanoRPLC-CZE-MS/MS for bottom-up proteomics of 5000 HEK293T cells. The SP3 method has high sample recovery for preparation of mass-limited proteome samples based on the literature.[29,30] To our best knowledge, this is the first report of bottom-up proteomics of a small number of human cells using CZE-MS/MS.

3.2 Experiment

3.2.1 Material and reagents

All reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. Urea was purchased from Alfa Aesar (Haverhill, MA). LC/MS grade water, methanol, acetonitrile (ACN), HPLC grade acetic acid (AA), formic acid (FA), and hydrofluoric acid (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 purchased from Polymicro Technologies (Phoenix, AZ). Carboxylate-modified paramagnetic beads (Sera-Mag SpeedBeads (hydrophilic) CAT # 45152105050250, and Sera-Mag SpeedBeads (Hydrophobic), CAT # 65152105050250) were purchased from GE Healthcare. 0.6 mL Eppendorf tubes (CAT # C-2176) were purchased from Denville (Saint-Laurent, Canada). CZE insert tubes (CAT # C4010-630P) were purchased from Thermo Scientific.

3.2.2 MCF7 cell culture and proteome digestion

The MCF7 cells were kindly provided by Dr. Xuefei Huang's lab from chemistry department, Michigan State University. The cells were cultured at 37°C under a 5% CO_2

in ATCC-formulated Eagle's Minimum Essential Medium, supplemented with 0.01 mg/ml human recombinant insulin and 10% fetal bovine serum. After cell culture, cells were harvested and rinsed with 10 mL PBS buffer a couple times. The cells were then lysed in 8 M urea buffer (100 mM NH_4HCO_3, pH = 8.0 protease inhibitor) with ultrasonication (Branson Sonifier 250, VWR Scientific, Batavia, IL) on ice for 10 min. After lysis, cells were centrifuged at 10,000 g for 10 min. The supernatant was subjected for BCA assay for protein concentration measurement.

250 µg of MCF7 protein were precipitated with 4 times (v/v) ice cold acetone and stored in -20 °C overnight. The protein pellet was obtained through centrifugation at 14,000 g for 15 min and was washed once by cold acetone. The protein was then dissolved with lysis buffer (8 M urea, 100 mM NH_4HCO_3, pH = 8.0). Reduction was implemented with addition of dithiothreitol (DTT) at 37 °C for 30 min and alkylation was implemented with addition of iodoacetamide (IAA) at room temperature for 20 min in the dark. DTT was added again to quench the extra IAA. The protein solution was then diluted with 500 µL of NH_4HCO_3 buffer (100 mM, pH = 8.0). 8 µg of trypsin (Bovine pancreas TPCK-treated) was added. Digestion was performed under 37 °C overnight. The digestion was quenched by adding 2 µL of formic acid. The peptides were then desalted by Sep-Pak C18 Cartridge (Waters) and lyophilized. The peptides were stored in -80 °C for further use.

3.2.3 HEK293T cell culture and SP3-based sample preparation

The HEK293T cell was kindly provided by Prof. Jian Hu’s lab at the Department of Biochemistry of Michigan State University. The HEK293T cells were grown in T75

flasks contained 1x Gibco Dulbecco's Modified Eagle Medium (DMEM) without sodium pyruvate at 37°C under 5% CO_2.

HEK293T cells were harvested, washed with PBS buffer a few times and then were immediately subjected to flow cytometry for cell sorting (BD Influx, BD Bioscience). Cell were sorted into 1.7 mL Eppendorf tubes. 5 µL cell lysis buffer (4% SDS (w/v), Protease inhibitor, 100 mM NH_4HCO_3, pH 8) was added into 5000-cell aliquot so final SDS content was about 1% (w/v). Cell lysis was performed by ultrasonication (Branson Sonifier 250, VWR Scientific, Batavia, IL) on ice for 10 min and then 95 °C incubation for 5 min. 2 µL 0.5 mM DTT was added into the sample. Sample was incubated under 37 °C for 30 min for reduction. Sample was then added with 2.5 µL of 1 mM IAA and incubated under room temperature for 20 min in dark for alkylation. 1 µL of 0.5 mM DTT was added to quench IAA.

2 µL of each Carboxylate-modified paramagnetic beads stock solution (Sera-Mag SpeedBeads (hydrophilic) CAT # 45152105050250, and Sera-Mag SpeedBeads (Hydrophobic), CAT # 65152105050250) were combined and washed with water a few times. The beads mixture was then added into the 5000-cell lysis solution treated with DTT and IAA. Solution was quickly mixed through vortex. 120-µL ACN was then added so the final ACN content was over 70% (v/v). Cell solution with beads was incubated under room temperature for 18 min. Then, a magnet was placed under the tube for 2 min. The supernatant was taken with magnet on. Beads were then washed with 200 µL of 70% (v/v) ethanol for twice and 100 % ACN once to remove SDS. 10 µL of NH_4HCO_3 buffer (100 mM, pH 8) was added into the tube to resuspend the beads. 50 ng of trypsin was then added for digestion under 37 °C overnight. After digestion, the sample tube

was spun down quickly and was added with 200 µL of 100 % ACN. Sample was incubated at room temperature for 18 min and then on magnet for 2 min. The supernatant was then taken out. The beads were then rinsed with 200 µL of 100 % ACN and then resuspended with 3 µL of NH_4HCO_3 buffer (50 mM, pH8) to release the peptides off beads.

3.2.4 NanoRPLC fractionation

EASY-nLC 1200 (Thermo Fisher Scientific) equipped with a capillary column (75 µm i.d. x 50 cm Length, C18, 2 µm bead, 100 Å pore, Thermo Fisher Scientific) was used for fractionation. Mobile phase A contained 2% (v/v) acetonitrile (ACN), 98% (v/v) H_2O and 0.1% (v/v) formic acid (FA). Mobile phase B contained 80% (v/v) ACN, 20% (v/v) H_2O and 0.1 % (v/v) FA. Flow rate was 200 nL/min. One-column set up was selected on the nanoRPLC system. 0.6-mL low-retention Eppendorf tube (Denville, Canada) was used for fraction collection.

50-fraction procedure. 5 µg and 500 ng of MCF7 proteome digests were fractionated into 50 fractions with a 3-h gradient using nanoRPLC, respectively. The gradient was set up as follows: from 8% to 30% (v/v) B in 100 min, from 30% to 50% (v/v) B in 50 min, from 50% to 80% (v/v) B in 15 min and stay at 80% (v/v) B for 15 min. The first fraction collection started at the sample loading and traversed the first 20 min of gradient. Then each fraction was collected every 3 min. The last fraction was collected from the 164 min to the end of the gradient. For 5-µg sample fractionation, 2.4 µL of 50 mM NH_4HCO_3 (pH 8.5) was deposited to the bottom of the 0.6-mL Eppendorf tube. For all fractions except the first and the last ones, when 0.6 µL of eluate (200 nL/min for 3 min) flowed into the buffer, the final pH was around 8.0. For 500-ng sample

fractionation, the same procedure was used but only 1.4 µL of the NH_4HCO_3 buffer (pH 8.5) was deposited into the tube, resulting in a total sample volume of 2 µL for each fraction. For the first and last nanoRPLC fractions, the eluates were lyophilized and redissolved in 3 µL (for the 5-µg sample) and 2 µL (for the 500-ng sample) of a 50 mM NH_4HCO_3 buffer (pH 8.0) for CZE-MS/MS.

20-fraction procedure. A 500-ng MCF7 cell proteome digest was fractionated into 20 fractions with a 100-min gradient. The 100-min gradient was set up as follows: from 8% to 30% (v/v) B in 60 min, from 30% to 50% (v/v) B in 30 min, from 50% to 80% (v/v) B in 5 min and stay at 80% (v/v) B for 5 min. First fraction collection started from sample loading and the first 15 min of gradient. Then each fraction was collected every 4 min. The last fraction collection started from 87 min and kept collecting until the end of the gradient. A 1.4 µL of 50 mM NH_4HCO_3 buffer (pH 8.5) was deposited into the tube first. 0.8 µL of the nanoRPLC eluate was collected in the tube and the final sample volume of each fraction was 2.2 µL.

10-fraction procedure. The proteome digest from 5000 HEK293T cells was fractionated by the nanoRPLC system into 10 fractions using a 60-min gradient with a similar procedure to the 20-fraction collection. The gradient was set up as follows: from 8% to 30% (v/v) B in 30 min, from 30% to 50% (v/v) B in 20 min, from 50% to 80% (v/v) B in 5 min and stay at 80% (v/v) B for 5 min. First fraction collection started from sample loading and the first 15 min of gradient. Then each fraction was collected every 4 min. The last fraction collection started from 47 min and kept collecting until the end of the gradient. 1.4 µL of 50 mM NH_4HCO_3 buffer was deposited into the tube first. 0.8 µL of

the eluate was combined into the NH_4HCO_3 buffer with a final sample volume of 2.2 µL and a final pH of around 8.

High pH 36-fraction procedure. 500 ng of MCF-7 cell digest was fractionated by nanoRPLC system into 36 fractions using 100 min gradient with a similar procedure as mentioned above. The gradient was set as follows: from 8% to 30% (v/v) B in 45 min, from 30% to 50% (v/v) B in 40 min, from 50% to 80% (v/v) B in 5 min and stay at 80% (v/v) B for 10 min. First fraction collection started from sample loading and the first 15 min of gradient. Then each fraction was collected every 2 min. The last fraction collection started from 83 min and kept collecting until the end of the gradient. 2 µL of water solution containing 0.1% (v/v) formic acid was first deposited into the tube for fraction collection. The 36 fractions were then combined into 18 fractions (for example, combining fractions 1 and 19, 2 and 20, 3 and 21, ----, 18 and 36) for nanoRPLC-MS/MS analysis.

3.2.5 Pretreatment of sample vials of CZE-MS and nanoRPLC-MS with BSA

The sample injection vials of CZE-MS and nanoRPLC-MS were treated with a BSA solution to reduce non-specific adsorption of peptides on the inner wall of vials. 10 µL of 2 mg/mL BSA solution was added into each sample injection vial and was incubated at room temperature for 10 min. After the BSA solution was removed, each vial was rinsed with 500 µL of a NH_4HCO_3 buffer (10 mM, pH 8) twice, followed by air dry in the chemical hood. The treated sample vials were ready for use. We need to note that the BSA treated sample vials were only used for the 500-ng MCF7 proteome samples and the 5000 HEK293T cell sample.

3.2.6 CZE-MS/MS

The inner wall of the CZE capillary (50/360 μm i.d./o.d.) was coated with linear polyacrylamide (LPA) based on the procedure described in reference[31] and the LPA-coated capillary was etched with HF to reduce its outer diameter based on the protocol described in reference[32].

The commercialized electrokinetically pumped sheath flow CE-MS interface (CMP scientific, Brooklyn, NY) was used to couple CZE to MS.[33, 34] An ECE-001 autosampler (CMP scientific) was used to carry out the automated CZE operation. The background electrolyte (BGE) for CZE was 5 % (v/v) AA (pH = 2.4). The sheath liquid was 10% (v/v) methanol and 0.2% (v/v) FA. The glass emitter for the electrospray was pulled from borosilicate glass capillary (0.75 mm i.d., 1 mm o.d.) by a Sutter P-1000 flaming/brown micropipette puller. The orifice of the emitter was controlled at 20-40 μm. The distance of the etched capillary tip to the emitter orifice was less than 300 μm and the distance of the emitter orifice to the MS entrance was 2 mm.

A 70-cm-long capillary and a 100-cm-long capillary were used for CZE-MS/MS. For the 70-cm capillary, the sample was loaded with 5-psi pressure for 60 s, leading to the injection of approximately 500 nL of sample. For the 100-cm capillary, the sample was loaded with 5-psi pressure for 90 s, leading to the injection of about 500-nL sample. After sample loading, the sample injection end of the capillary was moved into a BGE vial and 30-kV voltage was applied for separation. For the 70-cm capillary, separation was carried out for 50 or 60 min and for the 100-cm capillary, separation was carried out for 115 min. A 15-psi pressure was applied afterwards for 5 min to flush and condition

the capillary. A 2-kV voltage was applied in the sheath buffer vial for electrospray ionization.

A Q-Exactive HF (Thermo Fisher Scientific) mass spectrometer was use for all CZE-MS/MS analyses. For the 5-µg proteome digest analysis, the full MS scan range was set 300–1800 *m/z*. Resolution was set 60,000 (at *m/z* 200) and AGC was set 3E6. Maximum injection time was set 50 ms. The resolution of MS/MS was set 30,000 and AGC was set 1E5. Maximum injection time was set 50 ms and the loop count was set 10 (top 10). Quadrupole isolation window was set 2 *m/z* and the normalized fragmentation energy was set 28. The ion intensity threshold was set 5E4 and the dynamic exclusion window was set 30 s. For the 500-ng proteome digest and the 5000 HEK293T cell sample, to have better identification of low abundant peptides, we modified the acquisition parameters. The resolution, AGC and maximum injection time of full MS were set the same as the 5-µg proteome sample. However, the full MS scan range was adjusted to 300–1200 *m/z*. For MS/MS scan, the resolution was set 60,000 and the maximum injection time was set 200 ms. Quadrupole isolation window was set 4 *m/z*. the ion intensity threshold was set 1E4. All the rest of parameters were kept the same.

3.2.7 NanoRPLC-MS/MS

The peptide eluate in each combined fraction was transferred into an insert tube coated with BSA for nanoRPLC-MS/MS. About 75% of the peptide material in each fraction was loaded onto an EASY nanoLC-1200 (Thermo Fisher Scientific) for low pH nanoRPLC separation. A capillary column (75 µm i.d. x 50 cm, C18, 2 µm, 100 Å, Thermo Fisher Scientific) was employed for RPLC. Buffer A containing 0.1% (v/v) FA

and 2% (v/v) ACN and buffer B containing 80% (v/v) ACN and 0.1% (v/v) FA were used to generate gradient separation. Each sample was loaded onto the RPLC column with buffer A at 800-bar pressure. Then the peptides retained on the column were separated by a linear gradient. The flow rate was 200 nL/min. The gradient for RPLC separation was as follows: from 8 to 30% (v/v) B in 60 min, from 30% to 50% (v/v) B in 45 min, from 50% to 80% (v/v) B in 5 min and maintain at 80% (v/v) B for 10 min.

The MS parameter settings of each nanoRPLC-MS/MS analysis were all consistent with CZE-MS/MS analysis for 500-ng proteome digest except 120 min was set for data acquisition.

3.2.8 Data Analysis

Database searching was performed on Proteome discoverer 2.2 (Thermo Fisher Scientific) with SEQUEST HT search engine.[35] Database search parameters were set as follows: precursor ion mass tolerance was set 20 ppm; product ion mass tolerance was set 0.05 Da; UniProt human proteome database (UP000005640) was used as database; the false discovery rate (FDR%) was evaluated through target-decoy database search approach.[36,37] Trypsin was set for enzyme digestion with maximum 2 missed cleavage. Oxidation on methionine, acetylation on protein N-terminal and deamination on asparagine or glutamine were set as variable modifications. Carbamidomethylation on cysteine was set as fixed modification. The peptide FDR was set as 1%. Protein grouping was enabled, and the strict parsimony principle was applied.

MaxQuant 1.5.5.1 was also used for database search to obtain label free quantification (LFQ) information between different experiments.[38,39] All the parameters

were set as default during database search except that when LFQ was enabled, normalization was skipped to present the true quantity difference between different experiments. UniProt human proteome database (UP000005640) was used as the database. ExPASy Bioinformatic Resource Portal (https://web.expasy.org/compute_pi/) was used to calculate peptide isoelectric point (pI) and molecular weight (MW). A GRAVY calculator (http://www.gravy-calculator.de/) was used to calculate the GRAVY values of peptides. Positive GRAVY values indicate hydrophobic and negative GRAVY values suggest hydrophilic.

3.3 Results and discussion

3.3.1 Comparing 100-cm-long and 70-cm-long capillaries for CZE-MS/MS

Longitudinal diffusion is usually the only factor causes band broadening in CZE. The longer time analytes spend in the capillary, the more band broadening would be. In our previous work, we used a 100-cm-long LPA-coated capillary for CZE separation.[27] The LPA coating significantly reduced the electroosmotic flow in the capillary, which is crucial for effective dynamic pH junction-based sample stacking and a wide separation window. In a typical run of dynamic pH junction-based CZE-MS/MS using a 100-cm-long LPA-coated capillary, we identified 8172 peptides and 1896 proteins from a MCF7 proteome digest with a Q-Exactive HF mass spectrometer in 95 min. 100-ng peptides in 500 nL of 50 mM NH_4HCO_3 (pH 8.0) were injected for analysis. The effective separation window was about 1 h, **Figure 3.1A**. The peak capacity was roughly 300 based on the full width at half maximum (FWHM) of peptides. We performed peak extraction at m/z 593.33 with a mass tolerance of 10 ppm and observed three peaks corresponding to three different peptides at varied migration time, **Figure 3.1B**. The peptide peak

became obviously wider with the increase of migration time, most likely due to more significant diffusion in the capillary. Interestingly, the basic peptides clearly tended to migrate faster than the acidic peptides in the dynamic pH junction-based CZE, **Figure 3.1C**.

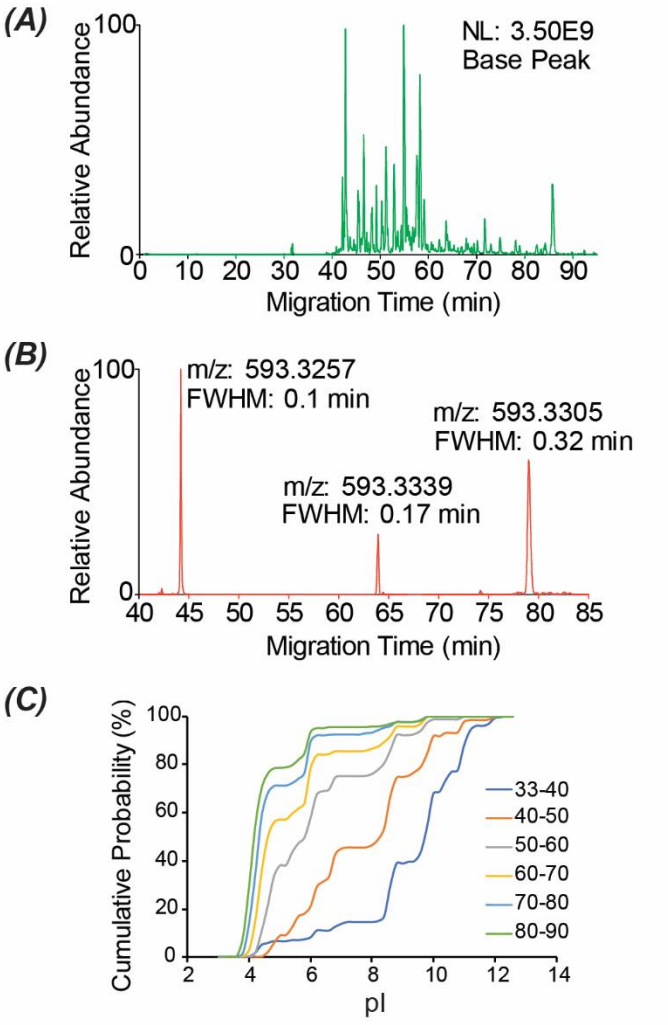

Figure 3. 1 CZE-MS/MS analysis of a 100-ng MCF7 proteome digest. (A) Base peak electropherogram of the CZE-MS/MS analysis. (B) Extracted ion electropherogram of three peptides with similar m/z. The m/z for peak extraction was 593.33 with a mass tolerance of 10 ppm. The m/z and full width at half maximum (FWHM) were labelled. (C) Cumulative distribution of isoelectric point (pI) of peptides identified by CZE-MS/MS in

different periods during a run. The identified peptides from 33 to 90 min were separated into six groups based on their migration time.

We believe that by reducing the migration time of peptides, sharper peptide peaks and higher peptide intensity could be achieved, which is important for CZE-MS/MS analysis of mass-limited samples. To reduce peptide migration time, one option is to use a shorter capillary. However, a shorter capillary can narrow the separation window, reduce the peak capacity, and decrease the number of MS/MS spectra from one CZE run. To ensure the overall peak capacity and number of MS/MS spectra of the nanoRPLC-CZE-MS/MS system, we need to collect a higher number of fractions for CZE-MS/MS analysis.

We first tested our idea by analyzing a 5-µg MCF7 proteome digest with nanoRPLC-CZE-MS/MS, in which a 70-cm-long LPA-coated capillary was employed for CZE-MS/MS analyses of 50 nanoRPLC fractions (70-cm-50-fractions). We also compared the data with our previous work that utilized a 100-cm-long LPA-coated capillary for CZE-MS/MS analyses of 20 nanoRPLC fractions (100-cm-20-fractions).[27]

For the 70-cm-50-fractions study, we loaded 5 µg of a MCF7 proteome digest onto a 50-cm-long nanoRPLC column for fractionation over a 3-h gradient as we did in our 100-cm-20-fractions study in reference 27. To reduce additional sample handling steps, here we put 2.4 µL of 50 mM NH_4HCO_3 buffer (pH 8.5) in each sample collection tube. When 0.6 µL of eluate (0.2 µL/min flow rate for 3 min) was combined with the NH_4HCO_3 buffer, the final pH was around 8 and final volume was 3 µL. 50 fractions were collected. Each fraction was then directly analyzed by the dynamic pH junction-based CZE-MS/MS with a 70-cm-long LPA-coated capillary. For each fraction, 500 nL

out of 3 µL sample was injected. The MS analysis time per fraction was 55 min. In total, we identified 7546 proteins and 66990 peptides from the 5-µg MCF7 proteome digest. The 70-cm-50-fractions system produced similar numbers of protein IDs and peptide IDs to the 100-cm-20-fractions system in reference 27 (7546 *vs.* 7512 proteins, 66990 *vs.* 59403 peptides) but with 50% less peptide material consumption (0.8 µg *vs.* 1.6 µg). The result indicates that the 70-cm-50-fractions system has better sensitivity than the 100-cm-20-fractions platform. Less peptide material consumption also saves more material for additional analysis.

We investigated the protein-level and peptide-level overlaps between the data in the 70-cm-50-fractions and 100-cm-20-fractions studies, **Figure 3.2A**. The two studies shared over 70% of protein IDs and only 26% of the peptide IDs. Considering the low peptide-level overlap, we further investigated the physicochemical properties of peptides identified in the two studies, **Figures 3.2B-2D**. The 70-cm-50-fractions system clearly tended to identify basic peptides compared to the 100-cm-20-fractions system, **Figure 3.2B**. The two systems showed no drastic differences in peptide molecular weight (MW) and GRAVY value, **Figures 3.2C** and **3.2D**.

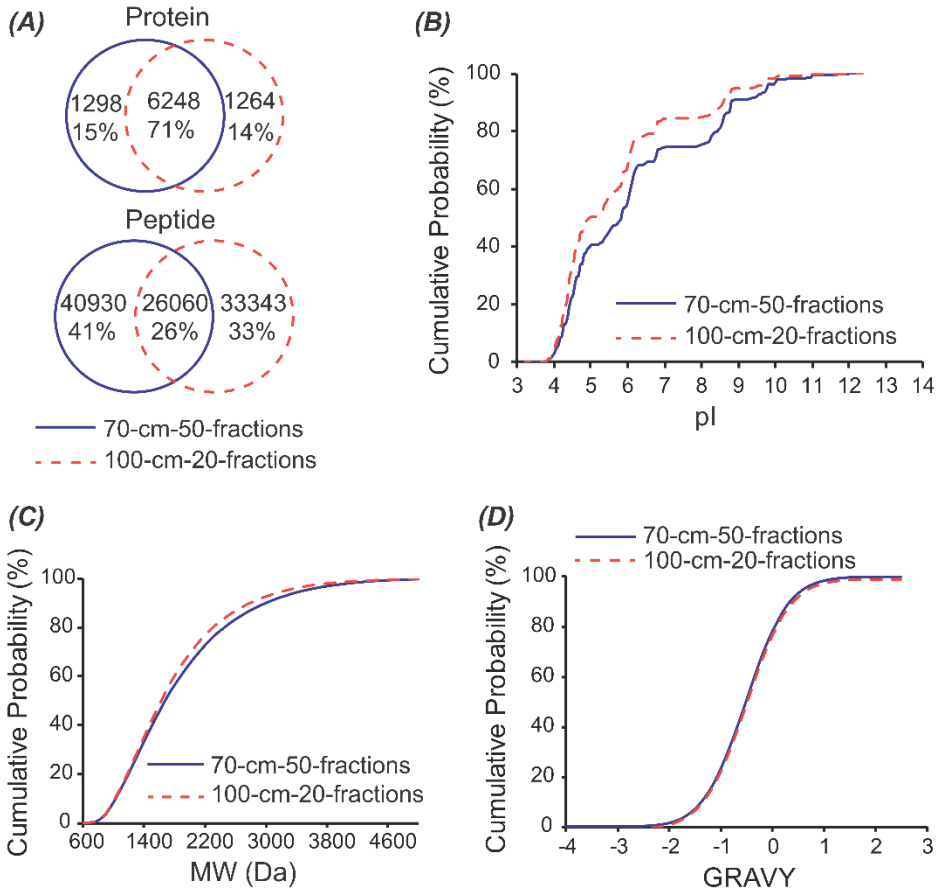

Figure 3. 2 Summary of the 5-µg MCF7 proteome digest data from 70-cm-50-fractions and 100-cm-20 fractions experiments. (A) Overlaps of identified proteins and peptides. (B) Cumulative distribution of isoelectric point (pI) of peptides. (C) Cumulative distribution of molecular weight (MW) of peptides. (D) Cumulative distribution of GRAVY (grand average of hydropathy) value of peptides. Positive GRAVY values indicate hydrophobic and negative GRAVY values suggest hydrophilic.

3.3.2 NanoRPLC-CZE-MS/MS for bottom-up proteomic analysis of 500-ng MCF7 proteome digests

The 70-cm-50-fractions system has better sensitivity than our previous 100-cm-20-fractions platform based on our data. Therefore, we further applied the 70-cm-50-fractions method in the analysis of a 500-ng MCF7 proteome digest.

We made some modifications on the nanoRPLC fraction collection and CZE-MS sample vial compared to the 5-µg-sample study to reduce the sample loss further. First, we only put 1.4-µL NH_4HCO_3 buffer in the Eppendorf tube for nanoRPLC fraction collection and we got 2-µL sample with pH about 8 in each Eppendorf tube after collection of 0.6-µL eluate from nanoRPLC. Second, we pretreated the sample vial used for CZE-MS/MS with a BSA solution to block nonspecific adsorption of peptides on the inner wall of the sample vial.

25% of peptides in each nanoRPLC fraction (500 nL out of the 2 µL sample) was loaded for CZE-MS/MS analysis, corresponding to consumption of only roughly 100-ng MCF7 proteome digest in total. We identified 6492 proteins and 47342 peptides using the 70-cm-50-fractions system. The total MS time was 45 h. The MS raw data have been deposited to the ProteomeXchange Consortium via the PRIDE[40] partner repository with the dataset identifier PXD014392.

Recently, the Mann group developed a highly sensitive 2D nanoRPLC-MS/MS system that employed a microscale capillary column for high-pH RPLC fractionation followed by low-pH nanoRPLC-MS/MS analysis. [26] The system identified 5724 proteins and 23765 peptides by MS/MS in 16 h using a Q-Exactive HF mass spectrometer with the consumption of 500-ng HeLa proteome digest. [26] Our improved nanoRPLC-CZE-MS/MS system identified 13% and nearly 100% more protein and peptide IDs than the 2D nanoRPLC-MS/MS system with four-times lower sample consumption (125 ng *vs.* 500 ng). We noted that our system consumed much longer MS time than the 2D nanoRPLC-MS/MS system (45 h vs. 16 h). We also tried to collect 20 nanoRPLC fractions for CZE-MS/MS to reduce the total MS analysis time. 4551 proteins and 24559

peptides were identified by nanoRPLC-CZE-MS/MS in 18 h with the consumption of about 100-ng peptides in total. The number of protein IDs is about 20% lower than that in reference 26 with comparable instrument time but 4-fold lower sample consumption.

To make a fair comparison between our nanoRPLC-CZE-MS/MS and the 2D-nanoRPLC-MS/MS for analysis of mass-limited proteome samples, we also analyzed a 500-ng MCF7 proteome digest using 2D-nanoRPLC-MS/MS with the same mass spectrometer. We fractionated the 500-ng digest into 36 fractions using high-pH nanoRPLC and combined them into 18 fractions. Each fraction was analyzed by low-pH nanoRPLC-MS/MS in a 2-h gradient. Considering the sample loading time of nanoRPLC-MS/MS, the total MS time for the 2D-nanoRPLC-MS/MS was about 45 h. The 2D-nanoRPLC-MS/MS identified 4758 proteins and 23589 peptides. Using the same mass spectrometer and instrument time, our nanoRPLC-CZE-MS/MS system identified 36% more proteins (6492 *vs.* 4758) and 100% more peptides (47342 *vs.* 23589) than the 2D-nanoRPLC-MS/MS from 500-ng MCF7 proteome digests. We noted that our nanoRPLC-CZE-MS/MS consumed 3-times lower (125 ng *vs.* 375 ng) peptides compared to the 2D-nanoRPLC-MS/MS. About 65% and 88% of the identified peptides and proteins from 2D-nanoRPLC-MS/MS were covered by the data from nanoRPLC-CZE-MS/MS, **Figure 3.3**.

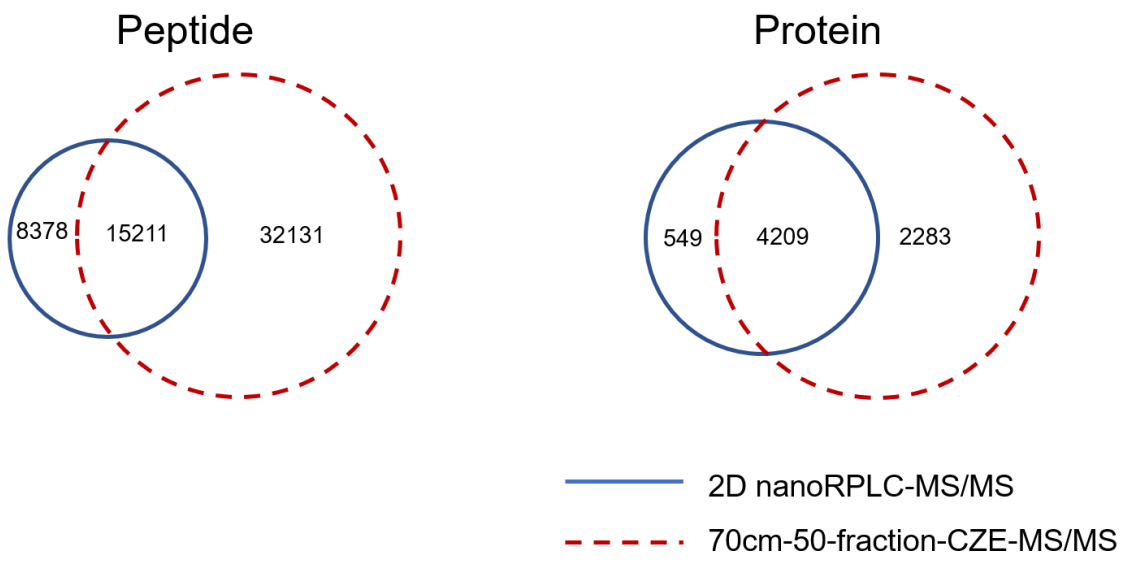

Figure 3. 3 The protein-level and peptide-level overlaps between the 2D-nanoRPLC-MS/MS and nanoRPLC-CZE-MS/MS analyses of 500-ng MCF-7 proteome digests.

The nanoRPLC-CZE-MS/MS system in this work showed high sensitivity due to several reasons. First, the system produced extremely high peak capacity for peptide separation. On average, each CZE run had a peak capacity of 170 in 55 min. The nanoRPLC-CZE with 50 fractions produced a peak capacity of 8500 for peptide separation. The extremely high peak capacity was most likely because nanoRPLC and CZE were well orthogonal for peptide separation. **Figure 3.4** shows the base peak chromatogram of nanoRPLC-MS/MS analysis of a 500-ng MCF7 proteome digest as well as base peak electropherograms of 20 out of 50 nanoRPLC fractions of the 500-ng digest after CZE-MS/MS analyses. It is clear that one 3-min nanoRPLC fraction can be further separated into an up to 40-min window by CZE.

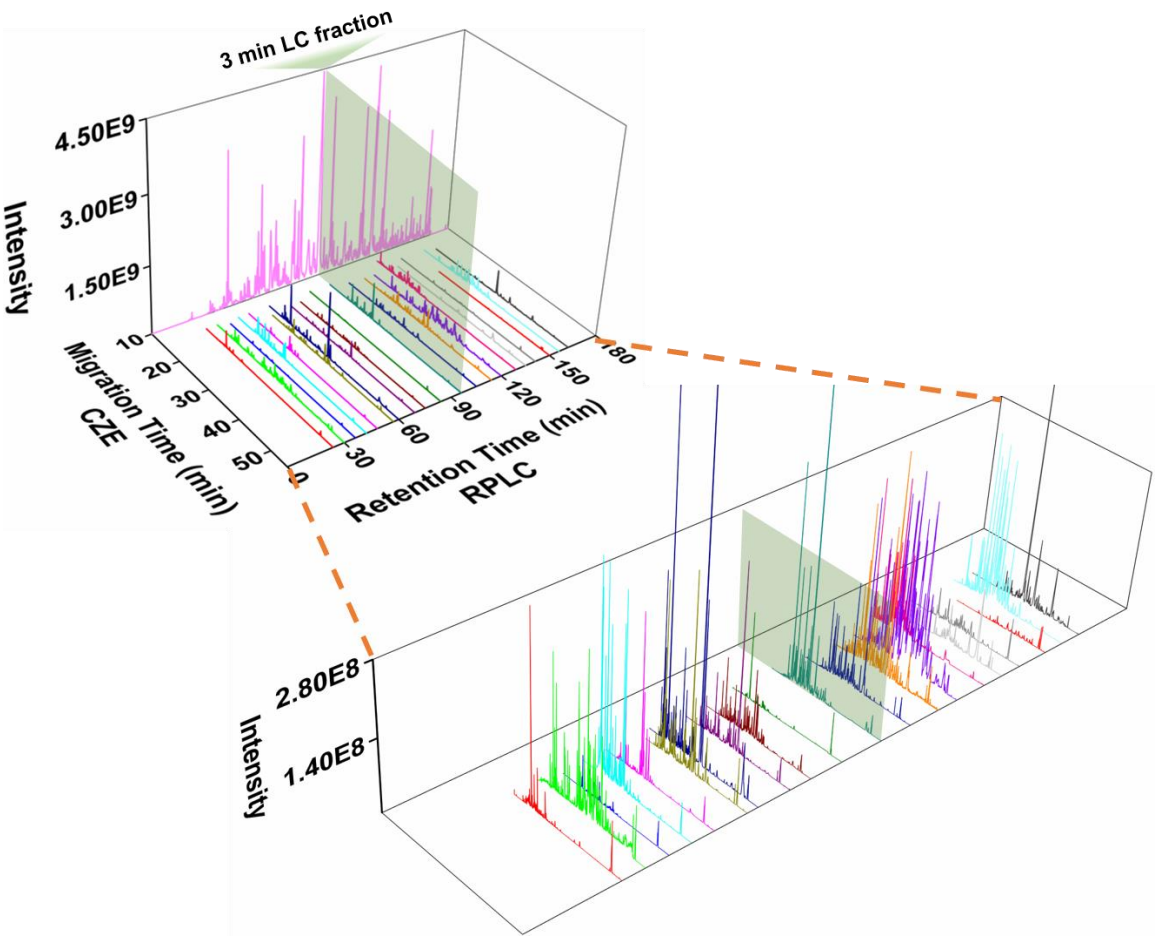

Figure 3. 4 Orthogonality of nanoRPLC and CZE. 3D plot of a base peak chromatogram of a 500-ng MCF7 proteome digest and base peak electropherograms of 20 nanoRPLC fractions from the 70-cm-50-fractions study of the 500-ng MCF7 proteome digest. A zoom-in plot is also shown in the figure.

Second, the pretreatment of CZE-MS sample vial with BSA drastically reduced sample loss. We compared the number of protein IDs, the number of peptide IDs, protein intensity, and peptide intensity between nanoRPLC-CZE-MS/MS studies with and without sample vial pretreatment. Twenty nanoRPLC fractions were collected for CZE-MS/MS analysis in both cases. The non-BSA treated study resulted in 39% and 58% lower number of protein (2758 vs. 4551) and peptide (10230 vs. 24559) IDs

compared to the BSA-treated study. We also compared the LFQ intensity of identified proteins in both studies, **Figure 3.5A**. The protein LFQ intensities had good correlation between the two conditions (R^2 = 0.93) with a slope of 4.5, indicating drastically higher protein intensity with BSA-treated sample vials compared to non-treated vials. The median of peptide intensities from the BSA-treated sample vial was 2.6-fold higher than that from the non-treated sample vial, **Figure 3.5B**.

Third, the 70-cm-50fractions system improved the sensitivity of nanoRPLC-CZE-MS/MS obviously compared to the 100-cm-20fractions system for analysis of 500-ng MCF7 proteome digest. We fractionated 500 ng of a MCF7 proteome digest into 20 fractions with nanoRPLC and analyzed each fraction with CZE-MS/MS using a 100-cm-long capillary (100-cm-20fractions). The CZE sample vials were treated with BSA. Each CZE-MS/MS run took 120 min, and the total MS time was 40 h. The 100-cm-20fractions system identified 4723 proteins and 24975 peptides from the 500-ng MCF7 proteome digest. The number of protein and peptide IDs was 27% and 47% lower compared to that from the 70-cm-50fractions system with comparable MS time (40 vs. 45 h). The protein LFQ intensity showed good correlation (R^2 = 0.93) between the two systems with a slope of 2.3, indicating that higher protein LFQ intensities were obtained with the 70-cm-50fractions system, **Figure 3.5C**.

We need to note that the number of protein and peptide IDs using our nanoRPLC-CZE-MS/MS system can be boosted drastically with a more advanced mass spectrometer. In this work, a Q-Exactive HF mass spectrometer was employed and we have shown that single-shot CZE-MS/MS analysis of a MCF7 proteome digest identified nearly 2000 proteins and 8000 peptides. Over 4000 protein IDs and 27000 peptide IDs

have been achieved by single-shot CZE-MS/MS using an Orbitrap Fusion Lumos mass spectrometer with an advanced-peak-determination algorithm.[41]

We also evaluated the reproducibility of our improved nanoRPLC-CZE-MS/MS system. We fractionated a 500-ng MCF7 proteome digest into 20 fractions followed by CZE-MS/MS analysis with a 70-cm-long capillary. We performed the experiment in duplicate. The system identified 4508±60 proteins and 25034±671 peptides. The duplicate analyses shared about 70% of the identified proteins. The system was highly reproducible regarding the LFQ protein intensity, **Figure 3.5D**. The medians of peptide intensities from the duplicate analyses were highly consistent (6.8E6 *vs.* 6.5E6). The data here clearly indicate that the nanoRPLC-CZE-MS/MS is qualitatively and quantitatively reproducible.

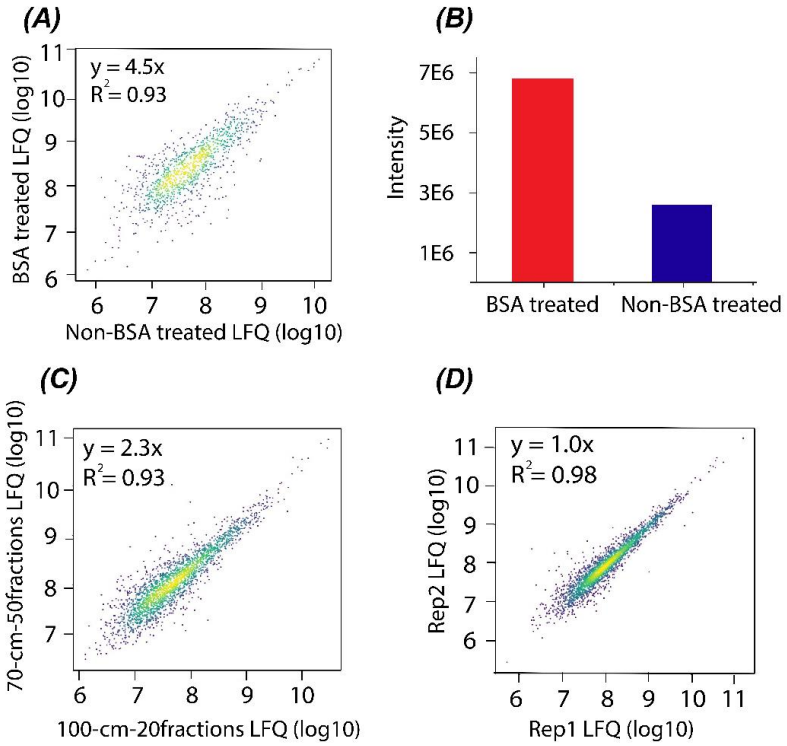

Figure 3. 5 Summary of the 500-ng MCF7 proteome digest data from different experiments regarding peptide intensity and protein LFQ intensity. The data were

from MaxQuant database search. (A) Correlation of protein LFQ intensity (log-log) between the studies using BSA treated and non-BSA treated sample vials. (B) The medians of peptide intensities from the studies using BSA treated and non-BSA treated sample vials. (C) Correlation of protein LFQ intensity (log-log) between the 100-cm-20fractions and 70-cm-50fractions studies. (D) Correlation of protein LFQ intensity (log-log) between two replicates of nanoRPLC-CZE-MS/MS analysis (Rep 1 and Rep 2).

3.3.3 Bottom-up proteomics of 5000 HEK293T cells

To further investigate the capability of our nanoRPLC-CZE-MS/MS system for bottom-up proteomic analysis of mass-limited samples, we collected 5000 HEK293T cells into an Eppendorf tube using flow cytometry and prepared the sample with the SP3 method [29] followed by nanoRPLC-CZE-MS/MS analysis. The SP3 method has been well characterized for preparation of low-µg and even sub-µg of complex proteome samples.[29,42] If we assume each HEK293T cell contains 0.1 ng of proteins,[43] 5000 cells contain 500-ng proteins in total. 16% of the peptides (500 nL out of 3 µL) corresponding to 800 HEK293T cells were first analyzed by single-shot CZE-MS/MS (Q-Exactive HF mass spectrometer). The single-shot analysis identified 1263 proteins and 5090 peptides in 2-h MS time. We then fractionated the rest of the sample using nanoRPLC into 10 fractions. Each fraction was analyzed by CZE-MS/MS using the 70-cm-long capillary in 65 min. We identified 3689 proteins in 11-h MS time using this approach with the consumption of a peptide amount corresponding to only roughly 1000 cells. We plotted the cumulative number of protein IDs versus the number of nanoRPLC fractions, **Figure 3.6**. We identified 2900 proteins from the first five nanoRPLC fractions

using about 5-h MS time with the consumption of 12% of the total peptides corresponding to only 500 cells.

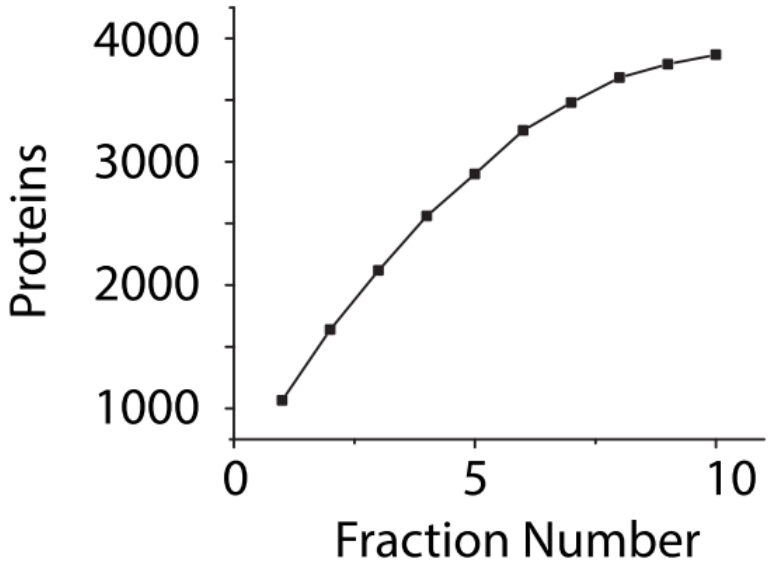

Figure 3. 6 Cumulative protein IDs vs. number of nanoRPLC fractions from the data of 5000 HEK293T cells.

We compared our data with that in the literature about bottom-up proteomic analyses of hundreds to thousands of human cells. Wang *et al.* identified 600 proteins from 5000 MCF-7 breast cancer cells using a highly efficient NP-40-based sample preparation method and nanoRPLC-MS/MS (Q-TOF mass spectrometer) in about 4-h MS time.[44] Wiśniewski *et al.* identified 1536 and 2055 proteins from 1000 and 3000 HeLa cells using the improved filter aided sample preparation method and nanoRPLC-MS/MS (Orbitrap Velos mass spectrometer) in 4-h MS time.[45] Li *et al.* developed a highly sensitive bottom-up proteomic procedure for analysis of small numbers of human cancer cells by employing a single-tube AFA (adaptive focused acoustics)-based sample preparation method and PLOT (porous layer open tube)-nanoRPLC-MS/MS (Q-Exactive mass spectrometer).[46] A sample containing 2000 MCF7 cells was processed

and a peptide aliquot corresponding to 500 cells was analyzed by nanoRPLC-MS/MS in 4-h MS time, leading to the identification of 3370 proteins. Chen *et al.* developed an integrated spintip-based sample preparation method for processing mass-limited proteome samples with high recovery, and 1270 proteins were identified from 2000 HEK 293T cells using nanoRPLC-MS/MS (Orbitrap Fusion mass spectrometer) in 1.4-h MS time.[47] Our SP3-nanoRPLC-CZE-MS/MS system achieved comparable performance compared to that developed by Li *et al.* and Chen *et al.*[46,47] regarding the number of protein IDs, MS time, and the number of human cells. The results clearly indicate that the SP3-nanoRPLC-CZE-MS/MS is an alternative bottom-up proteomic system for deep bottom-up proteomic analysis of mass-limited samples. Because CZE-MS/MS and nanoRPLC-MS/MS are complementary for protein and peptide IDs,[5-9,24,25,27] and because we only used 25% of the peptide sample in each nanoRPLC fraction for CZE-MS/MS analysis, the leftover peptide material in each nanoRPLC fraction can be further analyzed by nanoRPLC-MS/MS to boost the total number of protein and peptide IDs.

We further compared our 5000-cell data with a comprehensive bottom-up proteomic study of the same cell line published by Bekker-Jensen *et al.* in 2017.[48] In that study, 1 mg of HEK 293T cell proteome digest was analyzed by high-pH RPLC fractionation followed by nanoRPLC-MS/MS with the identification of 9246 unique protein-coding genes. Our 5000-cell study identified 3629 protein-coding genes and only 77% of them were also identified in the Bekker-Jensen's work, suggesting good complementarity between these two platforms for protein IDs. We then compared proteins identified in the 5000-cell study to a human transcription factor list that contains over 1600 transcription factors.[49] Our result covered 121 human transcription factors. It

has been estimated that in mammalian cells, transcription factors have copy numbers per cell in the range of 10,000-300,000.[50] In the 5000-cell study, the peptides from the 5000 cells were eventually dissolved in 3 µL of solution before nanoRPLC-CZE-MS/MS analysis. The transcription factor concentration in the 3-µL solution should be in the range of 1-50 pM, indicating high sensitivity of our system.

The proteomic sample preparation of small numbers of human cells has been drastically improved recently. Zhu *et al.* developed nanoPOTS (nanodroplet processing in one pot for trace samples) sample preparation method that enabled efficient preparation of 1-100 human cells for bottom-up proteomic analysis.[51,52] They achieved the identification of 3000 proteins from 140 HeLa cells by coupling the nanoPOTS method with nanoRPLC-MS/MS (Orbitrap Fusion Lumos mass spectrometer).[51] The Zhang group developed novel integrated devices for proteomic preparation of 100 cells and even single cells with high sample recovery.[53,54] Over 300 proteins were identified from single HeLa cells via coupling the sample preparation method to nanoRPLC-MS/MS (Orbitrap Fusion mass spectrometer).[53] We expect that coupling these advanced sample preparation methods with CZE-MS/MS will further improve the proteomic characterization of single human cells because it has been well demonstrated that the advanced CZE-MS/MS outperformed nanoRPLC-MS/MS for mass-limited proteomic sample analysis regarding the number of protein IDs.[7-10]

3.4 Conclusions

In this work, we presented an improved nanoRPLC-CZE-MS/MS system with high peak capacity and sensitivity for bottom-up proteomics of nanograms of human cell proteome samples. The system produced 6500 protein IDs from only 100-ng MCF7

proteome digest and identified 3700 proteins from 1000 HEK 293T cells. We expect that coupling more advanced sample preparation methods with our improved nanoRPLC-CZE-MS/MS will be a powerful tool for comprehensive bottom-up proteomics of small numbers of human cells.

CZE-MS/MS has been recognized as an alternative approach to LC-MS/MS for top-down proteomics,[55,56] metabolomics,[57] and global characterization of glycans[58]. We believe that our improved nanoRPLC-CZE-MS/MS system will also benefit the scientific communities for global delineation of proteoforms, metabolites, and glycans in mass-limited biological samples.

3.5 Acknowledgments

We thank Prof. Xuefei Huang's group and Prof. Jian Hu's group at Michigan State University for kindly providing the MCF7 cells and HEK293T cells for our research. The research was funded by the National Science Foundation (CAREER Award, DBI-1846913). L. Sun was supported by the National Institutes of Health (R01GM125991) and the National Science Foundation (CAREER Award, DBI-1846913).

REFERENCES

REFERENCES

1. Yuan, H.; Jiang, B.; Zhao, B.; Zhang, L.; Zhang, Y., *Analytical chemistry* **2019,** *91* (1), 264-276.

2. Motoyama, A.; Yates, J. R., 3rd, *Analytical chemistry* **2008,** *80* (19), 7187-93.

3. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., 3rd, *Chemical reviews* **2013,** *113* (4), 2343-94.

4. Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J., *Molecular & cellular proteomics : MCP* **2014,** *13* (1), 339-47.

5. 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., *Angewandte Chemie* **2014,** *53* (50), 13931-3.

6. Zhang, Z.; Qu, Y.; Dovichi, N. J., *TrAC Trends in Analytical Chemistry* **2018,** *108*, 23-37.

7. Wang, Y.; Fonslow, B. R.; Wong, C. C.; Nakorchevsky, A.; Yates, J. R., 3rd, *Analytical chemistry* **2012,** *84* (20), 8505-13.

8. Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J., *Analytical chemistry* **2013,** *85* (5), 2569-73.

9. Ludwig, K. R.; Sun, L.; Zhu, G.; Dovichi, N. J.; Hummon, A. B., *Analytical chemistry* **2015,** *87* (19), 9532-7.

10. Lombard-Banek, C.; Moody, S. A.; Nemes, P., *Angewandte Chemie* **2016,** *55* (7), 2454-8.

11. Lewis, K. C.; Opiteck, G. J.; Jorgenson, J. W.; Sheeley, D. M., *Journal of the American Society for Mass Spectrometry* **1997,** *8* (5), 495-500.

12. Bergstrom, S. K.; Samskog, J.; Markides, K. E., *Analytical chemistry* **2003,** *75* (20), 5461-7.

13. Bergstrom, S. K.; Dahlin, A. P.; Ramstrom, M.; Andersson, M.; Markides, K. E.; Bergquist, J., *The Analyst* **2006,** *131* (7), 791-8.

14. Zhang, J.; Hu, H.; Gao, M.; Yang, P.; Zhang, X., *Electrophoresis* **2004,** *25* (14), 2374-83.

15. Chambers, A. G.; Mellors, J. S.; Henley, W. H.; Ramsey, J. M., *Analytical chemistry* **2011,** *83* (3), 842-9.

16. Mellors, J. S.; Black, W. A.; Chambers, A. G.; Starkey, J. A.; Lacher, N. A.; Ramsey, J. M., *Analytical chemistry* **2013,** *85* (8), 4100-6.

17. Jooß, K.; Scholz, N.; Meixner, J.; Neusüß, C., *Electrophoresis* **2019,** *40* (7), 1061-1065.

18. Li, Y.; Champion, M. M.; Sun, L.; Champion, P. A.; Wojcik, R.; Dovichi, N. J., *Analytical chemistry* **2012,** *84* (3), 1617-22.

19. Yan, X.; Sun, L.; Zhu, G.; Cox, O. F.; Dovichi, N. J., *Proteomics* **2016,** *16* (23), 2945-2952.

20. Faserl, K.; Kremser, L.; Muller, M.; Teis, D.; Lindner, H. H., *Analytical chemistry* **2015,** *87* (9), 4633-40.

21. Faserl, K.; Sarg, B.; Sola, L.; Lindner, H. H., *Proteomics* **2017,** *17* (22), doi: 10.1002/pmic.201700310.

22. Faserl, K.; Sarg, B.; Gruber, P.; Lindner, H. H., *Electrophoresis* **2018,** *39* (9-10), 1208-1215.

23. Choi, S. B.; Lombard-Banek, C.; Munoz, L. P.; Manzini, M. C.; Nemes, P., *J Am Soc Mass Spectrom* **2018,** *29* (5), 913-922.

24. Chen, D.; Shen, X.; Sun, L., *Analytica chimica acta* **2018,** *1012*, 1-9.

25. Chen, D.; Ludwig, K. R.; Krokhin, O. V.; Spicer, V.; Yang, Z.; Shen, X.; Hummon, A. B.; Sun, L., *Analytical chemistry* **2019,** *91* (3), 2201-2208.

26. Kulak, N. A.; Geyer, P. E.; Mann, M., *Molecular & cellular proteomics : MCP* **2017,** *16* (4), 694-705.

27. Yang, Z.; Shen, X.; Chen, D.; Sun, L., *Analytical chemistry* **2018,** *90* (17), 10479-10486.

28. Britz-McKibbin, P.; Chen, D. D., *Analytical chemistry* **2000,** *72* (6), 1242-52.

29. Hughes, C. S.; Foehr, S.; Garfield, D. A.; Furlong, E. E.; Steinmetz, L. M.; Krijgsveld, J., *Molecular systems biology* **2014,** *10*, 757.

30. Hughes, C. S.; Moggridge, S.; Muller, T.; Sorensen, P. H.; Morin, G. B.; Krijgsveld, J., *Nature protocols* **2019,** *14* (1), 68-85.

31. McCool, E. N.; Lubeckyj, R.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., *Journal of visualized experiments : JoVE* **2018,** 140, e58644, doi:10.3791/58644.

32. Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J., *Angewandte Chemie* **2013,** *52* (51), 13661-4.

33. Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J., *Rapid communications in mass spectrometry : RCM* **2010,** *24* (17), 2554-60.

34. Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J., *Journal of proteome research* **2015,** *14* (5), 2312-21.

35. Eng, J. K., McCormack, A. L., Yates, J. R., *Journal of The American Society for Mass Spectrometry* **1994**, 5(11), 976-89.

36. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., *Analytical chemistry* **2002,** *74* (20), 5383-92.

37. Elias, J. E.; Gygi, S. P., *Nature methods* **2007,** *4* (3), 207-14.

38. Cox, J.; Mann, M., *Nature Biotechnology* **2008**, 26(12), 1367-72.

39. Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M., *Molecular Cellular Proteomics* **2014**, 13, 2513–2526.

40. Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A., *Nucleic Acids Research* **2019**, 47(D1), D442-D450.

41. Zhang, Z.; Hebert, A. S.; Westphall, M. S.; Qu, Y.; Coon, J. J.; Dovichi, N. J., *Analytical chemistry* **2018,** *90* (20), 12090-12093.

42. Sielaff, M.; Kuharev, J.; Bohn, T.; Hahlbrock, J.; Bopp, T.; Tenzer, S.; Distler, U., *Journal of proteome research* **2017,** *16* (11), 4060-4072.

43. Cohen, D.; Dickerson, J. A.; Whitmore, C. D.; Turner, E. H.; Palcic, M. M.; Hindsgaul, O.; Dovichi, N. J., *Annual review of analytical chemistry* **2008,** *1*, 165-90.

44. Wang, N.; Xu, M.; Wang, P.; Li, L., *Analytical chemistry* **2010,** *82* (6), 2262-71.

45. Wisniewski, J. R.; Ostasiewicz, P.; Mann, M., *Journal of proteome research* **2011,** *10* (7), 3040-9.

46. Li, S.; Plouffe, B. D.; Belov, A. M.; Ray, S.; Wang, X.; Murthy, S. K.; Karger, B. L.; Ivanov, A. R., *Molecular & cellular proteomics : MCP* **2015,** *14* (6), 1672-83.

47. Chen, W.; Wang, S.; Adhikari, S.; Deng, Z.; Wang, L.; Chen, L.; Ke, M.; Yang, P.; Tian, R., *Analytical chemistry* **2016,** *88* (9), 4864-71.

48. Bekker-Jensen, D. B.; Kelstrup, C. D.; Batth, T. S.; Larsen, S. C.; Haldrup, C.; Bramsen, J. B.; Sorensen, K. D.; Hoyer, S.; Orntoft, T. F.; Andersen, C. L.; Nielsen, M. L.; Olsen, J. V., *Cell systems* **2017,** *4* (6), 587-599 e4.

49. Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T., *Cell* **2018,** *172* (4), 650-665.

50. Biggin, M. D., *Developmental cell* **2011,** *21* (4), 611-26.

51. Zhu, Y.; Piehowski, P. D.; Zhao, R.; Chen, J.; Shen, Y.; Moore, R. J.; Shukla, A. K.; Petyuk, V. A.; Campbell-Thompson, M.; Mathews, C. E.; Smith, R. D.; Qian, W. J.; Kelly, R. T., *Nature communications* **2018,** *9* (1), 882.

52. Zhu, Y.; Clair, G.; Chrisler, W. B.; Shen, Y.; Zhao, R.; Shukla, A. K.; Moore, R. J.; Misra, R. S.; Pryhuber, G. S.; Smith, R. D.; Ansong, C.; Kelly, R. T., *Angewandte Chemie* **2018,** *57* (38), 12370-12374.

53. Shao, X.; Wang, X.; Guan, S.; Lin, H.; Yan, G.; Gao, M.; Deng, C.; Zhang, X., *Analytical chemistry* **2018,** *90* (23), 14003-14010.

54. Chen, Q.; Yan, G.; Gao, M.; Zhang, X., *Analytical chemistry* **2015,** *87* (13), 6674-80.

55. Lubeckyj, R. A.; McCool, E. N.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., *Analytical Chemistry* **2017**, *89*(22), 12059-12067.

56. McCool, E. N.; Lubeckyj, R. A.; Shen, X.; Chen, D.; Kou, Q.; Liu, X.; Sun, L., *Analytical Chemistry* **2018**, *90*(9), 5529-5533.

57. Onjiko, R. M.; Moody, S. A.; Nemes, P., *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*(21), 6545-50.

58. Qu, Y.; Sun, L.; Zhang, Z.; Dovichi, N. J., *Analytical Chemistry* **2018**, *90*(2), 1223-1233.

CHAPTER 4. Nanoparticle-aided nanoreactor for large-scale proteomics of few mammalian cells

4.1 Introduction

The advance of separation technique and mass spectrometer instrumentation have achieved comprehensive proteome profiling in bottom-up proteomics. High throughput identification of >10,000 gene products from human cells was achieved in multiple laboratories.[1-5] To reach the “dark corner” of the human proteome, that is proteins with low copy numbers per cell, it usually requires hundreds of micrograms to milligrams of material and extensive fractionation.[6] It was reported that, with a charge state of 5 to 10, single ion detection is possible using orbitrap instrumentation.[7] The state-of-the-art MS platforms have achieved zmol level limit of detection for peptides in complex samples.[8-14] If a single mammalian cell contains 12,000 to 15,000 gene products with a dynamic range spanning from one copy per cell to millions copies per cell, more than 4000 gene products could be identified from a single mammalian cell if protein recoveries from sample preparation and separation are 100%.[9,15] However, only a couple of labs in the world reported identification of hundreds of proteins from a single mammalian cell.[14,16,17] The challenge of single-cell or few-cell proteomics lies significantly in the sample preparation step. A mammalian somatic cell is only 10 to 20 μm in size and contains only hundreds of picogram of proteins in mass.[9,18] A full recovery of the protein material at such trace amounts is extremely challenging with regular sample processing techniques due to significant sample loss caused by dead adsorption of proteins/peptides on surfaces, such as processing containers and pipette tips. Researchers have been

making great efforts in exploring novel sample processing techniques that are suitable for mass-limited proteome samples (e.g., single or small numbers of cells).[14,16,19-25]

The basic idea of the new sample preparation methods for mass-limited samples is to decrease sample processing volume and eliminate sample transfer. The NanoPOTS (nanodroplet processing in one pot for trace samples) method developed by Zhu *et al.* is a nice example.[14] The method performed all the bottom-up sample preparation steps in a nano-well with a total volume of only 200 nL. It has achieved hundreds of to over one thousand protein identifications (IDs) from a single HeLa cell.[16,26] Due to an extreme small volume of sample handling, NanoPOTs required careful operations in a humid chamber. MicroFASP is another example developed by Zhang *et al.*,[25] and it is a modified version of the FASP (filter-aided sample preparation) method.[27] For the microFASP, a miniature filter membrane of 0.1 mm^2 was installed into a 20 µL pipette tip for sample processing, drastically reducing sample loss during sample preparation. The sample processing volume was maintained at a low microliter level. Over 1800 proteins were identified when a sample containing 100 HeLa cells was processed by the microFASP. More importantly, microFASP does not require special instrumentation.

The SP3 (Single-Pot Solid-Phase-enhanced Sample Preparation) is also one of the new sample processing approaches suitable for mass-limited sample.[22,23] All sample preparation steps were performed in a single Eppendorf tube. Under high organic content (>70% acetonitrile (ACN)), proteins were effectively captured on paramagnetic beads through hydrophilic interaction while salts and detergents were effectively removed. After that, proteins on beads were digested by trypsin, followed by MS analysis. The SP3 outperformed FASP in terms of proteome coverage when low µg of protein materials were

processed, and high quantitative reproducibility was also documented for the SP3 method.[28] Over 15,000 unique peptides were identified when a sample containing only 1000 HeLa cells was processed by SP3.[22] However, SP3 is mainly operated in Eppendorf tubes and requires a microliter-level solution for sample processing, limiting its performance for processing low nanograms of complex proteome samples.

Unlike NanoPOTs, microFASP, and SP3 where samples were processed in an open environment, some microreactors with small volumes in fused silica capillaries have been developed to process proteins in a closed environment and are potential alternatives for trace material processing.[29-31] The microreactors in fused silica capillaries can be easily sealed and provide a closed environment for all sample preparation steps.

Inspired by the microreactors and the SP3 method, in this work, we present a new sample processing technique, nanoparticle-aided nanoreactor for nanoproteomics (Nano3), for bottom-up proteomics of mass-limited samples. The Nano3 method employs the same nanoparticles and principle as the SP3 method, but carries out the sample processing in a nanoreactor with a total volume less than 30 nL. Paramagnetic beads were packed into a fused silica capillary to form a nanoreactor for capturing proteins from cells lysed by a lysis buffer containing SDS and ACN. After flushing the nanoreactor with ACN to remove SDS, proteins captured on nanoparticles were digested into peptides by a plug of trypsin solution, followed by peptide collection from the nanoreactor via flushing the reactor with a buffer containing 2% (v/v) ACN and 0.1 % (v/v) formic acid and nanoRPLC-MS/MS analysis. We compared the performance of Nano3 and SP3 for processing 50, 10 and 2 ng of mouse brain proteome samples. Nano3 outperformed SP3 regarding the number of protein IDs and intensity, indicating better overall sample

recovery. We further validated the Nano3 method for processing 1000, 100, and 10 HeLa cells, corresponding to 100, 10, and 1 ng proteins in mass with the assumption of 100 pg proteins per HeLa cell.

4.2 Experimental section

4.2.1 Material and reagent

All reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. LC/MS grade water, acetonitrile (ACN), ethanol, HPLC graded formic acid (FA) were purchased from Fisher Scientific (Pittsburgh, PA). Fused silica capillaries were purchased from Polymicro Technologies (Phoenix, AZ). Carboxylate-modified paramagnetic beads (Sera-Mag SpeedBeads (hydrophilic) CAT # 45152105050250, and Sera-Mag SpeedBeads (Hydrophobic), CAT # 65152105050250) were purchased from GE Healthcare. C18 column packing material (ReproSil-Pur 120Å C18-AQ 1.9μm) was purchased from ESI Source Solution (Woburn, MA). Frit kits were purchased from Next Advance (Troy, NY). A syringe pump that can be operated in withdraw and infuse modes was purchased from KD Scientific (Holliston, MA).

All the capillaries involved in Nano3 and protein transferring were pre-treated with BSA (2 mg/mL) solution to reduce sample loss due to dead adsorption based on our recent study.[32] Briefly, 2 mg/mL BSA solution was injected into the capillary and stored in capillary for 10 min at room temperature. The BSA solution was then flushed out by water. Flushing continued for 30 min. The capillary was then flushed with methanol and air dried before use.

4.2.2 Fabrication of the nanoreactor for the Nano3

A capillary (200 μm i.d., 360 μm o.d., 10 cm long) was installed with a polymer frit first for packing hydrophilic paramagnetic beads. The polymer solution for frit was made

according to the manufactural protocol. Briefly, 15 µL of Kasil-1624 and 5 µL Kasil-1 were mixed. After that, 5 µL of formamide was added into the mixture and vortexed for a few seconds. Frit material was then introduced into the capillary through capillary action. Both ends of the capillary were sealed with a rubber. The capillary was then incubated in 80 °C water bath for overnight. Prior of paramagnetic beads packing, the capillary was rinsed with methanol. Both kinds of paramagnetic beads (hydrophilic and hydrophobic) were mixed with 1:1 ratio, rinsed with water and were resuspended in 70% (v/v) ACN. The beads solution was introduced into the capillary using a syringe and manual pump. The length of the beads in the capillary was controlled to be about 1 millimeter, corresponding to less than 30 nL total volume. The whole capillary was filled with 70% (v/v) ACN before use. The total volume of the capillary containing the nanoreactor was 3-4 µL.

4.2.3 Mouse brain protein preparation

Mouse brains were kindly provided by Dr. Chen Chen's lab in the Department of Animal Science, Michigan State University. The whole protocol related to the mouse brain samples were performed following guidelines defined by the Institutional Animal Care and Use Committee of Michigan State University. The mouse brain tissue was lysed through homogenization in a lysis buffer (2% SDS, 100 mM NH_4HCO_3, pH 8.0, complete protease inhibitors, phosphatase inhibitors). After homogenization, the mouse brain solution was centrifuged at 10,000 g for 10 min. The supernatant was taken out and subjected for BCA assay for protein concentration measurement. 200 µg of mouse brain proteins were reduced with 0.5 µmol dithiothreitol (DTT) at 37 °C for 30 min and then alkylated with 1 µmol Iodoacetamide (IAA) at room temperature for 20 min in dark. 0.5 µmol of DTT was added into the protein solution again to quench IAA. The protein sample was used for

validating the sensitivity of our LC-MS platform and comparing the SP3 and Nano3 methods.

4.2.4 HeLa cell preparation

The cells (originally from ATCC) were cultured at 37°C under 5% CO_2 in Dulbecco's Modified Eagle's Medium, supplemented with 10% fetal bovine serum, 1x glutamine, and 1000 units/ml penicillin/streptomycin (all of reagents above from Thermo Fisher Scientific). After cell culture, cells were harvested, and cell concentration was measured using a hemocytometer. After careful PBS rinsing, cells suspended in PBS were diluted into concentrations of 1000 cells/µL, 100 cells/µL, and 10 cells/µL using the PBS buffer based on the hemocytometer result. To minimize cell lysis in PBS, the entire process was performed within 20 min, and diluted cell samples were immediately processed by the Nano3 method.

Another batch of cultured HeLa cells (2×10^6 cells) were lysed in 200-µL lysis buffer (2% SDS, 100 mM NH_4HCO_3, pH 8.0, complete protease inhibitor, and phosphatase inhibitor) with ultrasonication for 10 min on ice and 95 °C for 5 min. The lysate was centrifuged at 14,000 g for 5 min. The supernatant was collected and subjected for BCA assay for protein concentration measurement. The sample was also used to validate the Nano3 method.

4.2.5 Mouse brain sample processing using the SP3 and Nano3 methods

For the SP3 method, 20 µL of each of the paramagnetic beads stock solution (50 µg/µL) was combined and was rinsed with water for a few times. 20-µL water was then used to resuspend the paramagnetic beads. Reduced and alkylated mouse brain proteins in 2% SDS lysis buffer was diluted with 70% (v/v) ACN into 50 ng/µL, 10 ng/µL and 2

ng/µL. 1 µL of protein solution from each dilution was combined with 1 µL of beads solution. Then 5 µL of ACN was added into the tube so ACN concentration was higher than 70% (v/v) in the SP3 system. The mixture was incubated for over 18 min for thorough protein binding. A magnet was placed under the tube for 2 min to separate beads from solution. Supernatant was carefully removed with the magnet on and beads were rinsed with 10 µL of 70% Ethanol for two times and 10 µL 100% ACN for one time. 8 µL of 100 mM NH_4HCO_3 (pH 8) was then added into the tube to resuspend the beads. 1 µL of trypsin solution in 100 mM NH_4HCO_3 (pH 8) with different concentrations (10 ng/µL, 2.5 ng/µL and 1 ng/µL) was added to 50-ng, 10-ng and 2-ng protein samples, respectively, for protein digestion (37 °C, overnight). After digestion, the supernatant (about 6 µL) was directly deposited into a nanoLC insert tube for nanoRPLC-MS/MS analysis. The experiment was performed twice at each sample amount.

For the Nano3 method, the basic workflow for sample processing is shown in **Figure 4.1**. Sample transfer was performed through an empty capillary (100 µm i.d. × 12.5 cm length, 1 µL total volume). For the mouse brain samples, the proteins were loaded into the nanoreactor capillary not the intact cells. Reduced and alkylated mouse brain protein solution in the 2% SDS lysis buffer was diluted with 70% ACN into 50 ng/µL, 10 ng/µL and 2 ng/µL. Each protein solution was pushed into the 1-µL transfer capillary and the protein solutions in the transfer capillary containing 50-ng, 10-ng, and 2-ng proteins were pushed onto the nanoreactor through a syringe filled with 70% ACN for the Nano3 sample processing. After protein loading, beads were continuously rinsed with 70% ACN for additional 8 µL to remove the SDS. 1 µL of trypsin solution (in 100 mM NH_4HCO_3) with the same concentration settings as the SP3 method for different amounts of mouse

brain proteins was pushed into the nanoreactor through the transfer capillary to cover the nanoreactor. Both ends of the nanoreactor capillary were sealed with a rubber and the nanoreactor was incubated in a 37 °C water bath overnight for tryptic digestion. After digestion, the nanoreactor was rinsed with 5-6 µL of a buffer containing 2% ACN and 0.1% formic acid to elute the peptides into a nanoLC insert tube for nanoRPLC-MS/MS analysis. The experiment was performed twice at each sample amount.

4.2.6 Sample processing of few HeLa cells with the Nano3 method

The HeLa cell processing with the Nano3 method was similar to the mouse brain protein processing with some modifications, **Figure 4.1**. The nanoreactor capillary was first filled with a buffer containing 2% SDS and 80% ACN. The syringe used for cell loading onto the nanoreactor was also filled with the same buffer. The cell solution was first transferred into the 1-µL transfer capillary to control the number of cells for processing. For the cell solutions with concentrations of 1000 cells/µL, 100 cells/µL and 10 cells/µL, the number of cells in the transfer capillary was approximately 1000 cells, 100 cells, and 10 cells. The transfer capillary was connected to the nanoreactor capillary at one end and to the syringe filled with a buffer containing 2% SDS and 80% ACN at the other end. The 1-µL cell solution surrounded by the buffer containing SDS and ACN was pushed into the nanoreactor capillary. After that, both ends of the nanoreactor capillary was sealed with a rubber and the capillary was sonicated for 10 min, followed by incubation in a 95 °C water bath for 10 min for cell lysis and protein denaturation. Then the protein solution was pushed onto the nanoreactor for protein capturing and the nanoreactor was flushed with at least 10-µL 70% ACN to remove SDS. The rest of the sample processing steps were the same as the mouse brain protein processing. Trypsin amount for digestion was 10 ng,

5 ng and 2 ng for 1000 cells, 100 cells and 10 cells, respectively. The experiment was performed with replicates for each cell amount and we repeated the whole experiment one more time starting from cell culture. The sample insert tube for nanoRPLC-MS/MS analysis was pretreated with the BSA solution as described in our previous work.[32]

We also examined the number of HeLa cells in 1 µL of 10 cells/µL cell solution under a microscope by depositing 1 µL of the cell solution onto a glass slide. The cell count varied from 3 cells to 20 cells in 1 µL of the solution across 6 different examinations with a median of 10 cells.

We also performed a blank experiment to confirm that the protein IDs from HeLa cells were not from contaminates in the cell culture medium. Before PBS rinsing, harvested cells were centrifuged. After gentle centrifugation, the cells were pelleted, and a small aliquot of the supernatant was collected. The supernatant aliquot was treated as cells for processing with the Nano3 method. The steps for processing the blank sample are the same as the 10-1000 cell samples.

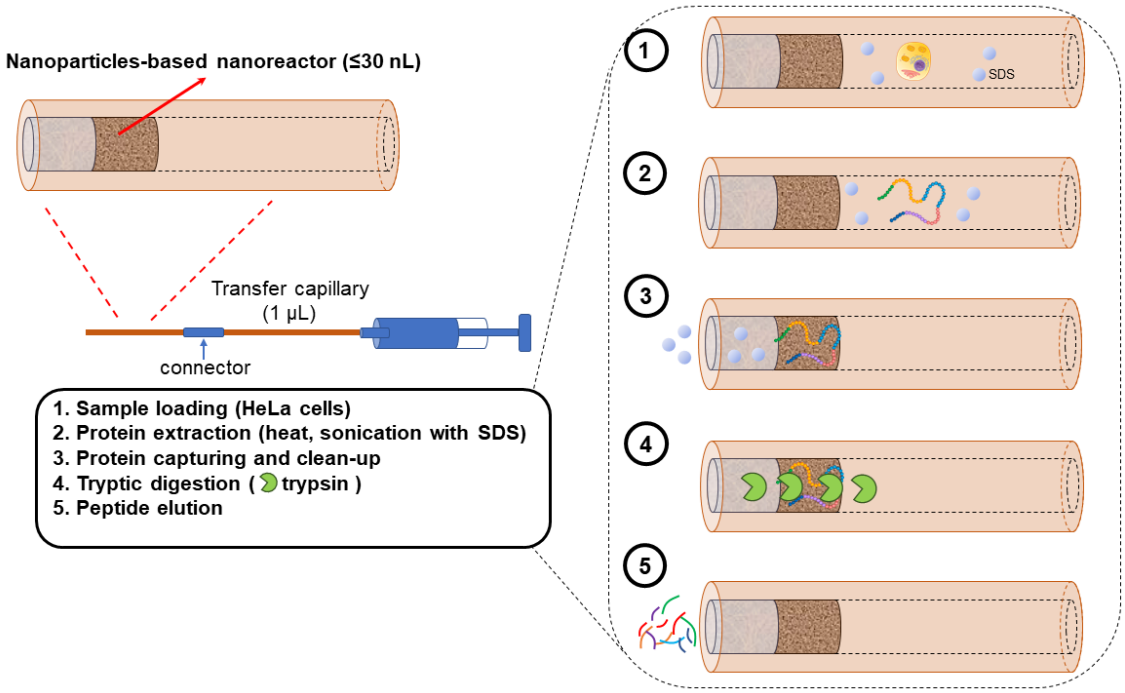

Figure 4. 1 Schematic of the general workflow of sample processing with the Nano3 method.

4.2.7 Direct sampling of mass-limited HeLa cell lysates in the Eppendorf tube for processing with the Nano3 method

A HeLa cell lysate was serially diluted to a concentration of 1 ng/µL with a buffer containing 2% SDS and 70% ACN. 1 µL of diluted protein solution containing only 1 ng of protein in mass was deposited into a 0.6 mL low retention Eppendorf tube for sample processing with the Nano3 method. A capillary (200 µm i.d.) containing a 100-µm-long nanoreactor was used. The capillary was first filled with 70% ACN. Because the capillary only contained a very small amount of paramagnetic beads, the back pressure was low enough for withdrawing the HeLa protein sample in the Eppendorf tube directly using a syringe pump operated in the withdraw mode. The protein solution was continuously drawn into the nanoreactor capillary until no solution was left in the tube. The protein

solution in the nanoreactor capillary was then pushed through the nanoreactor for protein loading with a buffer containing 70% ACN and was further processed using a similar procedure to the mouse brain samples. 1 µL of 70% ACN was added into the Eppendorf tube for protein rinsing after the protein solution was completely drawn. The 1 µL of rinsing solution was also loaded onto the nanoreactor for sample processing. Triplicate sample preparations were performed. For digestion, 2 ng of trypsin was loaded as 10-Hela cell analysis.

4.2.8 NanoRPLC-MS/MS

NanoRPLC separation was performed on an EASY nanoLC-1200 (Thermo Fisher Scientific) system with one column set-up. Two kinds of self-packed C18 capillary columns were used: 50 µm i.d. × 50 cm length and 75 µm i.d. × 50 cm length (ReproSil-Pur 120Å C18-AQ 1.9 µm). Buffer A containing 0.1% (v/v) FA and 2% (v/v) ACN and buffer B containing 80% (v/v) ACN and 0.1% (v/v) FA were used for gradient separation. Each sample was loaded onto the RPLC column with buffer A. Then the peptides retained on the column were separated by a linear gradient. For the 75-µm-i.d. column, the flow rate was set to 150 nL/min and for the 50-µm-i.d. column, the flow rate was set to 90 nL/min or 80 nL/min (depending on how high the back pressure was). The peptide samples were separated with a 75-min gradient or a 105-min gradient. >80% of peptide sample solution was loaded for each analysis.

A Q-Exactive HF (Thermo Fisher Scientific) mass spectrometer was used for all MS analysis in a data dependent acquisition (DDA) mode. The full scan range was 400-1200 *m/z*. The MS resolution was set as 60,000 (at *m/z* 200). The AGC target was set as 3e6 with a maximum injection time of 50 ms. The tandem MS resolution was also set at

60,000. The AGC target was set as 1e5 with a maximum injection time of 200 ms. The loop count was set as 10 (Top 10 DDA). Isolation window for MS/MS was set as 4 *m/z* and a normalized collision energy was 28. The intensity threshold was 1e4 for triggering MS/MS and the dynamic exclusion was 30 s.

4.2.9 Data analysis

All MS raw files were processed with MaxQuant 1.5.5.1.[33] Mouse Brain sample was searched against the database of UniPort Mus Musculus proteome (UP000000589) and human sample was searched against database of UniPort Homo Sapiens proteome (UP000005640). All the parameters were set default except for HeLa cell samples, the fixed modification carbamidomethyl on cysteine residues was turned off because we didn't perform reduction and alkylation for the HeLa cell samples. The match between runs (MBR) function was used.[34] The false discovery rates (FDRs) were controlled to be lower than 1% at the peptide and protein group levels.

4.3 Results and discussion

4.3.1 Comparisons of the SP3 and Nano3 methods for processing low-nanograms of a complex proteome sample

Before we compared the SP3 and Nano3 methods for mass-limited sample processing for bottom-up proteomics, we first evaluated the sensitivity of our nanoRPLC-MS/MS (Q-Exactive HF mass spectrometer) platform for the analysis of trace amounts of a complex proteome digest. A 100-µg aliquot of reduced and alkylated mouse brain protein sample was digested with trypsin following the SP3 procedure.[22,23] The digest was used for the system evaluation. A self-packed nanoRPLC capillary column (75 µm i.d. × 50 cm long, ReproSil-Pur, 120Å, C18-AQ, 1.9 µm beads) was used for peptide separation.

Different amounts of mouse brain peptides ranging from 0.2 ng to 50 ng were loaded onto the nanoRPLC-MS/MS system for analysis in triplicate and each LC-MS run used a 75-min gradient. As shown in the **Figure 4.2A**, by MS/MS only, fewer than 30 proteins were identified when only 0.2-ng protein digest was analyzed. That is about the amount of proteins from 1-2 single mammalian cells with a size of 10-20 μm.[18] The number of identified proteins increased to 627 when 2-ng of peptides were loaded. The numbers of protein IDs from 10-ng and 50-ng peptides were comparable (1313 *vs.* 1505). By performing the database search of all the raw files from 0.2-50 ng peptides together with the MaxQuant software and turning on the matching-between-runs function, the number of protein IDs from 0.2-ng peptides was boosted to 190, which is over 6-folds higher than that from MS/MS only. The number of protein IDs from the 2-ng peptide sample was also improved by nearly 80% compared to the data of MS/MS alone (1106 *vs.* 627). The data suggests that the match-between-runs function is extremely useful for mass-limited samples. Also, the nanoRPLC-MS/MS system had nice reproducibility for analysis of the 0.2-50-ng peptide samples regarding the number of protein IDs from triplicate analyses.

After evaluating the sensitivity of our nanoRPLC-MS/MS system, we compared the performance of the SP3 and Nano3 methods for processing 2-50 ng of mouse brain proteins, **Figure 4.2B**. The Nano3 method clearly outperformed the SP3 method for processing 2-50-ng mouse brain proteins regarding the number of protein IDs. For example, the Nano3 method identified 40-times and 6-times higher number of proteins based on MS/MS than the SP3 method starting from 2-ng (206 *vs.* 5 proteins) and 10-ng proteins (367 *vs.* 62), respectively. If we consider the protein IDs from both MS/MS and matching-between-runs, the Nano3 method still produced about 260% and 170% more

protein IDs than the SP3 method from the 2-ng and 10-ng protein samples. For the 50-ng protein sample, the Nano3 generated 40% more protein IDs than the SP3 (963 *vs.* 679).

The drastically better overall performance of the Nano3 method compared to the SP3 method is most likely due to much higher sample recovery from the Nano3 approach, demonstrated by the substantially higher protein intensity produced by the Nano3 method, **Figure 4.2C**. The much better protein recovery is due to the substantially smaller sample processing volume of the Nano3 approach, leading to less sample loss because of dead adsorption to surfaces and higher protein concentration for more efficient tryptic digestion. The SP3 method is a good option for preparation of sub-micrograms of proteome samples and it requires at least a 10-µL solution for processing the samples to make sure that the paramagnetic beads are freely suspended.[23] In our experiment, the sample processing volume using the SP3 method was about 10 µL. The protein concentration in the 10-µL solution was 0.2-5 ng/µL for the 2-50-ng protein samples. For the Nano3 method, the proteins were digested in the nanoreactor, which had a smaller than 30-nL volume. The protein concentration in the nanoreactor was over two orders of magnitude higher than that during the SP3 processing. Minimized sample processing volume is a key point for preparation of mass-limited proteome samples for large-scale proteome profiling.[35]

We need to highlight several advantages of the Nano3 method. First, like the SP3 method, the Nano3 method is compatible with various detergents and chaotropic reagents because proteins can be captured by the hydrophilic nanoparticles with high efficiency under a high ACN concentration environment and detergents and chaotropic reagents can be removed via flushing with ACN. Second, the total volume of nanoreactor

is only 30 nL in the experiment. The extremely small volume not only reduces protein loss during processing due to dead adsorption onto surfaces, but it also affords a relative high protein concentration for digestion, ensuring high enzymatic digestion efficiency. We need to note that the volume of the nanoreactor can be further reduced easily via packing lower amounts of nanoparticles in the capillary or using capillaries with a smaller inner diameter. We expect a nanoreactor with a smaller volume will further boost the performance of the Nano3 method for processing mass-limited samples. Third, the Nano3 technique employs a relatively closed environment for sample processing, making the protein and peptide storage easy before MS analysis and Nano3 does not require special instrumentation.

The Nano3 method was quantitatively reproducible for processing low ng of mouse brain proteome samples regarding intensity of quantified proteins from two replicates, **Figure 4.2D**. For the 10-ng and 50-ng samples, the linear correlation coefficients of protein intensity were 0.97 and 0.99, respectively. For the 2-ng sample, a reasonable correlation coefficient of 0.85 was still achieved.

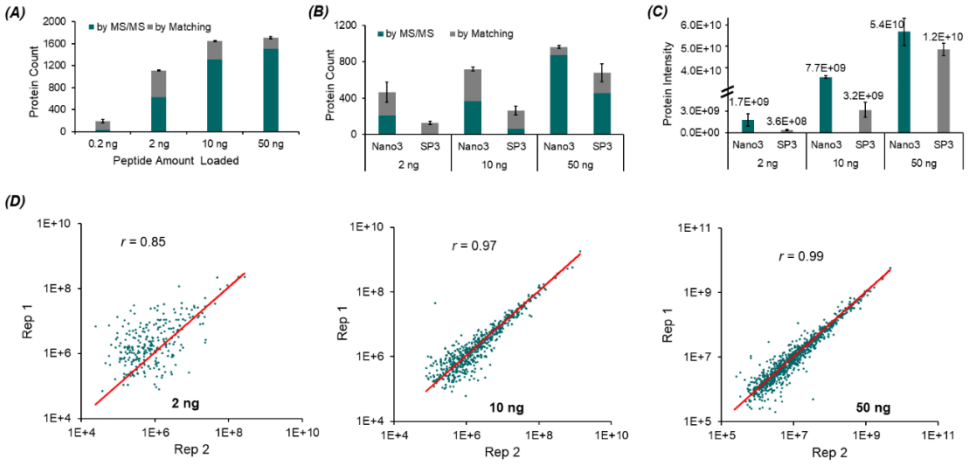

Figure 4. 2 Summary of the data of low-nanograms of mouse brain samples. (A) The number of protein ID from nanoRPLC-MS/MS as a function of the loaded peptide

amount. The proteins identified by MS/MS and matching (match between runs) are labelled separately. The error bars represent the standard deviations of the number of protein ID from triplicate analyses. (B) The number of protein ID from 2-50 ng of mouse brain protein samples processed by the Nano3 and SP3 methods. The error bars represent the standard deviations of the number of protein ID (MS/MS+matching) from duplicate sample preparation. (C) Total protein intensity from the 2-50 ng of mouse brain protein samples processed by the Nano3 and SP3 methods. The error bars represent the standard deviations of the total protein intensity from duplicate sample preparation. (D) Correlations of protein intensity between duplicate preparations of 2, 10 and 50 ng of mouse brain protein samples using the Nano3 method.

We further examined the peptide length and missed cleavage from the Nano3 method and compared them with those from the regular in-solution digestion and SP3 method. We selected point of 10-ng for further analysis. Protein to trypsin ration of Nano3, SP3 and in-solution digestion is 4:1, 4:1 and 30:1, respectively. For in-solution digestion, a bulk amount of mouse brain protein (200 µg) was processed and 10 ng peptides were analyzed by LC-MS. For Nano3 and SP3, 10 ng of mouse brain protein was processed, and the digested peptides were analyzed by LC-MS. As shown in **Figure 4.3,** high ratio of trypsin on nanograms of protein shows very similar result of peptide length distribution **(A)** and missed cleavage **(B)** compared to regular in-solution digestion, indicating efficient trypsin digestion with limited missed cleavage that is comparable to regular in-solution digestion on bulk protein material. We noticed that even though we controlled same amount of trypsin introduced into the Nano3 system as SP3, the actual amount of the trypsin surrounded the paramagnetic beads was hard to estimate. However, the similar

missed cleavage and peptide length percentage distribution across all three conditions indicated non-compromised digestion in Nano3 system. Further optimization regarding trypsin amount and digestion time may lead better outcome of Nano3 processing.

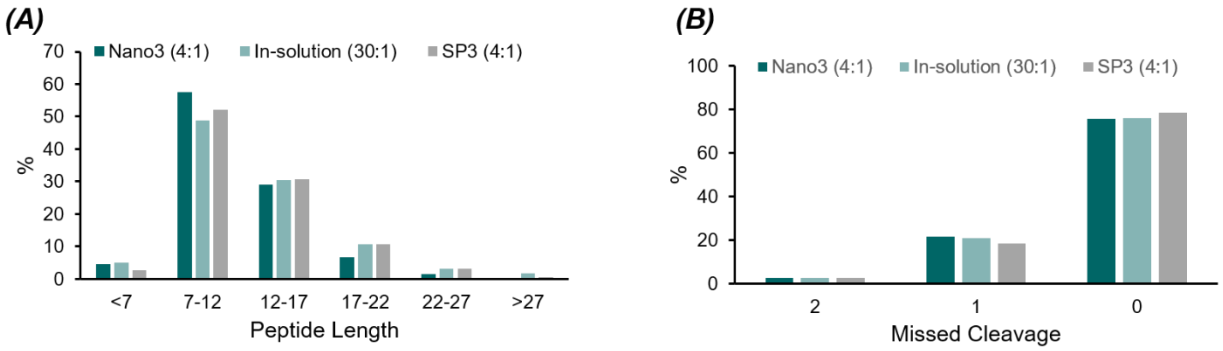

Figure 4. 3 comparison of digestion between Nano3, SP3 and regular in-solution method. (A) % distribution of peptide length of the Nano3, SP3 and in-solution method. (B) % distribution of missed cleavage of the Nano3, SP3 and in-solution method. The protein to trypsin ratio is indicated in the parenthesis.

Proteomics analysis of mass-limited samples requires both a highly efficient sample preparation method and a highly sensitive LC-MS/MS platform for peptide measurements. Our RPLC-MS/MS system with a 75-µm-i.d. capillary column only identified 26 ± 4 (N=3) protein IDs based on MS/MS when analyzing 0.2-ng mouse brain peptides, **Figure 4.2A**. To further improve the sensitivity of the platform, we tested another capillary column with a smaller inner diameter of 50 µm and a length of 50 cm. A lower flow rate of 90 nL/min or 80 nL/min was employed for the 50-µm-i.d. column compared to the 75-µm-i.d. column. We increased the length of the gradient to 105 min because of the delay of the chromatography caused by the low flow rate. We analyzed 0.2 ng and 2 ng of the mouse brain digest using the new RPLC-MS/MS system with the 50-µm-i.d. column. As shown in **Figure 4.4A**, we identified 224 ± 24 (N=2) proteins based

on MS/MS from the 0.2-ng sample, which is about 10-times higher than that from the 75-µm-i.d. column. Considering match-between-runs, 668 ± 41 (N=2) were identified from the 0.2-ng sample. The new RPLC-MS/MS system showed a drastically better sensitivity for protein ID from mass-limited samples, and it was used for the following experiments. **Figure 4.4B** shows an example chromatogram of the 2-ng mouse brain peptide sample after analyzed by the new nanoRPLC-MS/MS system. The system achieved strong peptide signals from the tiny amount of peptide sample and the peak capacity was estimated to be about 500 based on FWHM.

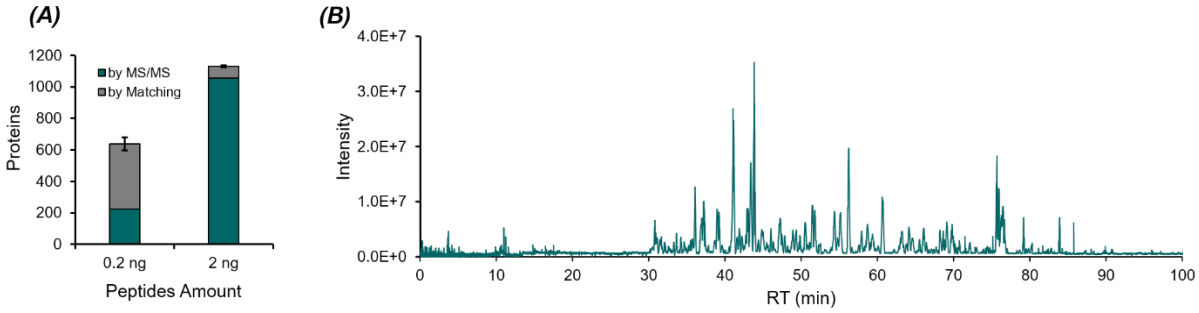

Figure 4. 4 Identification result from 2 ng and 0.2 ng of mouse brain peptides using optimized LC-MS platform. (A) number of proteins identified from 0.2 ng and 2 ng mouse brain peptides (N=2). The error bars represent the standard deviations of the number of protein ID (MS/MS+matching) from duplicate sample preparation. (B) chromatography of LC-MS result from 2-ng mouse brain peptides. Peak capacity was estimated over 500 by FWHM.

4.3.2 Application of the Nano3 method in processing 10-1000 HeLa cells

After concentration determination by hemocytometer method, the harvested HeLa cells were rinsed with a PBS buffer. The HeLa cell solution was then serially diluted with the PBS buffer into three different concentrations (1000 cells/µL, 100 cells/µL, 10 cells/µL).

Then 1 µL of the HeLa cell solutions with different concentrations were transferred into the nanoreactor capillary for sample processing with the Nano3 method, **Figure 4.1**. The approximate number of cells processed by the Nano3 method was 1000, 100 and 10 for the 1000 cells/µL, 100 cells/µL, and 10 cells/µL samples. We performed two replicates of the sample processing and repeated the whole experiment once starting from the cell culture. As shown in **Figure 4.5A,** we achieved 1816 ± 45 (N=5 for two batches) and 1109 ± 188 (N=5) protein IDs by MS/MS from the 1000 cells and 100 cells, respectively. For the 10-cells samples, we identified 451 ± 283 (N=4) proteins by MS/MS. Considering the match-between-runs feature, 1052 ± 317 proteins were identified in total from the 10 HeLa cells. We also processed one blank sample and only 16 proteins were identified.

We further evaluated the quantitative reproducibility of the workflow for analyzing few HeLa cells with the protein intensity data, **Figure 4.5B**. Analyses of 100 and 1000 HeLa cells produced reasonably good reproducibility regarding the protein intensity with linear correlation coefficients of 0.94 and 0.97 and analyses of 10 HeLa cells showed significant variations of protein intensity across replicate. We also determined higher variations of the number of protein IDs from the 10 cells compared to the 100 and 1000 cell samples, **Figure 4.5A**. We speculate that the higher variations of protein ID and intensity from the 10 cell samples are due to the significant variations of the number of cells in the 10-cell samples. We injected 1 µL of 10 cells/µL solution into the nanoreactor for processing and we assumed that the cells uniformly distributed in the solution, corresponding to 10 cells injected for sample preparation. To mimic the actual numbers of cells injected, we examined the cell count under a microscope by depositing 1 µL of 10 cells/µL solution

onto a glass slide. The cell count varied from 3 cells to 20 cells across 6 different examinations with a median of about 10 cells.

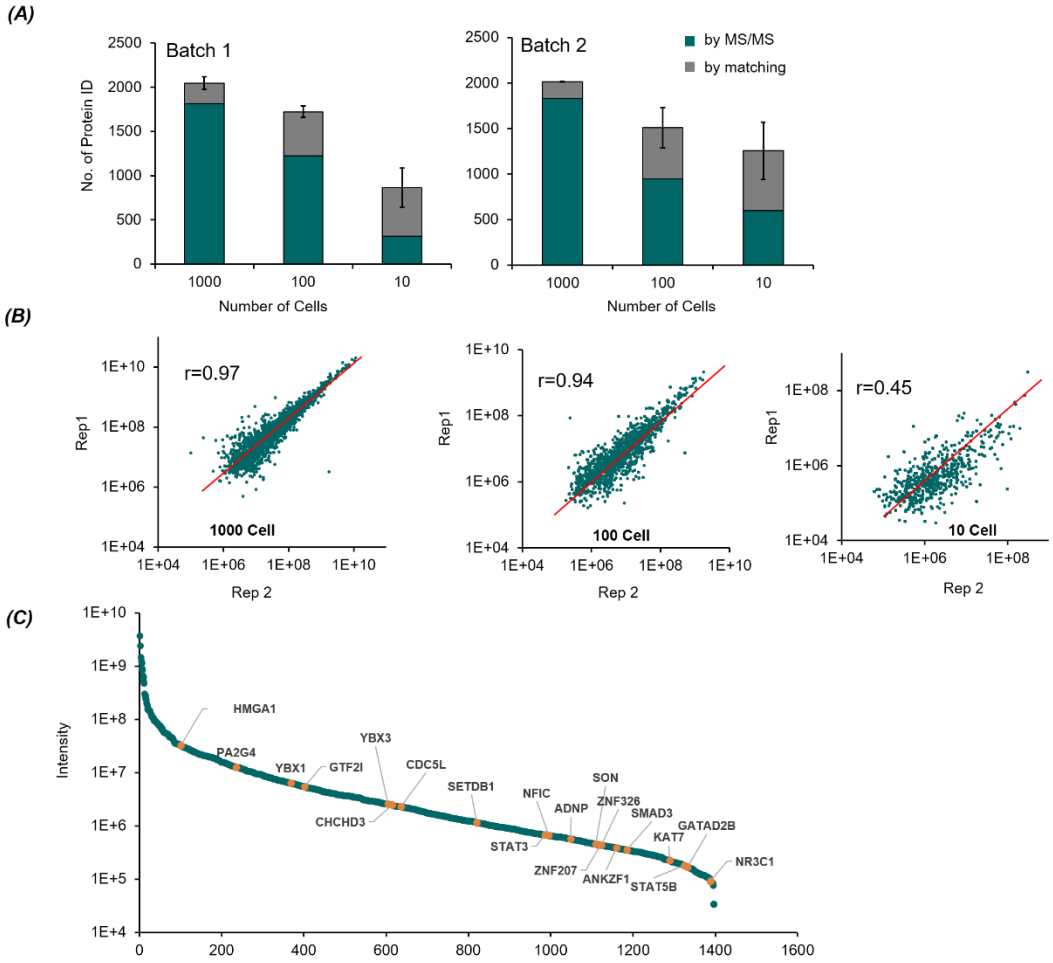

Figure 4. 5 Summary of the data of 10-1000 HeLa cells processed by the Nano3 method. (A) The number of identified proteins from 10, 100, 1000 HeLa cells from two batches of cultured cells. The proteins identified by MS/MS and matching (match between runs) are labelled separately. The error bars represent the standard deviations of the number of protein ID (MS/MS+matching) from duplicate sample preparation. (B) Protein intensity correlations between duplicate preparations of 1000, 100 and 10 HeLa cells using the Nano3 method. The data was from the Batch 1. (C) Protein Intensity plot of identified proteins from one analysis of the 10-cells sample showing the protein

intensity dynamic range. The highlighted ones are 20 transcription factors identified in the run.

We need to highlight that the intensity of identified proteins from the 10-cell sample spanned across five orders of magnitude and 20 transcription factors were confidently identified from only 10 HeLa cells, **Figure 4.5C**. The identified transcription factors are labeled in the figure. The transcription factors were determined through comparing the identified proteins from the 10-cell sample with a transcription factor database reported in the literature containing over 1600 transcription factors.[36] It has been estimated that in mammalian cell, transcription factors have a copy number of 10,000 to 300,000 per cell.[37] Peptides from 10 HeLa cells in our study were dissolved in about 5 µL prior of LC-MS analysis. From there we estimated the concentration of transcription factors in the 5-µL solution was in the range of 33 fM to 1 pM, indicating a high sensitivity of the overall workflow.

We need to point out that the experiments of 10-1000 cells were carried out by employing a 1-µL transfer capillary for controlling the number of cells injected for the Nano3 processing. This approach is appropriate to evaluate the performance of the Nano3 method for preparation of few mammalian cells. However, it may be not straightforward for processing mass-limited samples from laser capture microdissection (LCM) of tissues and cells isolated by fluorescence activated cell sorting (FACS) using the Nano3, because under those situations the samples are usually transferred into Eppendorf tubes or wells. To demonstrate the potential of the Nano3 method for processing trace protein samples placed in Eppendorf tubes, we put a 1-µL aliquot of a HeLa cell lysate containing 1-ng proteins in a 0.6 mL low retention Eppendorf tube and

processed the sample with the Nano3 method with the assistance of a syringe pump operated under a withdraw mode. The 1-ng HeLa cell proteins were dissolved in a buffer containing 2% SDS and 70% ACN.

In this experiment, we used a nanoreactor with a total length of only 100 µm with a total volume of 3 nL to reduce the backpressure of the nanoreactor for withdrawing the solution into the nanoreactor capillary directly using a syringe pump. After the 1-ng HeLa cell proteins were withdrawn into the nanoreactor capillary, the sample was processed using the similar approach as the mouse brain sample. After nanoRPLC-MS/MS analyses, 692±182 proteins (N=3) were identified from the 1-ng HeLa cell protein sample based on MS/MS. The data suggest the feasibility of coupling the Nano3 method with the LCM or FACS for analysis of trace amounts of proteome samples.

4.4 Conclusions

In this work, we developed a novel sample preparation method (Nano3) for processing trace complex proteome samples with high efficacy for large-scale bottom-up proteomics. The Nano3 method employed the basic concept of minimizing the sample processing volume. Proteins extracted from cells in a lysis buffer containing high-concentration SDS were concentrated and digested into peptides in the nanoreactor (≤30 nL total volume). The sample processing was performed in a relatively closed environment, facilitating the sample storage before LC-MS analysis. The Nano3 method identified 40-times higher number of proteins based on MS/MS than the SP3 method starting from 2-ng mouse brain proteins, most likely due to its more than 100 times smaller sample processing volume than the SP3 method, reducing sample loss and improving the tryptic digestion. Over 1000 proteins including 20 transcription factors were identified

from only 10 HeLa cells processed by the Nano3 method, demonstrating the potential of the Nano3 method for advancing large-scale bottom-up proteomics of few mammalian cells.

We expect the number of protein ID from few mammalian cells (*e.g.,* 10 HeLa cells) processed by the Nano3 method can be boosted obviously through several improvement. First, in this proof-of-principle study, we did not systematically optimize the Nano3 method for processing few cells. We expect the sample recovery can be improved significantly after optimizing the volume of the nanoreactor, the trypsin concentration for digestion, the tryptic digestion time, and the procedure for peptide elution from the nanoreactor. Second, the number of protein ID can be increased through employing a liquid-phase separation -MS/MS system with much better sensitivity than the system used in the current study. For example, a nanoRPLC-MS/MS system with a 20-μm-i.d. or 30-μm-i.d. RPLC column and one of the most advanced mass spectrometers (*e.g.,* Orbitrap Fusion Lumos) will be certainly helpful for pursuing a better proteome coverage from trace samples.[14,16] Capillary zone electrophoresis (CZE)-MS/MS could be a useful alternative for analyzing the trace samples processed by the Nano3 method because it outperformed nanoRPLC-MS/MS for the characterization of mass-limited samples regarding the number of protein ID and it has shown low zmole even ymole level limit of detections for peptides.[8, 38-42] We believe that the optimized Nano3 method coupled with an advanced liquid-phase separation-MS/MS system will be a useful tool for large-scale proteome profiling of few even single mammalian cells.

4.5 Acknowledgements

We thank the support from the National Institute of General Medical Sciences (NIGMS) through Grant R01GM125991 and the National Science Foundation through Grant DBI1846913 (CAREER Award) for the research project.

REFERENCES

REFERENCES

1. Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M., *Molecular systems biology* **2011,** *7*, 548.

2. Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M., *Molecular & cellular proteomics : MCP* **2012,** *11* (3), M111 014050.

3. Bekker-Jensen, D. B.; Kelstrup, C. D.; Batth, T. S.; Larsen, S. C.; Haldrup, C.; Bramsen, J. B.; Sorensen, K. D.; Hoyer, S.; Orntoft, T. F.; Andersen, C. L.; Nielsen, M. L.; Olsen, J. V., *Cell systems* **2017,** *4* (6), 587-599 e4.

4. Bache, N.; Geyer, P. E.; Bekker-Jensen, D. B.; Hoerning, O.; Falkenby, L.; Treit, P. V.; Doll, S.; Paron, I.; Muller, J. B.; Meier, F.; Olsen, J. V.; Vorm, O.; Mann, M., *Molecular & cellular proteomics : MCP* **2018,** *17* (11), 2284-2296.

5. Orre, L. M.; Vesterlund, M.; Pan, Y.; Arslan, T.; Zhu, Y.; Fernandez Woodbridge, A.; Frings, O.; Fredlund, E.; Lehtio, J., *Molecular cell* **2019,** *73* (1), 166-182 e7.

6. Zubarev, R. A., *Proteomics* **2013,** *13* (5), 723-726.

7. Zubarev, R. A.; Makarov, A., *Analytical chemistry* **2013,** *85* (11), 5288-5296.

8. Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J., *Angewandte Chemie* **2013,** *52* (51), 13661-13664.

9. Zhang, P.; Gaffrey, M. J.; Zhu, Y.; Chrisler, W. B.; Fillmore, T. L.; Yi, L.; Nicora, C. D.; Zhang, T.; Wu, H.; Jacobs, J.; Tang, K.; Kagan, J.; Srivastava, S.; Rodland, K. D.; Qian, W. J.; Smith, R. D.; Liu, T.; Wiley, H. S.; Shi, T., *Analytical chemistry* **2019,** *91* (2), 1441-1451.

10. Shen, Y.; Tolic, N.; Masselon, C.; Pasa-Tolic, L.; Camp, D. G., 2nd; Hixson, K. K.; Zhao, R.; Anderson, G. A.; Smith, R. D., *Analytical chemistry* **2004,** *76* (1), 144-154.

11. Li, S.; Plouffe, B. D.; Belov, A. M.; Ray, S.; Wang, X.; Murthy, S. K.; Karger, B. L.; Ivanov, A. R., *Molecular & cellular proteomics : MCP* **2015,** *14* (6), 1672-1683.

12. Smith, R. D.; Shen, Y.; Tang, K., *Accounts of chemical research* **2004,** *37* (4), 269-278.

13. Sun, X.; Kelly, R. T.; Tang, K.; Smith, R. D., *The Analyst* **2010,** *135* (9), 2296-302.

14. Zhu, Y.; Piehowski, P. D.; Zhao, R.; Chen, J.; Shen, Y.; Moore, R. J.; Shukla, A. K.; Petyuk, V. A.; Campbell-Thompson, M.; Mathews, C. E.; Smith, R. D.; Qian, W. J.; Kelly, R. T., *Nature communications* **2018,** *9* (1), 882.

15. Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M., *Nature* **2011,** *473* (7347), 337-342.

16. Cong, Y.; Liang, Y.; Motamedchaboki, K.; Huguet, R.; Truong, T.; Zhao, R.; Shen, Y.; Lopez-Ferrer, D.; Zhu, Y.; Kelly, R. T., *Analytical chemistry* **2020,** *92* (3), 2665-2671.

17. Brunner, A. D.; Thielert, M.; Vasilopoulou, C.; Ammar, C.; Coscia, F.; Mund, A.; Horning, O. B.; Bache, N.; Apalategui, A.; Lubeck, M.; Raether, O.; Park, M. A.; Richter, S.; Fischer, D. S.; Theis, F. J.; Meier, F.; Mann, M., *bioRxiv* **2020**, 2020.12.22.423933.

18. Cohen, D.; Dickerson, J. A.; Whitmore, C. D.; Turner, E. H.; Palcic, M. M.; Hindsgaul, O.; Dovichi, N. J., *Annu. Rev. Analytical chemistry* **2008**, *1*, 165– 190.

19. Li, Z. Y.; Huang, M.; Wang, X. K.; Zhu, Y.; Li, J. S.; Wong, C. C. L.; Fang, Q., *Analytical chemistry* **2018,** *90* (8), 5430-5438.

20. Shao, X.; Wang, X.; Guan, S.; Lin, H.; Yan, G.; Gao, M.; Deng, C.; Zhang, X., *Analytical chemistry* **2018,** *90* (23), 14003-14010.

21. Budnik, B.; Levy, E.; Harmange, G.; Slavov, N., *Genome biology* **2018,** *19* (1), 161.

22. Hughes, C. S.; Foehr, S.; Garfield, D. A.; Furlong, E. E.; Steinmetz, L. M.; Krijgsveld, J., *Molecular systems biology* **2014,** *10*, 757.

23. Hughes, C. S.; Moggridge, S.; Muller, T.; Sorensen, P. H.; Morin, G. B.; Krijgsveld, J., *Nature protocols* **2019,** *14* (1), 68-85.

24. Williams, S. M.; Liyu, A. V.; Tsai, C. F.; Moore, R. J.; Orton, D. J.; Chrisler, W. B.; Gaffrey, M. J.; Liu, T.; Smith, R. D.; Kelly, R. T.; Pasa-Tolic, L.; Zhu, Y., *Analytical chemistry* **2020,** *92* (15), 10588-10596.

25. Zhang, Z.; Dubiak, K. M.; Huber, P. W.; Dovichi, N. J., *Analytical chemistry* **2020,** *92* (7), 5554-5560.

26. Cong, Y.; Motamedchaboki, K.; Misal, S. A.; Liang, Y.; Guise, A. J.; Truong, T.; Huguet, R.; Plowey, E. D.; Zhu, Y.; Lopez-Ferrer, D.; Kelly, R. T., *Chemical science* **2021,***12*, 1001-1006.

27. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., *Nature methods* **2009,** *6* (5), 359-362.

28. Sielaff, M.; Kuharev, J.; Bohn, T.; Hahlbrock, J.; Bopp, T.; Tenzer, S.; Distler, U., *Journal of proteome research* **2017,** *16* (11), 4060-4072.

29. Tian, R.; Wang, S.; Elisma, F.; Li, L.; Zhou, H.; Wang, L.; Figeys, D., *Molecular & cellular proteomics : MCP* **2011,** *10* (2), M110 000679.

30. Zhao, Q.; Liang, Y.; Yuan, H.; Sui, Z.; Wu, Q.; Liang, Z.; Zhang, L.; Zhang, Y., *Analytical chemistry* **2013**, *85* (18), 8507-8512.

31. Wang, F.; Wei, X.; Zhou, H.; Liu, J.; Figeys, D.; Zou, H., *Proteomics* **2012,** *12* (21), 3129-3137.

32. Yang, Z.; Shen, X.; Chen, D.; Sun, L., *Journal of proteome research* **2019,** *18* (11), 4046-4054.

33. Cox, J.; Mann M., *Nat Biotechnol*. **2008,** *26* (12), 1367-1372.

34. Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M., *Molecular & cellular proteomics : MCP* **2014,** *13* (9), 2513-2526.

35. Yang, Z.; Sun, L., *Analytical methods* **2021,** *13* (10), 1214-1225.

36. Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T., *Cell* **2018,** *172* (4), 650-665.

37. Biggin, M. D., *Developmental cell* **2011,** *21* (4), 611-626.

38. Zhu, G.; Sun, L.; Yan, X.; Dovichi, N. J., *Analytical chemistry* **2013,** *85* (5), 2569-2573.

39. Ludwig, K. R.; Sun, L.; Zhu, G.; Dovichi, N. J.; Hummon, A. B., *Analytical chemistry* **2015,** *87* (19), 9532-9537.

40. Wang, Y.; Fonslow, B. R.; Wong, C. C.; Nakorchevsky, A.; Yates, J. R. 3rd., *Analytical chemistry* **2012,** *84* (20), 8505-8513.

41. Lombard-Banek, C.; Moody, S. A.; Nemes, P., *Angew Chem Int Ed.* **2016,** *55* (7), 2454-2458.

42. Amenson-Lamar, E.A.; Sun, L.; Zhang, Z.; Bohn, P. W.; Dovichi, N. J., *Talanta* **2019,** 204, 70-73.

[4]CHAPTER 5. Towards a universal sample preparation method for denaturing top-down proteomics of complex proteomes

5.1 Introduction

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

[4] Part of this chapter was adapted with permission from: Yang, Z.; Shen, X.; Chen, D.; Sun, L., Journal of proteome research 2020, 19 (8), 3315–3325.

developed a 2D-RPLC system for high-capacity proteoform separation, and identified 2778 proteoforms from HeLa cell lysates.[12] The Paša-Tolić group developed a high-capacity RPLC system for proteoform separation via using an 80-cm long RPLC column, enabling 1665 proteoform IDs from bacteria with an Orbitrap mass spectrometer.[13] Recently, our group employed a 1.5-meters long capillary for CZE separation of proteoforms and coupling the CZE separation to an Orbitrap mass spectrometer enabled the identification and quantification of thousands of proteoforms from zebrafish brain samples using hundreds of nanograms of protein materials.[14]

The development of large-scale dTDP underlines the importance of a standardized and universal sample preparation method to achieve comprehensive extraction of proteins from biological samples with high recovery, good reproducibility, minimum bias and absence of MS incompatible salts, chaotropes and detergents. Protein extraction using a cell lysis buffer containing chaotropic agents or detergents, and protein sample cleanup before MS with ultrafiltration or precipitation have been suggested as efficient approaches for preparation of protein samples for MS.[22] Sodium dodecyl sulfate (SDS) is an extremely efficient detergent for solubilizing and denaturing proteins, making it widely used in proteomics studies for protein extraction.[23] However, higher than 0.01% (w/v) SDS can be detrimental to chromatography separation and suppress the ESI.[24] Highly efficient depletion of SDS before MS analysis is critical. Multiple methods have been evaluated for SDS removal for bottom-up proteomics and/or dTDP, including membrane ultrafiltration,[25] chloroform-methanol precipitation (CMP),[26] and single-spot solid-phase sample preparation using magnetic beads (SP3).[27,28]

Membrane ultrafiltration (MU) has been widely used by the bottom-up proteomics community for the filter-aided sample preparation (FASP) method to remove SDS before enzymatic digestion of proteins.[25] Basically, a protein sample in 1-5% (w/v) SDS solution is loaded onto a commercialized membrane filter unit with a 10-30-kDa molecular weight cut off (MWCO), followed by washing with a 8 M urea solution to remove SDS, which is based on the fact that 8 M urea can destroy the hydrophobic interaction between SDS and proteins. The MU has also been routinely deployed for buffer exchange for TDP sample preparation.[22] CMP is a well-recognized method for removing SDS from proteins in the dTDP workflow, and the Kelleher group has utilized the CMP for cleaning the protein samples after GELFrEE fractionation in their large-scale dTDP works.[5,6,8] Briefly, a protein sample dissolved in a SDS solution is mixed with methanol, chloroform, and water. After centrifugation, three phases form and the proteins precipitate at the interphase. After removing the upper phase, more methanol is added and the purified protein pellet is obtained after centrifugation. SP3 has been suggested as an efficient sample preparation method for bottom-up proteomics and various detergents can be removed from proteins using the SP3 method.[27,28] Recently, the Webb group evaluated the SP3 method for preparing intact protein samples for dTDP, demonstrating the great potential of the SP3 method as a universal sample preparation method for both bottom-up proteomics and dTDP.[29] For SP3, a protein sample in a SDS buffer was mixed with magnetic beads and acetonitrile (ACN). Under a high concentration of ACN, proteins are adsorbed on the beads. Then the beads are washed with organic solvents (i.e., ethanol and ACN) to clean up the proteins, followed by on-bead digestion for bottom-up proteomics[27,28] or recovering proteins from beads with cold 80% (v/v) formic acid for dTDP.[29]

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

5.2 Experiment

5.2.1 Materials and Reagents

All reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. Urea was purchased from Alfa Aesar (Haverhill, MA). LC/MS grade water, methanol, chloroform, HPLC grade acetic acid (AA), formic acid (FA), and hydrofluoric acid (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 purchased from Polymicro Technologies (Phoenix, AZ). Carboxylate-modified paramagnetic beads (Sera-Mag SpeedBeads (hydrophilic) CAT # 45152105050250, and Sera-Mag SpeedBeads (Hydrophobic), CAT # 65152105050250) were purchased from GE Healthcare. Centrifugal filter units with a 30-kDa molecular weight cutoff for ultrafiltration were purchased from Millipore (MRCF0R030). CZE insert tubes (CAT # C4010-630P) were purchased from Thermo Scientific. The polyacrylamide gel for SDS-

PAGE was purchased from Bio-rad (CAT # 4561094). The gel staining buffer was purchased from Bio-rad (CAT # 1610803).

5.2.2 Protein Extraction from Escherichia coli and HepG2 cells

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

5.2.3 Protein sample cleanup with various methods before MS analysis

5.2.3.1 SP3 method

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

5.2.3.2 CMP method

The CMP procedure was processed based on the literature.[26] Briefly, 400 µL methanol, 100 µL chloroform and 300 µL water were added into 100-µg *E. coli* cell lysate (1 µg/µL, 1% (w/v) SDS) successively. Every addition of reagent was followed by a thorough vortex. The mixture was then centrifuged at 14 000 g for 1 min. Solution separated into three layers after centrifugation. The top aqueous layer was carefully removed without disturbing the protein flake. 400 µL of methanol was then added into the solution followed by a thorough vortex. The mixture was then centrifuged at 20 000 g for

5 min. Supernatant was removed. The protein pellet was suspended in a 50-µL buffer containing 100 mM NH_4HCO_3 (pH 8) with or without 1% (w/v) SDS with gentle pipetting. We also vortexed and sonicated the sample solution gently for a short period of time to improve the protein recovery. After centrifugation, the protein solution was analyzed by the BCA assay to determine the protein concentration.

5.2.3.3 MU method

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

5.2.4 SDS-PAGE and CZE-MS/MS analysis

The *E. coli* and HepG2 cell lysates before and after cleanup with the three methods were analyzed by SDS-PAGE according to the procedure in the literature.[30] The gel was first rinsed with D.I. water for 5 min for 3 times. Coomassie blue staining buffer was used

for overnight staining with gentle swing. The destaining was processed with D.I. water rinsing.

A 100-cm linear polyacrylamide (LPA)-coated capillary (50/360 μm i.d./o.d.) was used for CZE separation. The LPA coating was prepared following the published procedure.[31,32] One end of the capillary was etched with HF following the published procedure to reduce the outer diameter of the capillary to 70-80 μm.[33] The commercialized electrokinetically pumped sheath flow CE-MS interface (EMASS II, CMP scientific, Brooklyn, NY) was used to couple CZE to MS.[34,35] The automated CZE operations were implemented with an ECE-001 autosampler (CMP scientific). The sheath buffer contained 10% (v/v) methanol and 0.2% (v/v) FA. The sample buffer was 100 mM NH_4HCO_3 (pH 8) and the background electrolyte (BGE) was 20% or 40% (v/v) acetic acid in water. The glass emitter for the electrospray was pulled from borosilicate glass capillary (0.75 mm i.d., 1 mm o.d.) by a Sutter P-1000 flaming/brown micropipette puller. The orifice of the emitter was controlled at 20-40 μm. The distance of the etched capillary tip to the emitter orifice was less than 300 μm and the distance of the emitter orifice to the MS entrance was around 2 mm. The sample was loaded with 5 psi for 90 s so approximately 500 nL of the sample was loaded into the capillary. The capillary sample injection end was immersed in a CE vial containing 100 mM NH_4HCO_3 (pH 8) for 10 s to neutralize the leftover acetic acid on the outer surface of the capillary injection end before it was moved into the sample vial for sample injection. After sample loading, the capillary sample injection end was moved into a BGE vial and the CZE separation was carried out by applying a +30-kV voltage for 115 min. A 15-psi pressure was applied afterwards for 5

min to flush and condition the capillary. A +2-kV voltage was applied in the sheath buffer vial for ESI.

A Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) was used for all CZE-MS/MS analyses. A data-dependent acquisition (DDA) method was employed. The full MS scan range was 600–2000 *m/z*. Mass resolution was 120,000 (at *m/z* 200) and AGC was 1E6. Maximum injection time was 50 ms. The top 3 most abundant ions in a MS spectrum were isolated in the quadrupole with a 4-m/z isolation window sequentially and fragmented with higher-energy collisional dissociation (HCD) with the normalized collision energy 20. The mass resolution for MS/MS was 60,000 (at *m/z* 200). The maximum injection time was 200 ms and the AGC was 1E5. The ion intensity threshold was 2E4 for triggering MS/MS. The dynamic exclusion was turned on and set at 30 s. Charge exclusion was enabled and ions with charges from +1 to +3 as well as ions with unassigned charge states were excluded from MS/MS.

5.2.5 Data analysis

The TopPIC (Top-down mass spectrometry based proteoform identification and characterization) software was applied for proteoform IDs via database search for all *E. coli* and HepG2 data.[19] Briefly, the RAW files were converted into mzML files using the msconvert tool.[36] The mzML files was then processed by the TopFD (Top-down mass spectrometry feature detection) tool for spectral deconvolution. The resulted msalign files was then processed by TopPIC (v1.3.1) for database searching. UniProt databases of *E. coli* (UP000000625) and Human (UP000005640) were used for search. For the database search, the maximum number of mass shift was 1. All other parameters were kept as default. The target-decoy approach was employed to evaluate the false discovery rate

(FDR) of proteoform spectrum match (PrSM) and proteoform IDs.[37,38] The database search results were filtered with a 1% PrSM-level FDR and a 5% proteoform-level FDR. The MS raw data have been deposited to the ProteomeXchange Consortium via the PRIDE[39] partner repository with the data set identifier PXD018248.

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

5.3 Results and discussion

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

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

performance of the three methods. To make the sample preparation method compatible with follow-up dynamic pH junction-based CZE-MS/MS analysis,[41] 100 mM NH_4HCO_3 (pH 8.0) was used to redissolve the proteins after removing SDS with the three methods.

For the SP3 method, we first tested the loading capacity of magnetic beads by incubating 100 µg of *E. coli* proteins with three different amounts of magnetic beads, 10 µg, 100 µg and 500 µg. The protein recovery based on the BCA assay was about 60% and had no obvious difference among the three different bead amounts, **Figure 5.1A**. We also measured the amount of proteins that were not bound to the magnetic beads at the first step with the BCA assay, **Figure 5.1B**. The unbound protein amount was about 5 µg, indicating that the magnetic beads captured proteins with high efficiency. Considering the recovered proteins (~60 µg) and unbound proteins (~ 5 µg), we noted that about 35% of the loaded proteins were lost somewhere during the SP3 process. We speculated that those proteins were still adsorbed on the magnetic beads and were not eluted by the 100 mM NH_4HCO_3 (pH 8.0) buffer. We further analyzed the proteins prepared by the SP3 method with the three different bead amounts using SDS-PAGE, **Figure 5.2A**. The three *E. coli* protein samples after the SP3 cleanup show no significant difference regarding the molecular weight (MW) distributions. The results indicate that 10 µg of magnetic beads are good enough to prepare 100-µg proteins from a complex proteome, which agrees well with the data in the literature.[27,28] We utilized 10-µg beads for all the following SP3 experiments.

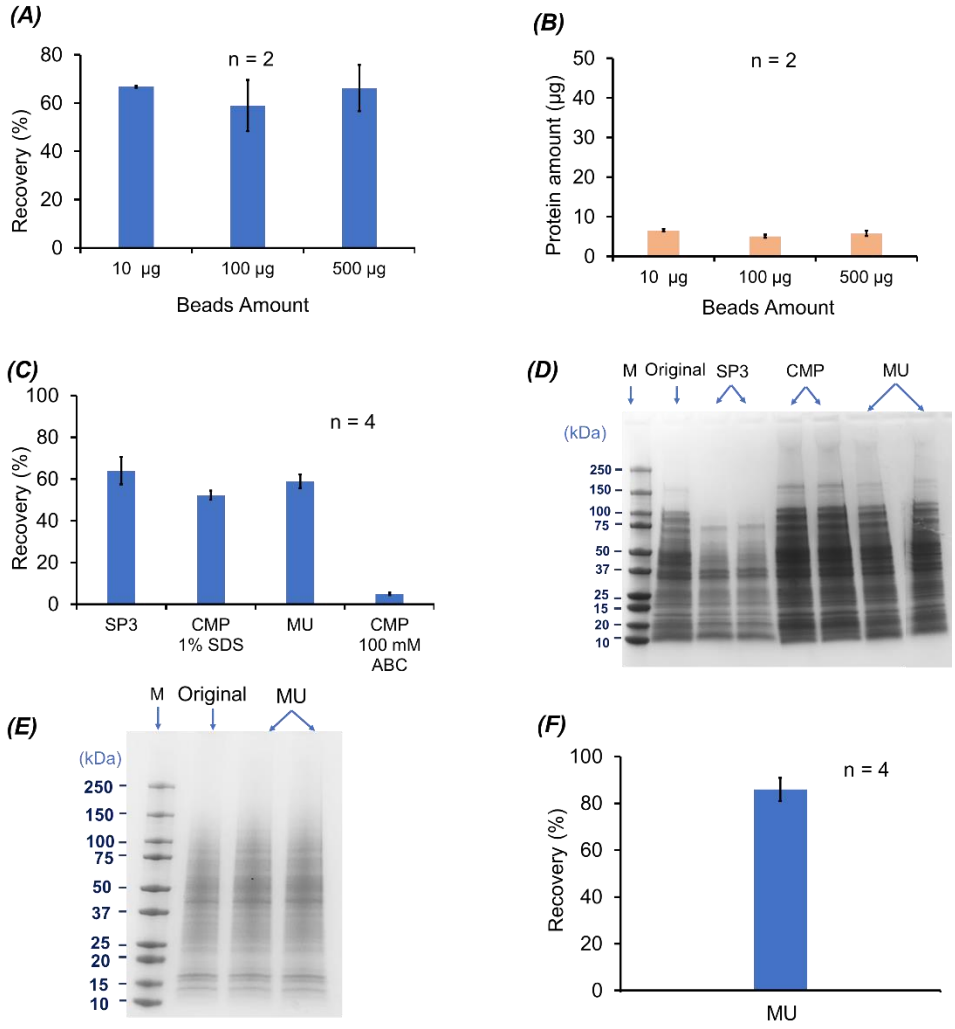

Figure 5. 1 BCA and SDS-PAGE results on the E. coli cell proteins (A-D) and HepG2 cell proteins (E and F) when different SDS removal methods were applied.

(A) Protein recovery (%) of the SP3 method for 100-µg E. coli proteins when different amounts of magnetic beads were used (n=2). (B) Amounts of unbound proteins to magnetic beads as a function of the magnetic bead amount (n=2). (C) Protein recovery (%) of the SP3, CMP and MU methods. The protein pellets from the CMP method were dissolved in 100 mM NH4HCO3 (ABC is short for ammonium bicarbonate) (pH 8) with or without 1% (w/v) SDS (n=4). (D) SDS-PAGE data of the recovered E. coli proteins using the SP3, CMP and MU methods (n=2) as well as the E. coli cell lysate in 1% (w/v)

SDS before sample cleanup (Original). For the CMP method, the protein pellet dissolved in 100 mM NH4HCO3 (pH 8) with 1% (w/v) SDS was used for the analysis. For each sample, an aliquot of 10-μg proteins was loaded for SDS-PAGE. (E) SDS-PAGE data of the HepG2 cell protein samples before (Original) and after sample cleanup with the MU method (n=2). For each sample, an aliquot of 6-μg proteins was loaded for SDS-PAGE. (F) Protein recovery data of the HepG2 cell samples after the MU method-based sample cleanup (n=4). The error bars in the figures represent the standard deviations of protein recovery or protein amount.

We also noted that the SP3 method-based sample cleanup introduced an obvious bias against large proteins (higher than 50 kDa) compared to the sample before cleanup, **Figure 5.2A**. The bias was also observed in the HepG2 human cell lysate processed by the SP3 method, **Figure 5.2B**. We used 100 mM NH_4HCO_3 (pH 8.0) to extract the proteins from the beads in order to make the method compatible with follow-up CZE-MS/MS analysis, which might lead to relatively low efficiency of redissolving large proteins, because it has been suggested that a buffer containing detergents is essential for completely extracting proteins bound to beads in SP3.[27-29]

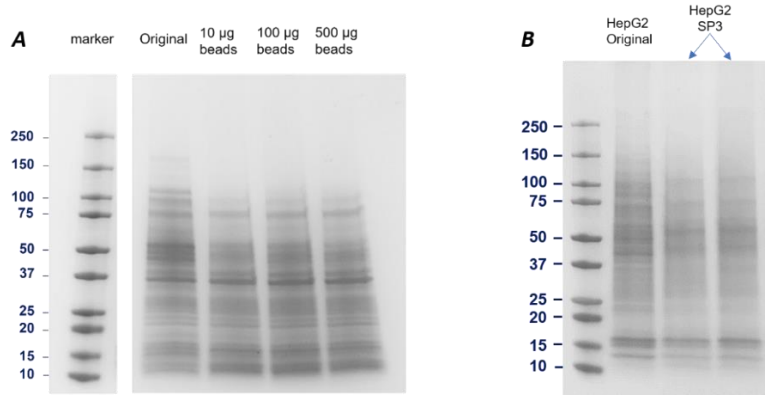

Figure 5. 2 SDS-PAGE analysis of SP3 processed proteins. (A) SDS-PAGE analysis of E. coli proteins processed by the SP3 method with three different amounts of beads.

The original sample is the cell lysate in 1% (w/v) SDS without sample cleanup. For all the four E. coli samples, 6-µg E. coli protein was loaded for analysis. (B) SDS-PAGE analysis of HepG2 proteins processed by the SP3 method (n=2). The original sample is the cell lysate in 1% (w/v) SDS without sample cleanup. 7 µg of total proteins were loaded on each lane.

We then employed the MU, CMP, and SP3 methods for preparing aliquots of the *E. coli* cell lysate dissolved in 1% (w/v) SDS. Each aliquot contained 100-µg proteins, and four aliquots were prepared by each method. The MU and SP3 methods generated much higher protein recovery than the CMP method (~60% *vs.* 5%) with good reproducibility (RSD<12%) when a solution containing 100 mM NH_4HCO_3 (pH 8.0) was used to redissolve the protein pellet from CMP, **Figure 5.1C**. We noted that the protein pellet from CMP was hard to be dissolved in the NH_4HCO_3 buffer, which resulted in a low protein recovery. We further tried to use a buffer containing 1% (w/v) SDS and 100 mM NH_4HCO_3 (pH 8.0) to redissolve the protein pellet and obtained a 50% protein recovery with high precision (RSD, 4%). We then analyzed the *E. coli* cell lysates before and after cleanup using the three methods by SDS-PAGE, **Figure 5.1D**. For the CMP method, we used the protein sample redissolved in the 1% (w/v) SDS solution for SDS-PAGE. Two batches of prepared samples with the three methods were analyzed. The MU and CMP method show comparable protein MW distributions, which are similar to the original *E. coli* sample without cleanup. As we discussed before, the SP3 method had trouble recovering large proteins with the 100 mM NH_4HCO_3 (pH 8.0) buffer. All the three methods show good reproducibility regarding the SDS-PAGE data. Based on the discussed protein recovery, protein bias, and compatibility with the CZE-MS/MS analysis of the three sample cleanup

methods, the MU method outperformed the CMP and SP3 methods. We further employed the MU method for preparation of the HepG2 cell lysate in 1% (w/v) SDS. The SDS-PAGE and BCA assay data clearly show that the MU method can achieve reproducible preparation of the human cell lysate with high protein recovery and precision (86±5%), **Figure 5.1E** and **5.1F**. All the results demonstrate that the MU method could be a universal method for sample preparation in dTDP of complex proteomes. We obtained a higher protein recovery for the human cell lysate than the *E. coli* cell lysate (86% *vs.* 60%) using the MU method, presumably due to the fact that *E. coli* proteins tend to be smaller than human proteins in the length range of 1-250 amino acids based on the data in Swiss-Prot database, **Figure 5.3**, resulting in a higher chance for protein flow-through the membrane (30-kDa MWCO) for the *E. coli* sample.

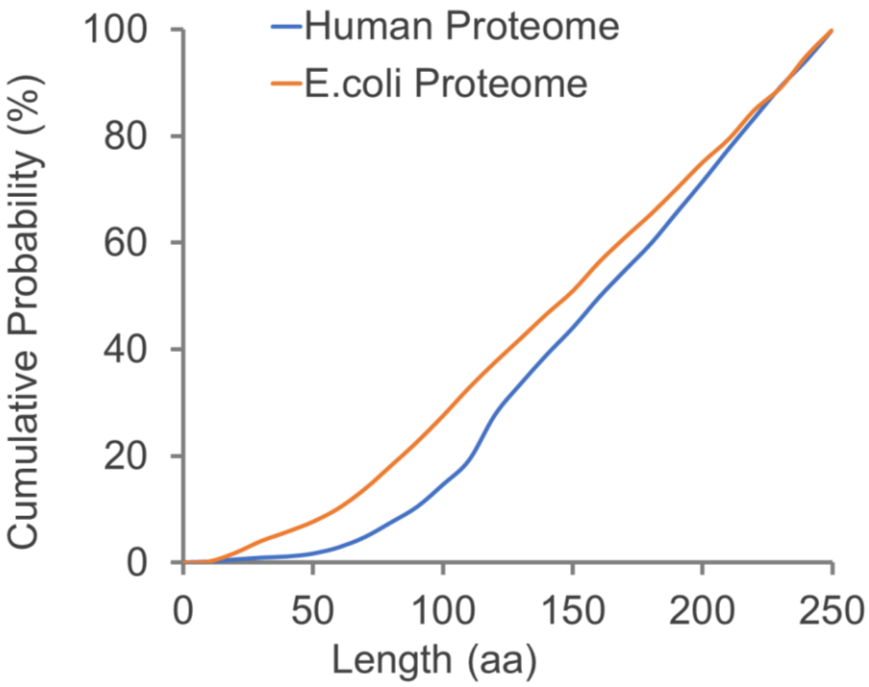

Figure 5. 3 Cumulative distribution of the length of E. coli proteins and human proteins in the Swiss-Prot database in a length range of 1-250 amino acids (aa).

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

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

We further coupled the SDS-based protein extraction and the MU-based protein sample cleanup to our dynamic pH junction-based CZE-MS/MS for dTDP of *E. coli* cells. About 500 nL of the *E. coli* protein solution in 100 mM NH_4HCO_3 (pH 8.0) after cleanup was injected into the CZE capillary for analysis. The injected protein amount was roughly 400 ng. The BGE of CZE was 20% (v/v) acetic acid. We performed CZE-MS/MS analysis of two batches of the *E. coli* sample prepared by the MU method. **Figure 5.4A** shows the base peak electropherograms of the two *E. coli* samples and **Figure 5.4B** shows the numbers of identified proteins, proteoforms, and PrSMs. Single-shot CZE-MS/MS identified 832±65 proteoforms (n=2) with a 5% proteoform-level FDR. When we used a 1% proteoform-level FDR, 821± 67 (n = 2) proteoforms corresponding to 219± 21 proteins were identified in a single CZE-MS/MS run, **Figure 5.4B**. On average, about 20 fragment ions were matched to each identified proteoform, **Figure 5.4C**, suggesting the high confidence of the proteoform identifications. We noted that mass of identified proteoforms ranged from 1 kDa to 25 kDa and over 70% of the identified proteoforms had mass smaller

than 6 kDa. We also analyzed the GO information of the identified proteins, **Figure 5.4D**, and about 30% of the proteins were membrane proteins. We finally analyzed the hydrophobicity of the identified proteoforms and compared it with our previous work, in which 8M urea was used for protein extraction from *E. coli* cells.[41] As shown in **Figure 5.4E**, the *E. coli* proteoforms identified in this work show higher hydrophobicity than the ones identified in our previous work, most likely due to the fact that SDS has stronger solubility for hydrophobic proteins than 8M urea. We also noted that compared to the protein samples extracted with 8M urea,[41] the samples from the 1% (w/v) SDS extraction required a higher acetic acid concentration in the BGE of CZE (20% *vs.* 5% (v/v) acetic acid) to achieve reproducible CZE separations, which might be due to the higher hydrophobicity of proteoforms from the 1% (w/v) SDS extraction. We need to point out that when high concentration of acetic acid (*i.e.,* 20%) is used as the BGE for CZE separation, the sample dissolved in the NH_4HCO_3 buffer in a sample vial could be acidified by the BGE during the sample injection process, which will influence the dynamic pH junction sample stacking obviously. When the sample volume is small (*i.e.,* <5 µL), the issue becomes severe. Immersing the sample injection end of the capillary in a 100 mM NH_4HCO_3 buffer for seconds before moving it into the sample vial for sample injection can eliminate the issue based on our experience.

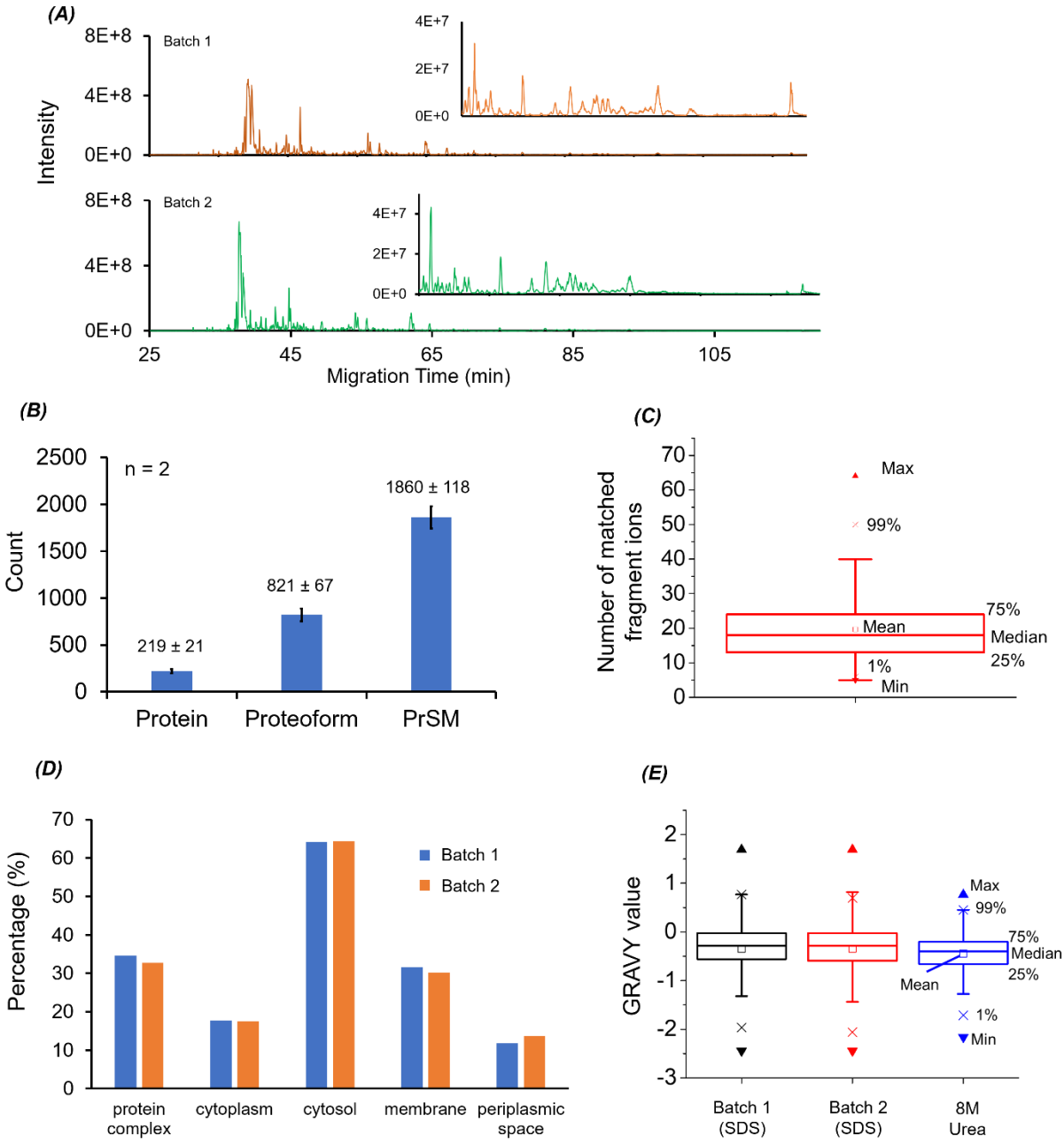

Figure 5. 4 CZE-MS/MS data of E. coli samples prepared with the MU method. (A) Base peak electropherograms of two batches of prepared E. coli protein samples after CZE-MS/MS analysis. (B) Numbers of protein, proteoform, and PrSM identifications from the two CZE-MS/MS runs. The error bars represent the standard deviations of the number of identifications. (C) Box chart of the number of matched fragment ions of

identified E. coli proteoforms. (D) Gene Ontology cellular component analysis of identified E. coli proteins from the two CZE-MS/MS analyses. (E) Box charts of GRAVY values of the identified proteoforms from the two CZE-MS/MS analyses in this work (SDS-batch 1 and SDS-batch 2) and from our previous work in reference 41 (8M urea).

We also analyzed the HepG2 cell proteins prepared by the MU method using our dynamic pH junction-based CZE-MS/MS. The same CZE and MS conditions as the *E. coli* samples were used here except that we employed 40% (v/v) acetic acid as the BGE of CZE due to much higher complexity of the human cell line sample compared to the *E. coli* sample. The CZE-MS/MS identified 534 proteoforms and 248 proteins in a single run with a 5% proteoform-level FDR. When a 1% proteoform-level FDR was used, 516 proteoforms corresponding to 241 proteins were identified. **Figure 5.5A** shows the base peak electropherogram of the CZE-MS/MS run. The mass of identified proteoforms ranged from about 1 kDa to roughly 24 kDa, **Figure 5.5B**. Over 200 proteoforms had mass higher than 10 kDa. Out of the 248 identified proteins, 125 proteins are membrane proteins, 112 proteins are located in nucleus, and 22 proteins belong to chromatin according to the information from the UniProt Knowledgebase (https://www.uniprot.org/). Sequences and fragmentation patterns of two transmembrane proteins (6.8 kDa mitochondrial proteolipid and Cytochrome c oxidase subunit 6A1, mitochondrial) are shown in **Figures 5.5C** and **5.5D**. The two membrane proteins were identified with high confidence and the TMDs were cleaved reasonably well in gas phase by HCD. **Figures 5.5E** and **5.5F** show the mass spectrum and fragmentation pattern of one proteoform of C1QBP (Complement component 1 Q subcomponent-binding protein, mitochondrial) having a mass of 23767.7 Da. The proteoform had clear signal in the mass spectrum and

was identified by MS/MS through the database search with 18 matched fragment ions and a 2.53E-11 E-value. An N-terminal truncation was determined for the proteoform.

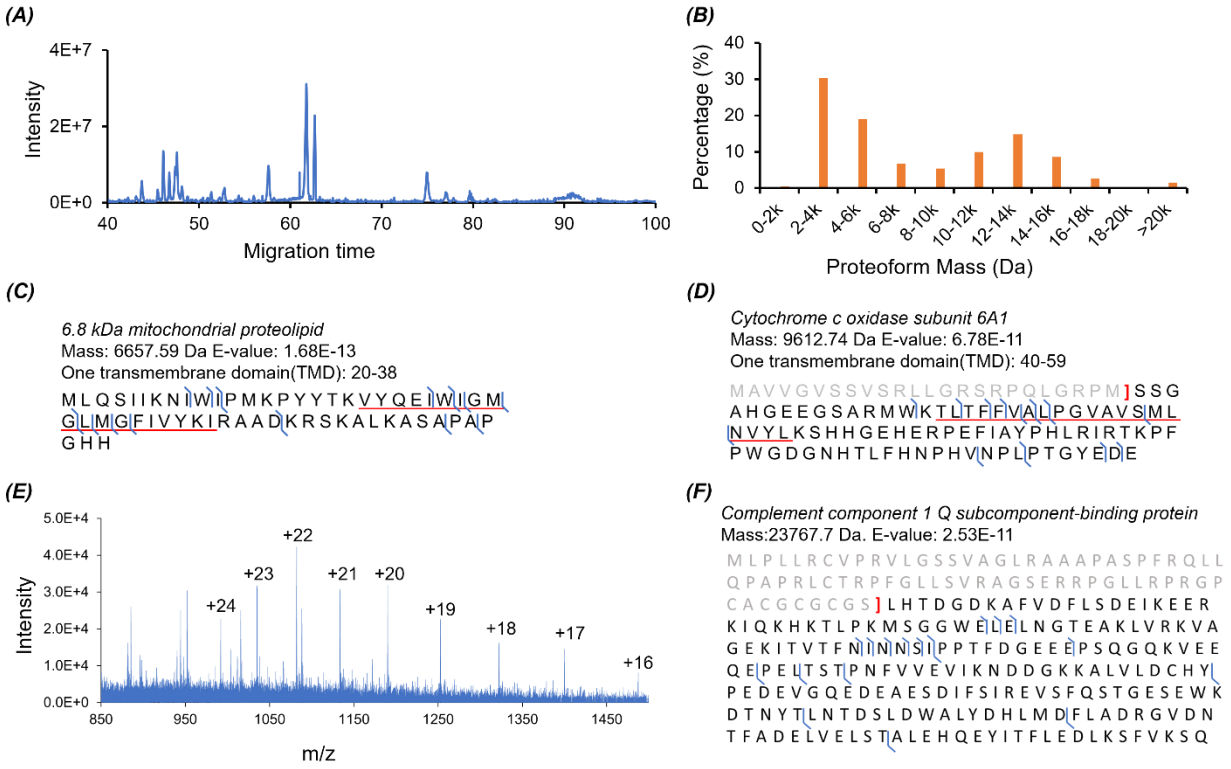

Figure 5. 5 CZE-MS/MS data of the HepG2 cell protein sample prepared with the MU method. (A) Base peak electropherogram of the protein sample after CZE-MS/MS analysis. (B) Mass distribution of the identified proteoforms from the HepG2 protein sample. (C) and (D): Sequences and fragmentation patterns of two transmembrane proteins with one TMD. The regions corresponding to TMDs are underlined. (E) Mass spectrum of the identified proteoform of C1QBP (Complement component 1 Q subcomponent-binding protein, mitochondrial) with a mass of 23767.7Da. (F) Sequence and fragmentation pattern of the C1QBP proteoform in (E).

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

5.3.3 Proteoforms with post-translational modifications (PTMs)

We also performed another CZE-MS/MS run of the prepared *E. coli* sample from the MU method under very clean CZE and MS conditions to pursue a higher number of proteoform identifications, leading to an identification of 1,336 proteoforms corresponding to 301 proteins with a 1% proteoform-level FDR. Various protein modifications were detected, including but not limited to N-terminal methionine removal, N-terminal truncation, N-terminal acetylation, and disulfide bond, **Figure 5.6A**. Two truncated proteoforms of 50S ribosomal protein L7/L12 at the N-terminus with or without lysine methylation are shown in **Figures 5.6B** and **5.6C**. The fragmentation patterns show extensive backbone cleavages of the two proteoforms. We also observed that the abundance of the methylated proteoform was about 50% of the non-methylated proteoform according to the mass spectrum in **Figure 5.6D**. The methylation at Lys-82 detected in our work agrees well with the data in the literature.[42] We identified 15 proteoforms with one or two disulfide bonds and those proteoforms. Sequences and fragmentation patterns of two proteoforms with one and two disulfide bonds are shown in **Figures 5.6E** and **5.6F**. Interestingly, for **Figure 5.6E**, the location of the disulfide bond was previously reported as zinc ion binding position.[43] For the 50S ribosomal protein L31, the literature data suggested that the C16 was responsible for zinc ion binding, but our data show that the C16, C18, C37 and C40

form two disulfide bonds, **Figure 5.6F**. The disulfide bonds might form endogenously or develop after cell lysis due to the loss of zinc ions during sample preparation.

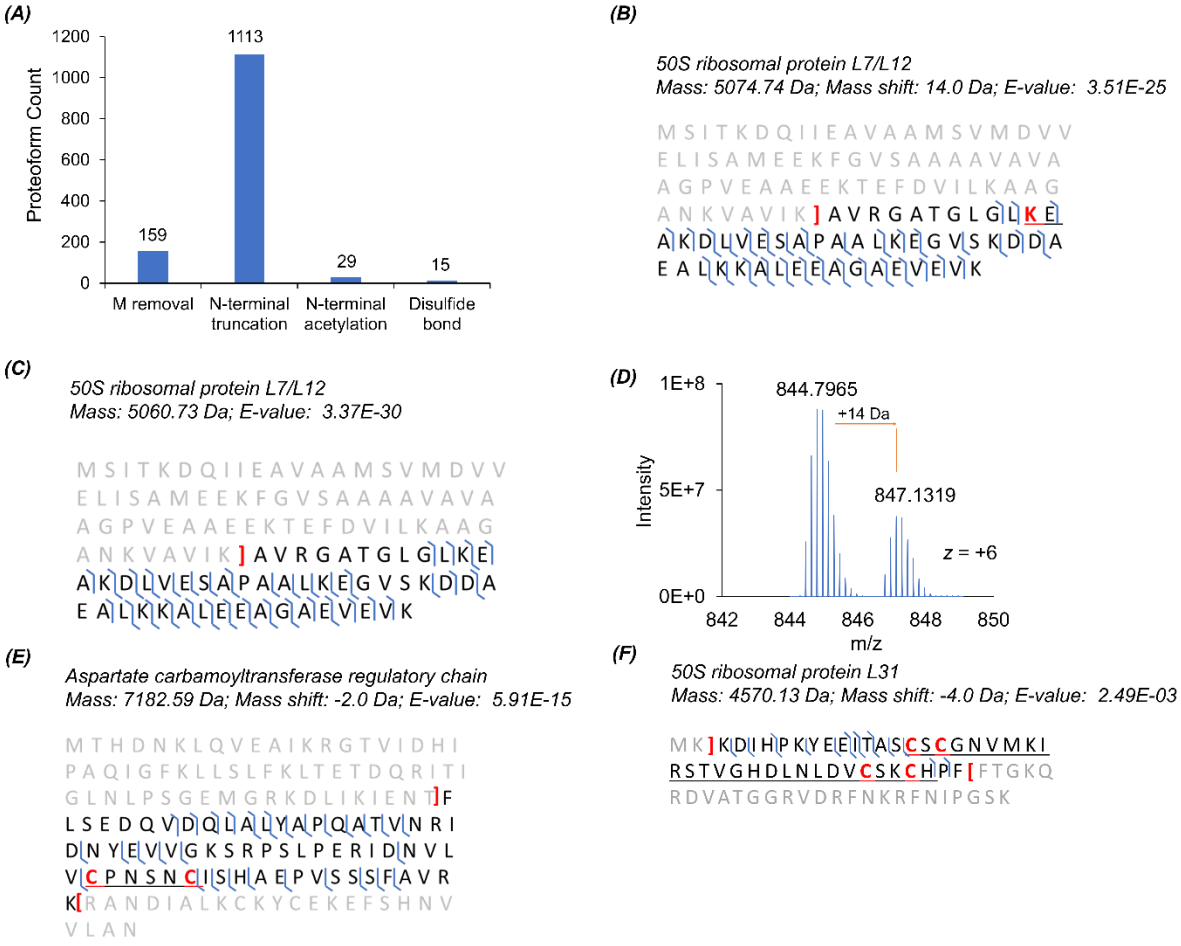

Figure 5. 6 CZE-MS/MS data of the E. coli sample regarding PTMs. (A) Distribution of some modifications on the identified proteoforms. (B) Sequence and fragmentation pattern of the 50S ribosomal protein L7/L12 proteoform with one methylation at the marked lysine residue and N-terminal truncation. (C) Sequence and fragmentation pattern of the 50S ribosomal protein L7/L12 proteoform with only N-terminal truncation. (D) Mass spectrum of one charge state (+6) of the 50S ribosomal protein L7/L12 proteoforms in (B) and (C). (E) Sequence and fragmentation pattern of one proteoform of aspartate carbamoyltransferase regulatory chain with one disulfide bond between the

two marked cystine residues and truncations at the termini. (F) Sequence and fragmentation pattern of one proteoform of 50S ribosomal protein L31 with two disulfide bonds among the four marked cystine residues, removal of two amino acid residues at the N-terminus, and truncation at the C-terminus.

We identified proteoforms with various PTMs in the HepG2 data, including but not limited to N-terminal acetylation (205), phosphorylation (11), and disulfide bonds (8), **Figure 5.7A**. We identified one proteoform of programmed cell death protein 5 with N-terminal acetylation and one serine phosphorylation (**Figure 5.7B**), one proteoform of 60S acidic ribosomal protein P2 with two serine phosphorylations (**Figure 5.7C**), and one proteoform of small ubiquitin-related modifier 1 with both acetylation and phosphorylation at the N-terminal serine residue (**Figure 5.7D**). The PTM information of these three proteoforms match well with the UniProt Knowledgebase (https://www.uniprot.org/). We noted that the three serine residues marked in red in the underlined region in **Figure 5.7C** could be phosphorylated according to the UniProt Knowledgebase, and our data show that only two of them are actually phosphorylated in the proteoform.

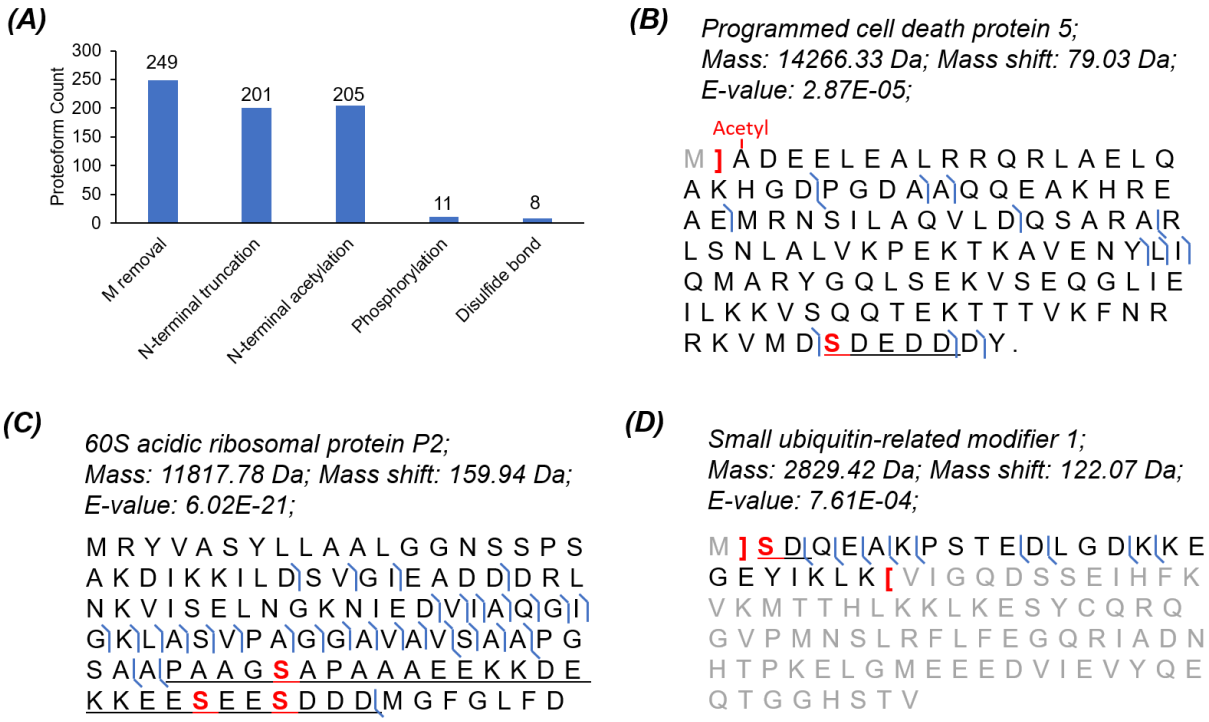

Figure 5. 7 CZE-MS/MS data of the HepG2 sample regarding PTMs. (A) Distribution of some modifications on the identified proteoforms. Sequences and fragmentation patterns of some proteoforms with one phosphorylation site and N-terminal acetylation (B), with two phosphorylation sites (C), and with phosphorylation and acetylation on the N-terminal serine residue (D).

Prothymosin alpha (PTMA) is a histone binding protein and it can regulate gene transcription.[44] Prothymosin alpha has eight phosphorylation sites according to the UniProt Knowledgebase. Our data revealed one phosphorylation site (mass shift 79.97 Da) in the underlined region (S85 or T87) in **Figure 5.8A**, which is not reported previously. We also compared the relative abundance of the identified phosphorylated proteoform of PTMA and the corresponding unphosphorylated proteoform based on the extracted base peak electropherogram, **Figure 5.8B**. The unphosphorylated proteoform had about 5-times higher abundance than the phosphorylated one. Additionally, CZE separated the

phosphorylated and unphosphorylated proteoforms very well with an 8-min difference in migration time and the phosphorylated one migrated obviously slower than the unphosphorylated one in CZE due to the charge reduction from the phosphorylation, which agrees well with the previous reports.[45-48] The migration time shift between unphosphorylated and phosphorylated proteoforms provides additional evidence for the phosphorylation PTM. **Figures 5.8C** and **5.8D** show mass spectra of the unphosphorylated and phosphorylated proteoforms, indicating a difference between them regarding charge distribution. We speculate that the phosphorylation could influence the ESI of prothymosin alpha to some extent.

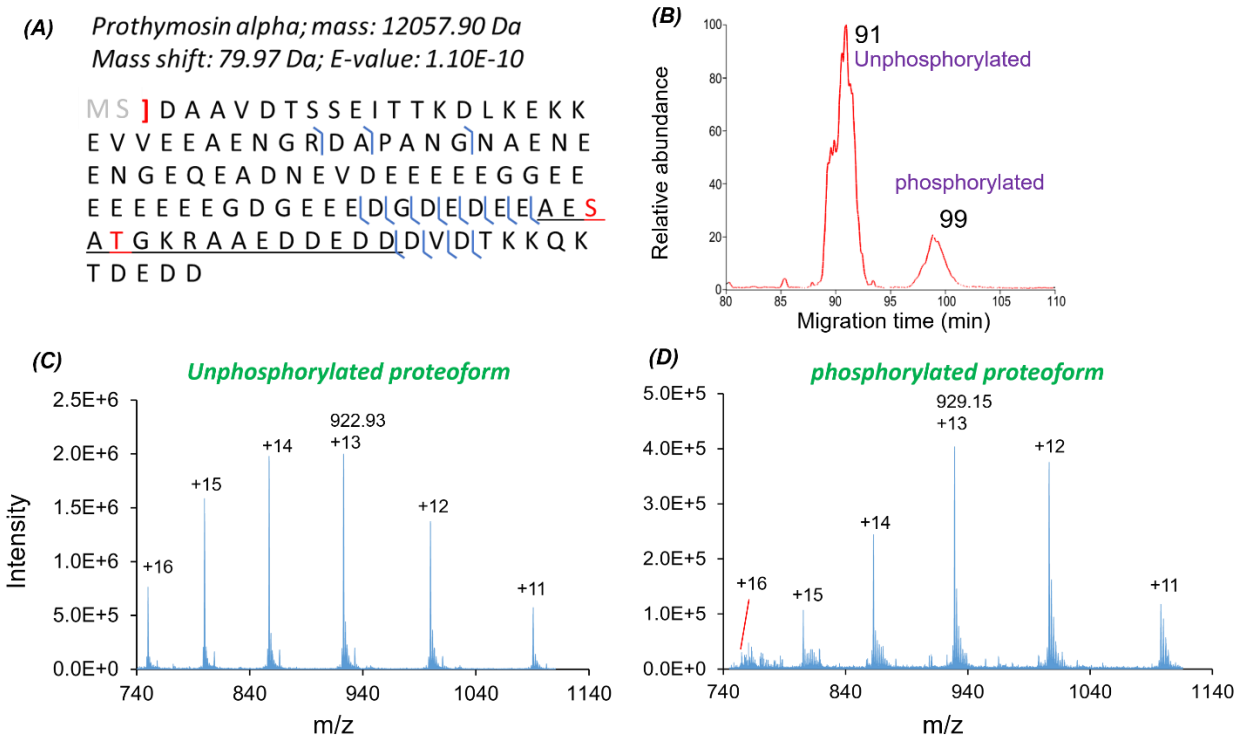

Figure 5. 8 Proteoform information of prothymosin alpha. (A) Sequence and fragmentation pattern of the prothymosin alpha proteoform with one phosphorylation at the marked serine or threonine residue. (B) Extracted base peak electropherogram of phosphorylated and unphosphorylated proteoforms of prothymosin alpha. The highest

abundant charge state (+13) was used for peak extraction with a 20-ppm mass tolerance. Mass spectra of the unphosphorylated (C) and phosphorylated (D) proteoforms of prothymosin alpha.

Histone PTMs are extremely important for regulating gene expression and dTDP is an invaluable approach for delineating the histone code in a proteoform specific manner.[49-52] In this work, we identified 94 histone proteoforms from the HepG2 sample in a single CZE-MS/MS run without any histone purification. The 94 histone proteoforms covered the five major histone variants, H1 (11), H2A (39), H2B (36), H3 (1) and H4 (7), **Figure 5.9A**. We observed various PTMs on the histone proteoforms, including acetylation, methylation, and phosphorylation, **Figures 5.9B-F**. Sequences and fragmentation patterns of two histone H4 proteoforms are shown in **Figure 5.9B** and **C**. We observed both N-terminal acetylation and a 28-Da mass shift most likely corresponding to two methylations within the underlined region in the two proteoforms. Due to the limited backbone cleavage coverages for the two proteoforms, it is difficult to localize the methylation PTM. Interestingly, there are no literature reports about methylation or di-methylation PTM in the two regions of histone H4 underlined in **Figures 5.9B** and **C** according to the UniProt Knowledgebase. We also identified one histone H4 proteoform with a 337-Da mass shift, **Figure 5.9D**. The mass shift corresponds to a region with four lysine residues (K6, K9, K13 and K17). According to the UniProt Knowledgebase, these four lysine residues could have acetylation (+42 Da), propionylation (+56 Da), crotonylation (+68 Da), butyrylation (+70 Da), succinylation (+100 Da), and glutarylation (+114 Da). We speculate that the 337-Da mass shift is most likely produced by a

combination of these various PTMs. The data further suggest the importance of improving the backbone cleavage coverage for comprehensive characterization of proteoforms.

We identified one proteoform of Histone H2A type 1-J with a 122-Da mass shift in the underlined region, **Figure 5.9E**. We speculate that the mass shift corresponds to an acetylation (+42 Da) and a phosphorylation (+80 Da). It has been reported that the K6 and K10 residues could be acetylated.[53] However, no literature information about the phosphorylation at T17, S19 or S20 in the mass shift corresponding region according to the UniProt Knowledgebase. We also identified one Histone H2A type 1 proteoform with an 83-Da mass shift in the underlined region, **Figure 5.9F**. The K96 and K100 in the mass shift corresponding region could be acetylated based on the previous reports [53,54] and the information from PhosphoSitePlus® v6.5.8 (https://www.phosphosite.org/). Two lysine acetylation modifications produce an 84-Da mass shift, which is 1-Da heavier than the observed mass shift. The 1-Da difference could be due to a misassignment of the monoisotopic peak of the protein, which resulted in a 1-Da error of the proteoform's monoisotopic mass. Therefore, the observed 83-Da mass shift is most likely due to the acetylation at both K96 and K100.

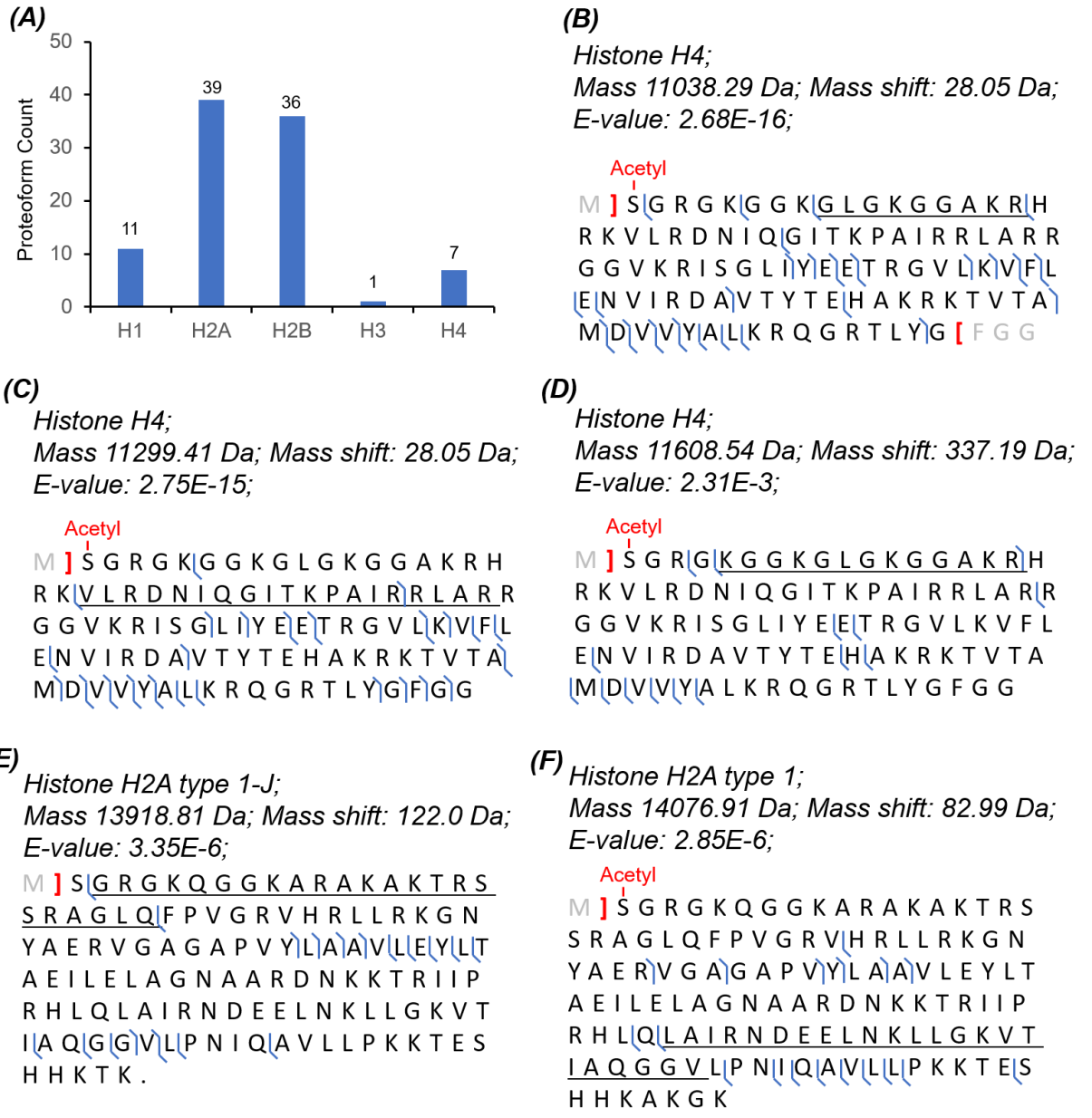

Figure 5. 9 CZE-MS/MS data of the HepG2 sample regarding histone proteoforms.

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

5.4 Conclusions

We performed comprehensive comparisons of the MU, CMP, and SP3 methods for cleanup of proteome samples in a lysis buffer containing SDS regarding protein recovery, protein bias, and compatibility with follow-up MS analysis. Our data indicate that the SDS-based protein extraction and the MU-based protein cleanup could be a universal sample preparation procedure for dTDP of complex proteome samples. The procedure produced reproducible sample preparation with high protein recovery for both *E. coli* and human cell line samples. Single-shot CZE-MS/MS analysis of the prepared *E. coli* and HepG2 cell proteome samples (400-ng proteins consumed) identified up to 1 336 proteoforms (301 proteins) and 516 proteoforms (241 proteins) with a 1% proteome-level FDR, respectively. Single-shot CZE-MS/MS analysis of the HepG2 cell sample identified 125 membrane proteins and 94 histone proteoforms. The sample preparation procedure including the SDS-based protein extraction and the MU-based protein cleanup should be also compatible with the widely used RPLC-MS/MS approach, although we only used CZE-MS/MS in this work.

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

5.5 Acknowledgements

We thank Prof. Heedeok Hong's group at the Department of Chemistry of Michigan State University and Prof. David M. Lubman's group at the University of Michigan for kindly providing the *E. coli* cells and HepG2 human cells for this project. We thank the

support from the National Science Foundation (CAREER Award, Grant DBI1846913) and the National Institutes of Health (Grant R01GM125991).

REFERENCES

REFERENCES

1. Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P., *Nature methods* **2013,** *10* (3), 186-7.

2. Toby, T. K.; Fornelli, L.; Kelleher, N. L., *Annual review of analytical chemistry* **2016,** *9* (1), 499-519.

3. Smith, L. M.; Kelleher, N. L., *Science* **2018,** *359* (6380), 1106-1107.

4. Ntai, I.; Fornelli, L.; DeHart, C. J.; Hutton, J. E.; Doubleday, P. F.; LeDuc, R. D.; van Nispen, A. J.; Fellers, R. T.; Whiteley, G.; Boja, E. S.; Rodriguez, H.; Kelleher, N. L., *Proceedings of the National Academy of Sciences of the United States of America* **2018,** *115* (16), 4140-4145.

5. Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L., *Nature* **2011,** *480* (7376), 254-8.

6. Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L., *Molecular & cellular proteomics : MCP* **2013,** *12* (12), 3465-73

7. Ljiljana, Paša-Tolić.; Pamela, K. Jensen.; Gordon, A. Anderson.; Mary, S. Lipton.; Kim, K. Peden.; Suzana, Martinović.; Nikola, Tolić.; James, E. Bruce.; Richard, D. Smith., *J. Am. Chem. Soc.* **1999**, *121*(34), 7949-7950

8. Anderson, L. C.; DeHart, C. J.; Kaiser, N. K.; Fellers, R. T.; Smith, D. F.; Greer, J. B.; LeDuc, R. D.; Blakney, G. T.; Thomas, P. M.; Kelleher, N. L.; Hendrickson, C. L., *Journal of proteome research* **2017,** *16* (2), 1087-1096.

9. Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y., *Analytical chemistry* **2017,** *89* (10), 5467-5475.

10. Liang, Y.; Jin, Y.; Wu, Z.; Tucholski, T.; Brown, K. A.; Zhang, L.; Zhang, Y.; Ge, Y., *Analytical chemistry* **2019,** *91* (3), 1743-1747.

11. McCool, E. N.; Lubeckyj, R. A.; Shen, X.; Chen, D.; Kou, Q.; Liu, X.; Sun, L., *Analytical chemistry* **2018,** *90* (9), 5529-5533.

12. Yu, D.; Wang, Z.; Cupp-Sutton, K. A.; Liu, X.; Wu, S., *Journal of the American Society for Mass Spectrometry* **2019,** *30* (12), 2502-2513.

13. Ansong, C.; Wu, S.; Meng, D.; Liu, X.; Brewer, H. M.; Deatherage Kaiser, B. L.; Nakayasu, E. S.; Cort, J. R.; Pevzner, P.; Smith, R. D.; Heffron, F.; Adkins, J. N.; Pasa-Tolic, L., *Proceedings of the National Academy of Sciences of the United States of America* **2013,** *110* (25), 10153-8.

14. Lubeckyj, R. A.; Basharat, A. R.; Shen, X.; Liu, X.; Sun, L., *Journal of the American Society for Mass Spectrometry* **2019,** *30* (8), 1435-1445.

15. Liu, Z.; Wang, R.; Liu, J.; Sun, R.; Wang, F., *Journal of proteome research* **2019,** *18* (5), 2185-2194.

16. Riley, N. M.; Westphall, M. S.; Coon, J. J., *Journal of proteome research* **2017,** *16* (7), 2653-2659.

17. Shaw, J. B.; Li, W.; Holden, D. D.; Zhang, Y.; Griep-Raming, J.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S., *Journal of the American Chemical Society* **2013,** *135* (34), 12646-51.

18. Shaw, J. B.; Malhan, N.; Vasil'ev, Y. V.; Lopez, N. I.; Makarov, A.; Beckman, J. S.; Voinov, V. G., *Analytical chemistry* **2018,** *90* (18), 10819-10827.

19. Kou, Q.; Xun, L.; Liu, X., *Bioinformatics* **2016,** *32* (22), 3495-3497.

20. Cai, W.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X.; Ge, Y., *Molecular & cellular proteomics : MCP* **2016,** *15* (2), 703-14.

21. Sun, R. X.; Luo, L.; Wu, L.; Wang, R. M.; Zeng, W. F.; Chi, H.; Liu, C.; He, S. M., *Analytical chemistry* **2016,** *88* (6), 3082-90.

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

23. Speers, A. E.; Wu, C. C., *Chemical reviews* **2007,** *107* (8), 3687-714.

24. Botelho, D.; Wall, M. J.; Vieira, D. B.; Fitzsimmons, S.; Liu, F.; Doucette, A., *Journal of proteome research* **2010,** *9* (6), 2863-70.

25. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., *Nature methods* **2009,** *6* (5), 359-62.

26. Wessel, D.; Flugge, U. I., *Analytical biochemistry* **1984,** *138* (1), 141-3.

27. Hughes, C. S.; Foehr, S.; Garfield, D. A.; Furlong, E. E.; Steinmetz, L. M.; Krijgsveld, J., *Molecular systems biology* **2014,** *10*, 757.

28. Hughes, C. S.; Moggridge, S.; Muller, T.; Sorensen, P. H.; Morin, G. B.; Krijgsveld, J., *Nature protocols* **2019,** *14* (1), 68-85.

29. Dagley, L. F.; Infusini, G.; Larsen, R. H.; Sandow, J. J.; Webb, A. I., *Journal of proteome research* **2019,** *18* (7), 2915-2924.

30. Blancher, C.; Jones, A., *Methods in molecular medicine* **2001,** *57*, 145-62.

31. McCool, E. N.; Lubeckyj, R.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., *Journal of visualized experiments: JoVE* **2018,** *140*, e58644.

32. Zhu, G.; Sun, L.; Dovichi, N. J., *Talanta* **2016,** *146*, 839-43.

33. Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J., *Angewandte Chemie* **2013,** *52* (51), 13661-4.

34. Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J., *Journal of proteome research* **2015,** *14* (5), 2312-21.

35. Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J., *Rapid communications in mass spectrometry : RCM* **2010,** *24* (17), 2554-60.

36. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., *Bioinformatics* **2008,** *24* (21), 2534-6.

37. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., *Analytical chemistry* **2002,** *74* (20), 5383-92.

38. Elias, J. E.; Gygi, S. P., *Nature methods* **2007,** *4* (3), 207-14.

39. Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A., *Nucleic acids research* **2019,** *47* (D1), D442-D450.

40. Arribas, J.; Castano, J. G., *The Journal of biological chemistry* **1990,** *265* (23), 13969-73.

41. Lubeckyj, R. A.; McCool, E. N.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., *Analytical chemistry* **2017,** *89* (22), 12059-12067.

42. Chang, C. N.; Chang, N., *Biochemistry* **1975,** *14* (3), 468-77.

43. Monaco, H. L.; Crawford, J. L.; Lipscomb, W. N., *Proceedings of the National Academy of Sciences of the United States of America* **1978,** *75* (11), 5276-80.

44. Karetsou, Z.; Kretsovali, A.; Murphy, C.; Tsolas, O.; Papamarcaki, T., *EMBO reports* **2002,** *3* (4), 361-6.

45. Chen, D.; Lubeckyj, R. A.; Yang, Z.; McCool, E. N.; Shen, X.; Wang, Q.; Xu, T.; Sun, L., *Analytical chemistry* **2020,** *92* (5), 3503-3507.

46. Wojcik, R.; Vannatta, M.; Dovichi, N. J., *Analytical chemistry* **2010,** *82* (4), 1564-7.
47. Kim, J.; Zand, R.; Lubman, D. M., *Electrophoresis* **2003,** *24* (5), 782-93.

48. Faserl, K.; Sarg, B.; Gruber, P.; Lindner, H. H., *Electrophoresis* **2018,** *39* (9-10), 1208-1215.

49. Zheng, Y.; Huang, X.; Kelleher, N. L., *Current opinion in chemical biology* **2016,** *33*, 142-50.

50. Janssen, K. A.; Sidoli, S.; Garcia, B. A., *Methods in enzymology* **2017,** *586*, 359-378.

51. Wang, T.; Holt, M. V.; Young, N. L., *Epigenetics* **2018,** *13* (5), 519-535.

52. Gargano, A. F. G.; Shaw, J. B.; Zhou, M.; Wilkins, C. S.; Fillmore, T. L.; Moore, R. J.; Somsen, G. W.; Pasa-Tolic, L., *Journal of proteome research* **2018,** *17* (11), 3791-3800.

53. Wu, Q.; Cheng, Z.; Zhu, J.; Xu, W.; Peng, X.; Chen, C.; Li, W.; Wang, F.; Cao, L.; Yi, X.; Wu, Z.; Li, J.; Fan, P., *Scientific reports* **2015,** *5*, 9520.

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

SUMMARY

Proteomics analysis of mass limited sample such as single cell proteomics is going through initial stage of development but is progressing very rapidly. We have seen the proteome coverage increased from a couple hundreds of identifications to over one thousand identifications from a single mammalian cell. Keys of large-scale protein characterization of single cells regarding sample preparation and separation are reducing sample loss and reducing flow rate (without sacrificing separation efficiency) to improve sample recovery and improve ionization efficiency during ESI, respectively. For MS instrumentation, in addition to improving acquisition rate and resolving power, the installation of ion mobility mass spectrometer such as FAIMS and TIMS provides additional level of fractionation and reduces background noise which are all beneficial for low input analysis. Not only instrumentation hardware, but the data acquisition strategy of MS also plays vital role to improve performance of single cell proteomics. Comparing to conventional data acquisition mode in most proteomics analysis(i.e., data dependent acquisition (DDA)), data independent acquisition (DIA) strategy has unique advantages. In DDA, peptides are sequentially picked for fragmentation, in order of precursor ion intensity that is determined in survey scan, leading to inherent randomness of peptide fragmentation and missing value problem. The missing value problem is more serious for low abundant peptides and leads to poor reproducibility for low-input sample analysis. DIA on the other hand, fragments all the peptides in broader isolation window, potentially improves sensitivity, dynamic range, reproducibility, and quantification accuracy.

Although bottom-up proteomics is the major strategy for single cell proteomics due to the fact that it has drastically better sensitivity than top-down proteomics, most recent top-down proteomic studies demonstrated the identification and quantification of hundreds to thousands of proteoforms from nanograms-input of proteome samples. The data suggests the potential of using top-down proteomics for the characterization of mass limited proteome samples, even single cells.

One significance of proteomics analysis is the PTM information that can't be revealed through transcriptomics and genomics studies. However, PTMs analysis of mass limited sample still have great room of development due to the low abundance of proteins containing specific PTMs. Enrichment regarding PTMs of interest is usually necessary, and it usually requires high protein sample input in mass. It determines an important direction of technique advancement in proteomics of mass limited sample.

Unlike transcriptomics/genomics analysis, proteomics analysis can't be parallelly processed with throughput of thousands cells/analysis, although the multiplexity technique for protein analysis is progressing. Continuous investment regarding automation, multiplexity and shortened analysis time are crucial for large-scale single cell proteomics studies.