DEVELOPMENT OF NOVEL TECHNIQUES FOR TOP-DOWN PROTEOMICS OF
COLORECTAL CANCER CELLS

By

Olivia Gordon

A THESIS

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

Chemistry – Master of Science

2024

ABSTRACT

Colorectal cancer (CRC) is one of the most life-threatening and prevalent forms of cancer worldwide. A better molecular-level understanding of CRC will produce novel protein biomarkers for CRC diagnosis and therapy development. Proteins play fundamental roles in modulating almost all the biological processes in cells, and different proteoforms of the same gene can have divergent biological functions. Therefore, large-scale studies of proteins in a proteoform-specific manner in CRC cells using mass spectrometry (MS)-based proteomics provide a wonderful opportunity for bettering our understanding of CRC progression and discovering new proteoform biomarkers. MS-based top-down proteomics (TDP) is ideal for the characterization of proteoforms because it measures intact proteoforms directly by employing liquid-phase separations and MS. However, TDP still faces many challenges. One of them is related to the measurement of large proteoforms (>30 kDa) from complex samples (i.e., CRC cells) due to their much wider charge state distributions and more isotopic peaks in each charge state compared to small proteoforms, leading to substantially lower signal-to-noise ratios. Many critical proteins related to CRC are larger than 30 kDa, *e.g.*, DNA mismatch repair proteins (MSH2), EPCAM, and TP53. Therefore, the development of new TDP techniques for large proteoforms is fundamental to advance our understanding of CRC progression and biomarker discovery. Capillary zone electrophoresis (CZE)-MS is an attractive technique for TDP and has the potential to address the issues of large proteoform separation and MS detection. However, in published CZE-MS-based TDP datasets, almost all the identified proteoforms are smaller than 30 kDa. It is suspected that this phenomenon is due to sample preparation techniques and CZE-MS conditions favoring smaller proteoforms. Here, we aim to provide an improved TDP workflow for large proteoforms using optimized sample preparation and CZE-MS/MS methods.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CRC – colorectal cancer

gFOBT – guaiac fecal occult blood test

FIT – fecal immunochemical test

PTM – post-translational modification

DNA – deoxyribonucleic acid

RNA – ribonucleic acid

KRAS – Kristen Rat Sarcoma Viral oncogene homolog

MAPK – mitogen-activated protein kinase

MS – mass spectrometry

m/z – size-to-charge ratio

ESI – electrospray ionization

LC – liquid chromatography

CE – capillary electrophoresis

BUP – bottom-up proteomics

TDP – top-down proteomics

SEC – size-exclusion chromatography

CZE – capillary zone electrophoresis

BGE – background electrolyte

AGC – automatic gain control

kDa – kilodaltons

EOF – electroosmotic flow

AA – acetic acid

LC – liquid chromatography

MeOH – methanol

FA – formic acid

CZE-MS – capillary zone electrophoresis-mass spectrometry

ACN – acetonitrile

BSA – bovine serum albumin

CA – carbonic anhydrase

MyO – myoglobin

Ubq – ubiquitin

ABC – ammonium bicarbonate

DPBS – Dulbecco's phosphate buffered saline

NP – nanoparticle

SDS – sodium dodecyl sulfate

PAGE – polyacrylamide gel electrophoresis

INTRODUCTION

Colorectal cancer (CRC) is a sizeable issue in the world of oncological treatment and research. Today it is the third most prevalent, and one of the most life-threatening forms of cancer, with a significant impact on global public health[1,2].

This disease typically arises from the abnormal and uncontrolled growth of cells, called polyps, strictly confined to the colon or rectal lining[3,4]. Often these two areas of occurrence are combined into one, hence the name "colorectal" (Figure 1).
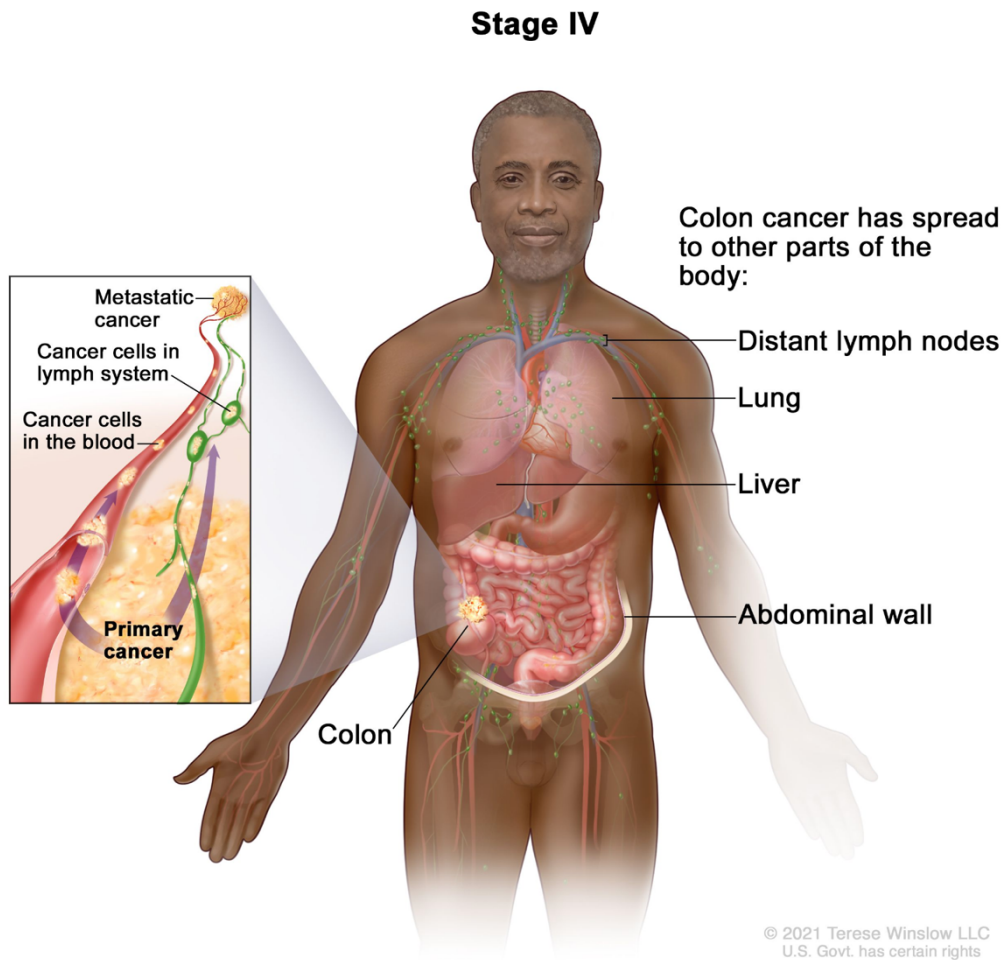

Figure 1. Image depicting stage IV colon cancer placement and progression. Image adapted from reference [3].

CRC has several distinct stages, each characterized by the extent of tumor formation and metastasis, that are important to understand when it comes to determining prognosis and treatment options[3–6]. Early diagnosis will significantly improve a patient's chances of successful treatment as well as long-term survival[5,6].

Currently, the standard method for colorectal cancer screening is through a colonoscopy[7]. A colonoscopy is a medical procedure that involves the insertion of a colonoscope through the rectum into the colon. The colonoscope then inflates the colon with air for a better view of the colorectal lining as a camera transmits a video image to a monitor for the examiner to study[8]. However, this method of screening has the drawback of poor patient compliance. Due to the invasive nature of the procedure, it carries risks such as hemorrhage, colonic perforation, and cardiorespiratory complications[7,8]. Besides the physical risks, another reason for the lack of compliance can be attributed to procedure discomfort, bowel preparation, or simply shame[8]. Looking towards a more non-invasive method of CRC screening, guaiac fecal occult blood tests (gFOBT) have been developed[7]. This screening method is based on the identification of hemoglobin peroxidase activity in stool. When guaiac is exposed to hemoglobin, the reaction produces a blue color indicating a positive result[9]. FOBTs are an easy and cost-effective at-home method for screening for CRC. Another method commonly used for testing the presence of occult blood in stool is by using a fecal immunochemistry test (FIT). This test employs the use of antibodies to detect blood rather than guaiac's color-changing reaction (Figure 2).
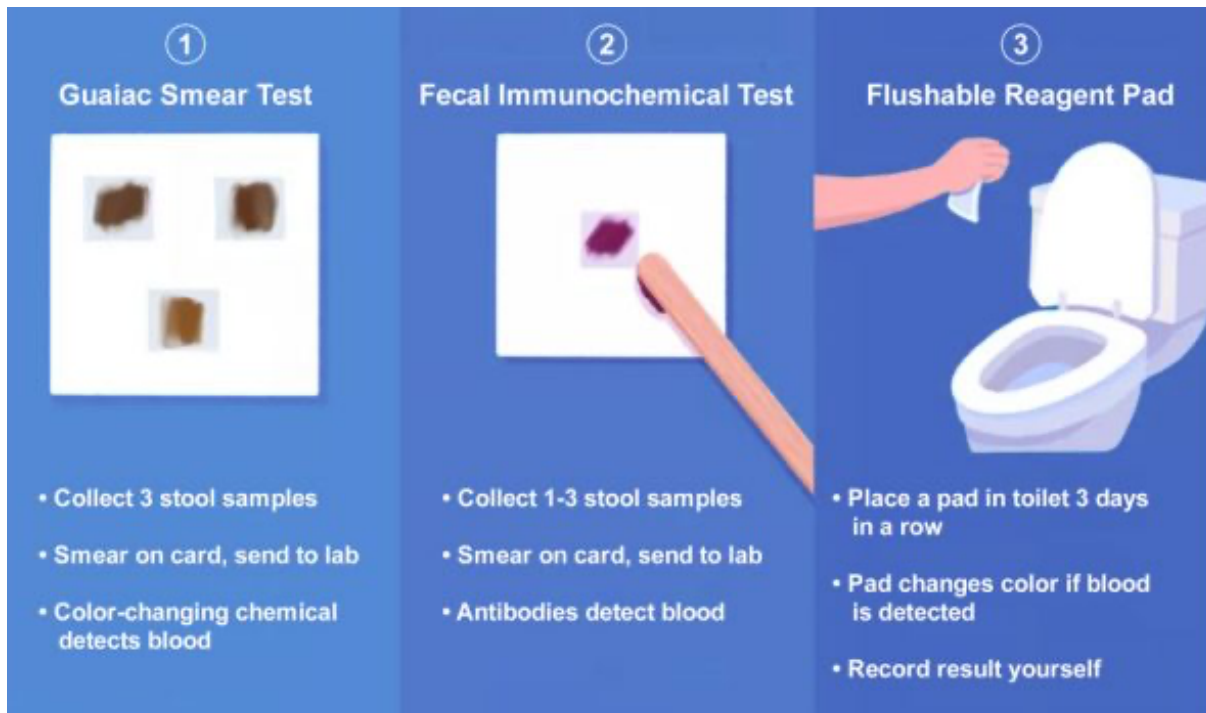
Figure 2. Image showcasing three different methods for non-invasive colorectal cancer screening. Image adapted from reference [9].

Unfortunately, both of these methods have similar drawbacks: the selectivity is poor and the rate of both false positives and negatives is high[7,9]. As for the sensitivity aspect, blood in the stool can be the result of many different medical conditions, not exclusively CRC. On the same note, false positives can be caused by a number of things such as high-meat diets and certain medications like ibuprofen or anticoagulants[9]. In any case, a positive result should be followed up by a colonoscopy for confirmation which, as discussed, has the issue of poor follow-up[7-9]. CRC, like many diseases, is difficult to make an all-inclusive diagnostic test for. This is due in part to the heterogeneity of the disease as it is a prominent and complex aspect of it[10,11]. This heterogeneity poses both challenges and opportunities in the understanding, diagnosis, and management of CRC.

Studies have shown that there are several opportunities for the alteration of CRC cells on both genetic and proteomic levels[10–15]. Tumor heterogeneity can manifest in both an intratumor and intertumor setting (Figure 3)[10].
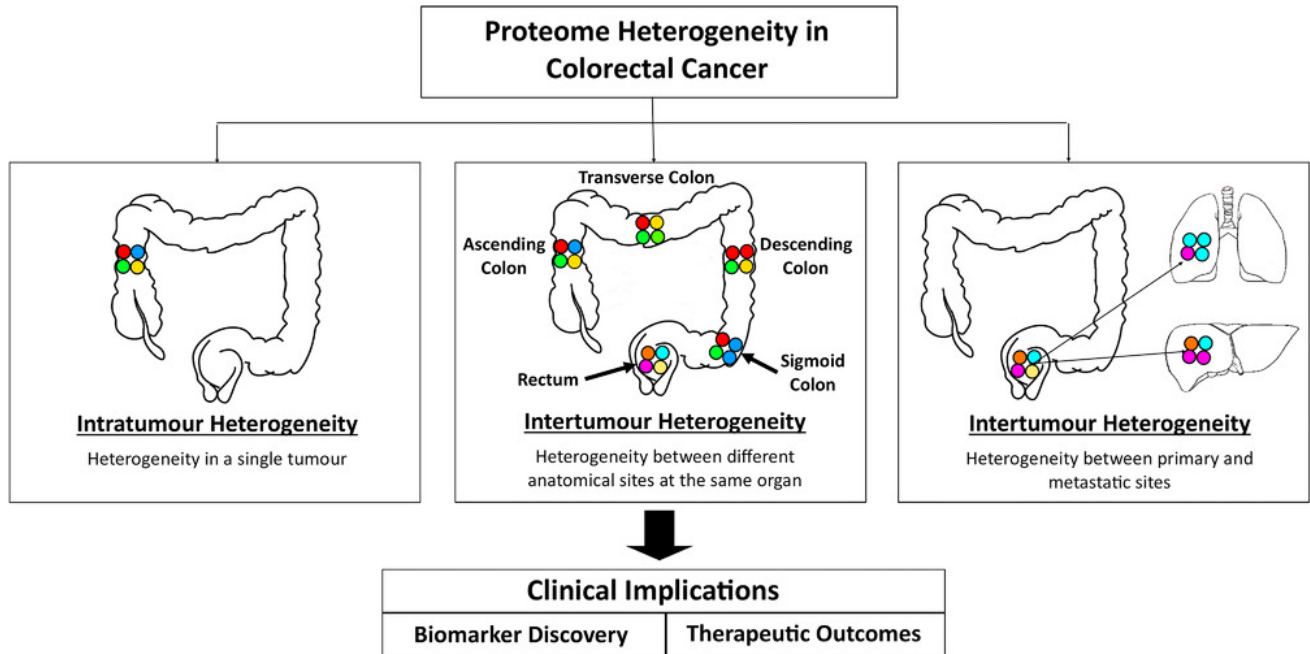

Figure 3. Figure showing how tumors in CRC can be heterogeneous both inter and intratumor. Each colored circle represents a different theoretical protein. Four colored circles together represent the composition of the proteome for one tumor. Figure adapted from reference [10].

Not only that, but these alterations can occur at any stage in the cancer's progression. The complete complexity of this disease is not entirely understood. There are a few different subclasses of CRC recognized based upon genetic changes that display differences in clinicopathological features, yet that does no justice to the true complexity[16]. CRC can be broken down even further into subtypes based on the presence or absence of specific signaling pathway mutations and/or post-translational modifications (PTM) on proteins[11]. Obtaining a better understanding of CRC at every stage in its progression is vital for the development of improved therapies and potential biomarker discovery.

A biomarker is a measurable substance whose presence is indicative of some biological phenomenon[17]. For the study of CRC, the focus is on deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and protein biomarkers. The most widely used biomarkers today are DNA biomarkers[18]. This includes mononucleotide markers, Bat-25 and Bat-26, and mutations in the Kristen Rat Sarcoma Viral oncogene homolog (KRAS) gene. These biomarkers indicate that there are accumulations of alterations in highly repeated DNA sequences and there is increased proliferation through prolonged mitogen-activated protein kinase (MAPK) signaling pathway activation respectively, indicative of CRC[19–22]. Another type of biomarker in clinical trials is RNA, or more specifically microRNA (miRNA), biomarkers[18]. MicroRNAs are small pieces of non-coding RNA that perform a variety of regulatory functions by binding to coding RNA and preventing protein production[23]. The unregulated expression of miRNAs, like miR-106a, may significantly affect cell migration and invasion further contributing to CRC progression[23,24]. However, it is important to note that an increase in RNA does not inherently mean that there will be an increase in protein synthesis and expression[22,25,26]. The final group of biomarkers utilized for the study of CRC are protein biomarkers. Unfortunately, this group severely lacks overall study and application[18]. Therefore, there is interest in expanding research toward the field of proteomic biomarker discovery as proteins are known to have a strong structure-to-function relationship[27].

An attractive method for the discovery of novel protein biomarkers is to utilize mass spectrometry (MS). MS has long been a popular choice of technique for protein analysis due to its ability to handle many of the complexities that are associated with the overall proteome[28,29]. The principle behind MS is to measure the size-to-charge ratio (m/z) of analyte ions in the gas phase. The instrument itself is made of three fundamental parts: the ion source, the mass

analyzer, and the detector[28–30]. Here, the function and selection of ion source and mass analyzer will be discussed. The ion source functions as the converter of analyte molecules into the gas phase ions if not already in that phase[28]. The ion source commonly utilized for large biomolecules such as proteins is electrospray ionization (ESI).

ESI operates by taking advantage of the electrostatic forces between ions in a solution and the surface tension of the solution forming what is known as a Taylor cone[31]. Once a voltage is applied to the system, droplets form, and as the solvent evaporates, the charged droplet reaches the Rayleigh limit and undergoes coulombic fission. This process is repeated until only a nakedly charged analyte is left (Figure 4)[31,32].
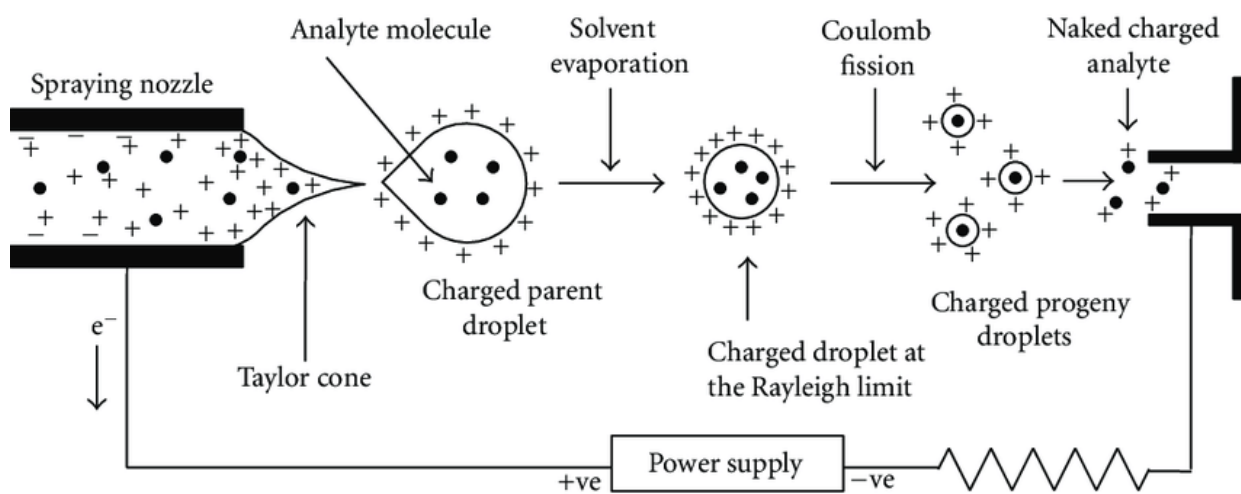

Figure 4. Diagram of ESI principle. Charges are represented by plus icons while analyte molecules are solid black circles. As analyte progresses toward the mass spectrometer inlet (right), the solvent is evaporated and the analyte is left charged alone. Figure adapted from reference [31].

This method of ionization is ideal for proteomic studies for a few reasons. It is a soft ionization technique meaning that it will generate ions without excessive fragmentation which is essential for the characterization of more liable protein interactions. This ionization technique also has a high mass range for tolerating large biomolecules such as proteins. Also, ESI is compatible with both liquid chromatography (LC) and capillary electrophoresis (CE). The mass analyzer

functions as a separator for the generated ions based upon the m/z. The current state-of-the-art

mass analyzer to be used in proteomic studies is the orbitrap[28,33].

An orbitrap is composed of a spindle-like central electrode with an outer electrode surrounding

it. An orbitrap works by harnessing the electrostatic attraction of ions towards the central

electrode compensated by the centrifugal force of the ions from the introduction into the orbitrap.

Ions with different m/z will oscillate at different frequencies. By employing image current as the

method of detection, the m/z of different ions can be determined from their respective oscillation

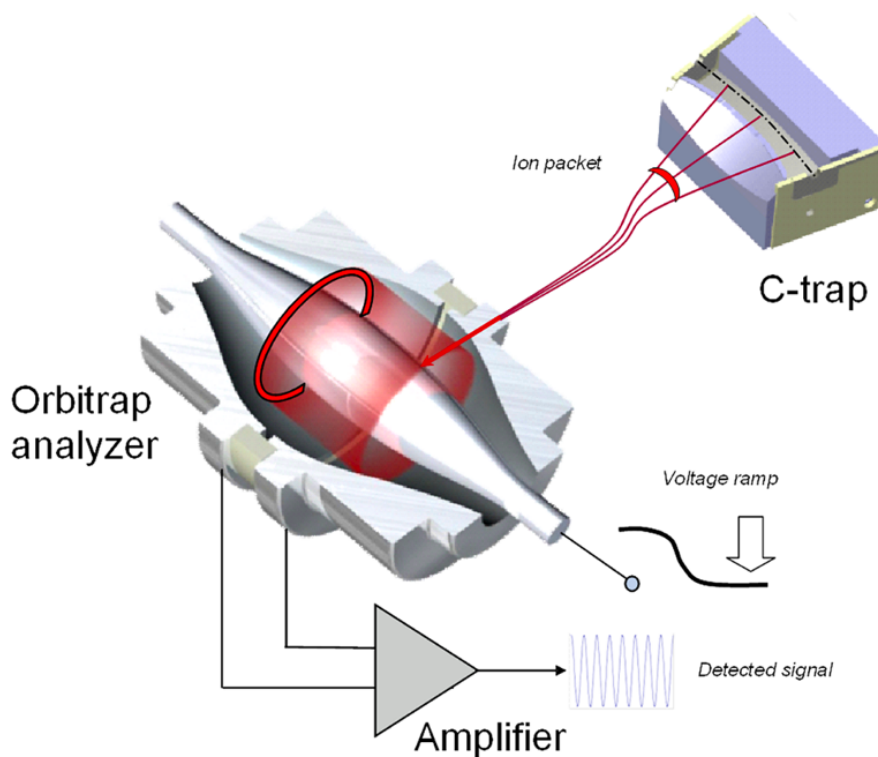frequencies after Fourier transformation (Figure 5)[34].

Figure 5. Diagram of orbitrap mass analyzer. As ions are tangentially introduced to the orbitrap, the electrostatic attraction to the central electrode allows the ions to oscillate at a frequency specific to their m/z. That oscillation frequency is then Fourier transformed to deduce an ion's m/z. Figure adapted from reference [34].

The main reason for using an orbitrap in proteomics is its improved resolution (480,000 at m/z 200) for accurate mass determination and its wide dynamic range (up to 5 orders of magnitude) allowing for the detection of low-concentration species[28,33,34].

As previously mentioned, mass spectrometry is often chosen for proteomic work because it can handle the complexities associated with proteins[28,29]. The existence of proteoforms is where much of this complexity arises. Proteoforms are different forms of a protein produced from the same gene with a variety of sequence variations due to gene mutations or PTMs[35,36]. The function of proteoforms with the same gene family can vary drastically; thus, the identification of specific proteoforms is vital to understanding rapidly changing diseases like CRC[10,27,35,36]. There are currently two approaches to proteoform identification via MS: a bottom-up proteomics (BUP) approach and a top-down proteomics (TDP) approach (Figure 6).
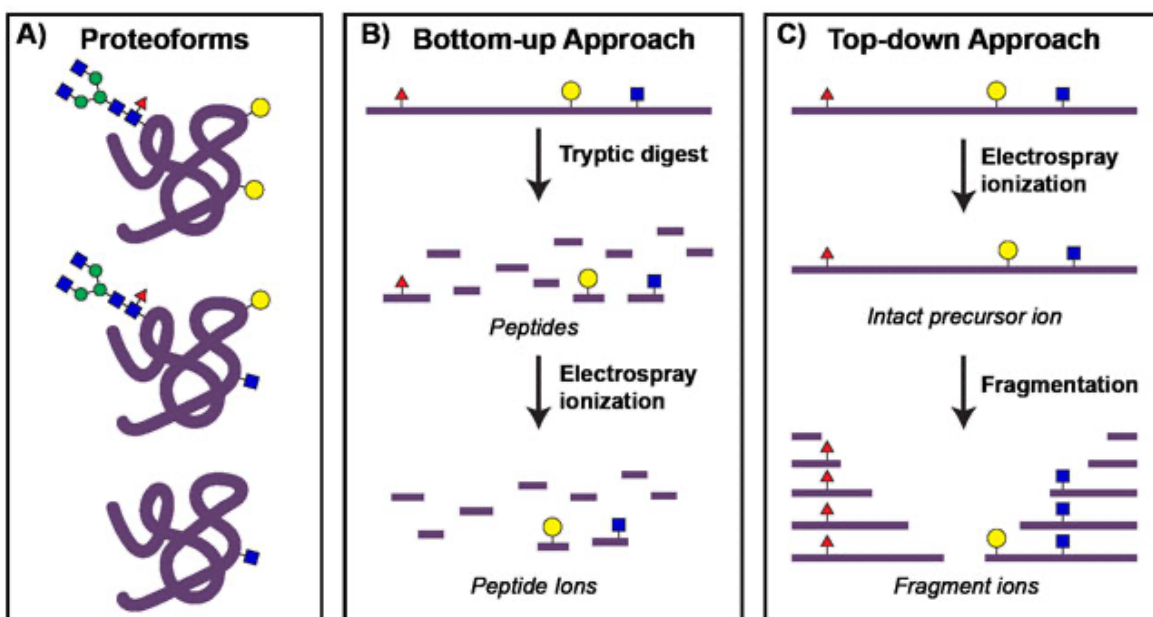

Figure 6. Image detailing aspects of proteoform analysis via MS. A) Examples of proteoforms. B) Workflow for bottom-up proteomics. C) Workflow for top-down proteomics. Illustration adapted from reference [37].

Each method begins with an intact proteoform-containing solution. For BUP, this proteoform extract is digested into much smaller, 8-15 residue peptide chains[38]. Following digestion, the sample is ionized by ESI and analyzed by MS. The difference in the TDP approach is that there is no digestion step. Proteoforms are ionized as intact biomolecules[28,35,36,39,40]. Like most things, there are advantages and disadvantages associated with each technique. BUP has the capability to be high throughput and has well-developed methods as well as large databases readily available. However, it suffers from potential peptide loss during the preparation steps and limited proteoform sequence coverage with the loss of information about the connectivity between PTMs. TDP on the other hand is reliable and comprehensive for all types of PTMs without prior knowledge and with full sequence coverage. The downside to this technique is that it may require rigorous pre-separation and purification dependent upon the sample and target analyte[35,36,39,40]. More importantly, TDP still faces many challenges related to the measurement of large proteoforms (>30 kDa) from complex samples (i.e., CRC cells). This is due to their much wider charge state distributions and more isotopic peaks in each charge state compared to small proteoforms, leading to substantially lower signal-to-noise ratios. Many critical proteins related to CRC are larger than 30 kDa, *e.g.*, DNA mismatch repair proteins (MSH2), EPCAM, and TP53. Therefore, the development of new TDP techniques for large proteoforms is fundamental to advance our understanding of CRC progression and biomarker discovery.

This study attempts to help fill in the gap of proteoform biomarker knowledge for CRC by developing an MS-based TDP technique for the characterization of large proteoforms in CRC cells. A number of separation and purification techniques before MS analysis will be considered. This includes the use of size-exclusion chromatography (SEC) to reduce the complexity of samples into fractions. Capillary zone electrophoresis (CZE)-MS will also be used as another

dimension of separation. CZE-MS conditions, i.e. compositions of background electrolyte (BGE) of CZE and CE-MS sheath buffer and the automatic gain control (AGC) target will be optimized in an effort to target and identify proteoforms that are larger than 30 kilodaltons (kDa) in size with a standard protein mixture. The optimized system will then be used to analyze a much more complex samples: CRC cell lines SW480 and SW620. Here, the entire workflow will be judged for its effectiveness when applied to the CRC stand-in.

CAPILLARY ZONE ELECTROPHORESIS – MASS SPECTROMETRY OPTIMIZATION

Capillary electrophoresis (CE) is an attractive technique for the separation of proteoforms before introduction to mass spectrometry (MS). This is due to the fact that the separation environment is aqueous and the separation times are rapid[41–43]. Fundamentally, CE is a separation technique that takes advantage of an analyte's electrophoretic mobility when subjected to an applied high voltage. These high voltages have the ability to generate an electroosmotic flow (EOF) of BGE and analyte species within a capillary. The most basic setup for this instrument involves a fused silica capillary, usually 20-200 μm inner diameter/360 μm outer diameter, a high voltage supply, buffer reservoirs of BGE, electrodes, and a detection system (Figure 7)[44].
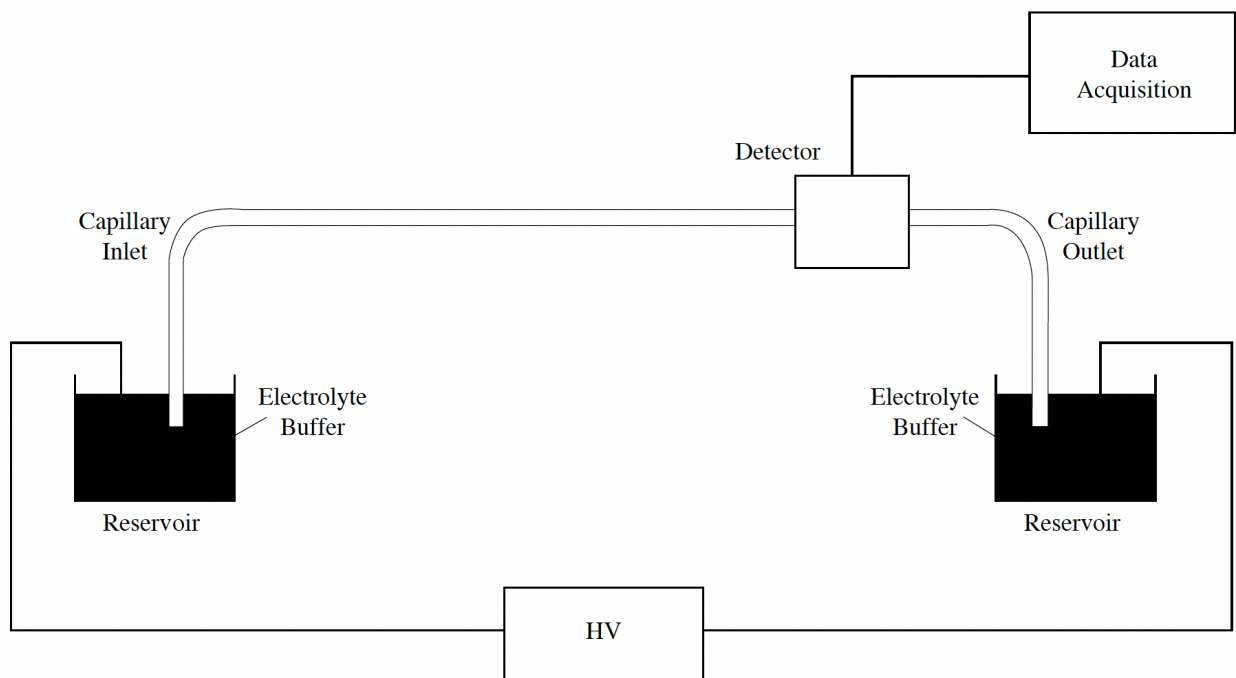

Figure 7. The basic set-up for a CE system. Figure adapted from reference [43].

The setup configuration in Figure 7 is mostly conserved for MS. The main difference is that the second buffer reservoir is adjusted to accommodate ESI, as depicted in Figure 4[31]. This reservoir

contains what is known as a sheath buffer which acts to stabilize the electrical connection for the outlet electrode of the CE[45].

CE is a diverse technique that has several different modes of operation depending on goal and conditions[41–44]. The simplest form of CE is capillary zone electrophoresis (CZE). After sample injection and application of voltage, the contents of the sample mixture are separated into zones based upon their charge-to-size ratios (Figure 8). Important components of this method include the BGE, sheath buffer for ESI, and electric field strength[44].
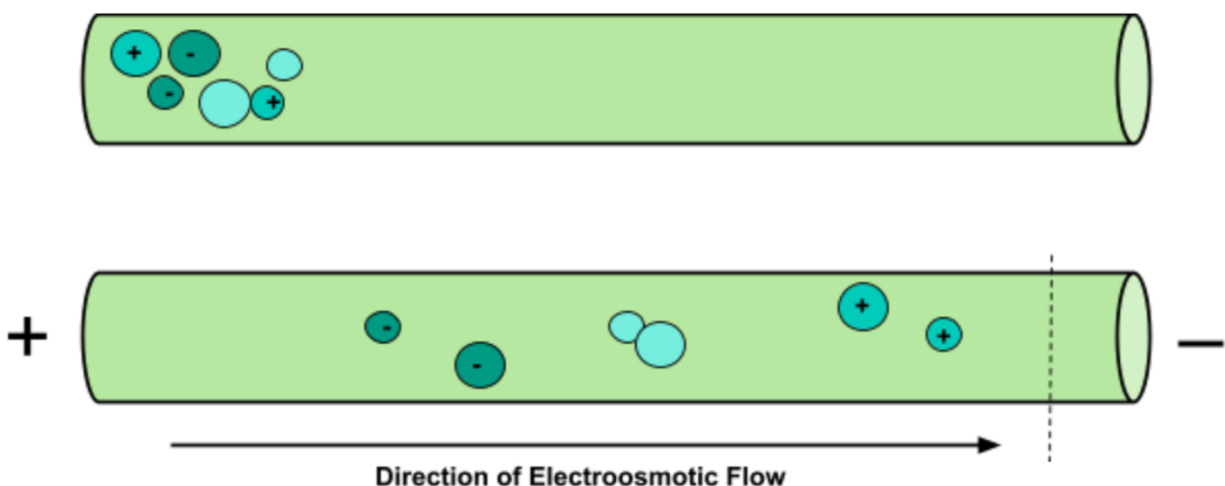

Figure 8. Diagram depicting CZE. Blue spheres may represent analyte. Each blue sphere is assigned a charge denoted by the plus, minus, or no symbol. A voltage being applied across the capillary is represented by the plus and minus symbols on either end.

For protein analysis, the typical BGE is 5% (v/v) acetic acid (AA) in liquid chromatography (LC) grade water, the sheath buffer is 10% (v/v) methanol (MeOH) and 0.2% (v/v) formic acid (FA) in LC grade water, and the voltage applied is 30 kV across the length of the capillary[46]. The average capillary is 1 m long and has an inner diameter of 50 μm.

Background Electrolyte Optimization

Previous studies have shown that the introduction of organic solvents to the BGE of their CZE-MS experiments has had favorable results[47,48]. The CZE-MS system was optimized to include an

organic solvent in the BGE on the grounds that it may enhance the solubility of more

hydrophobic proteoforms, induce conformational changes in proteoforms to reduce the loss due

to capillary adsorption, and improve the signal intensity by making the BGE more readily

volatile. The available literature was consulted to choose what organic solvent to introduce to the

established 5% (v/v) AA BGE. Staub et al. reports testing several organic solvents as add-ins to

their 75 mM ammonium formate buffers. This includes MeOH, ethanol, and acetonitrile (ACN)

ranging in percentages from 5 to 60 percent (v/v). The study presents its findings through the

context of their ACN studies because it "gave a maximal effect at a minimal concentration"[48].

All reported organic solvents had similar effects but at almost double the concentration of ACN.

Using Staub et al. as a guide, various concentrations of ACN from 0 to 30% (v/v) were tested in

addition to our standard BGE content of 5% AA (v/v) in water. Concentrations above 30% (v/v)

ACN were not tested due to the concern of precipitating hydrophilic proteoforms from solution.

A typical methanol-chloroform precipitation procedure will contain 40-90% organic at any given

time, thus these experiments will not approach that concentration[49]. A standard protein mixture

containing 0.7 mg/mL bovine serum albumin (BSA, 66 kDa), 0.3 mg/mL carbonic anhydrase

(CA, 29 kDa), 0.2 mg/mL myoglobin (MyO, 17 kDa), and 0.05 mg/mL ubiquitin (Ubq, 8 kDa)

in 50 mM ammonium bicarbonate (ABC, pH=8.0) was prepared for the experiment. All proteins

and materials were purchased from Sigma-Aldrich (St. Louis, MO). These proteins are standard

in the Sun lab and are often used to assess capillary performance. To determine the degree of

improvement that the amount of introduced organic solvent provides, two parameters were

assessed: resolution between the standard proteins and measured instrument signal intensity. For

an analyte peak to be considered fully resolved the resolution must be 1.5 or greater using

equation 1 where R is the resolution, $t_{R,1}$ and $t_{R,2}$ are the retention times for each peak on the condition that $t_{R1} < t_{R2}$, and $W_{0.5h,1}$ and $W_{0.5h,2}$ are the full width at half maximum of each peak.

$$R = 1.18 \times \frac{t_{R,2} - t_{R,1}}{W_{0.5h,1} + W_{0.5h,2}} \tag{1}$$

Figure 9 displays the result of four measured concentrations of ACN (0%, 10%, 20%, and 30% (v/v)) in BGE solution and its effect on sample separation.
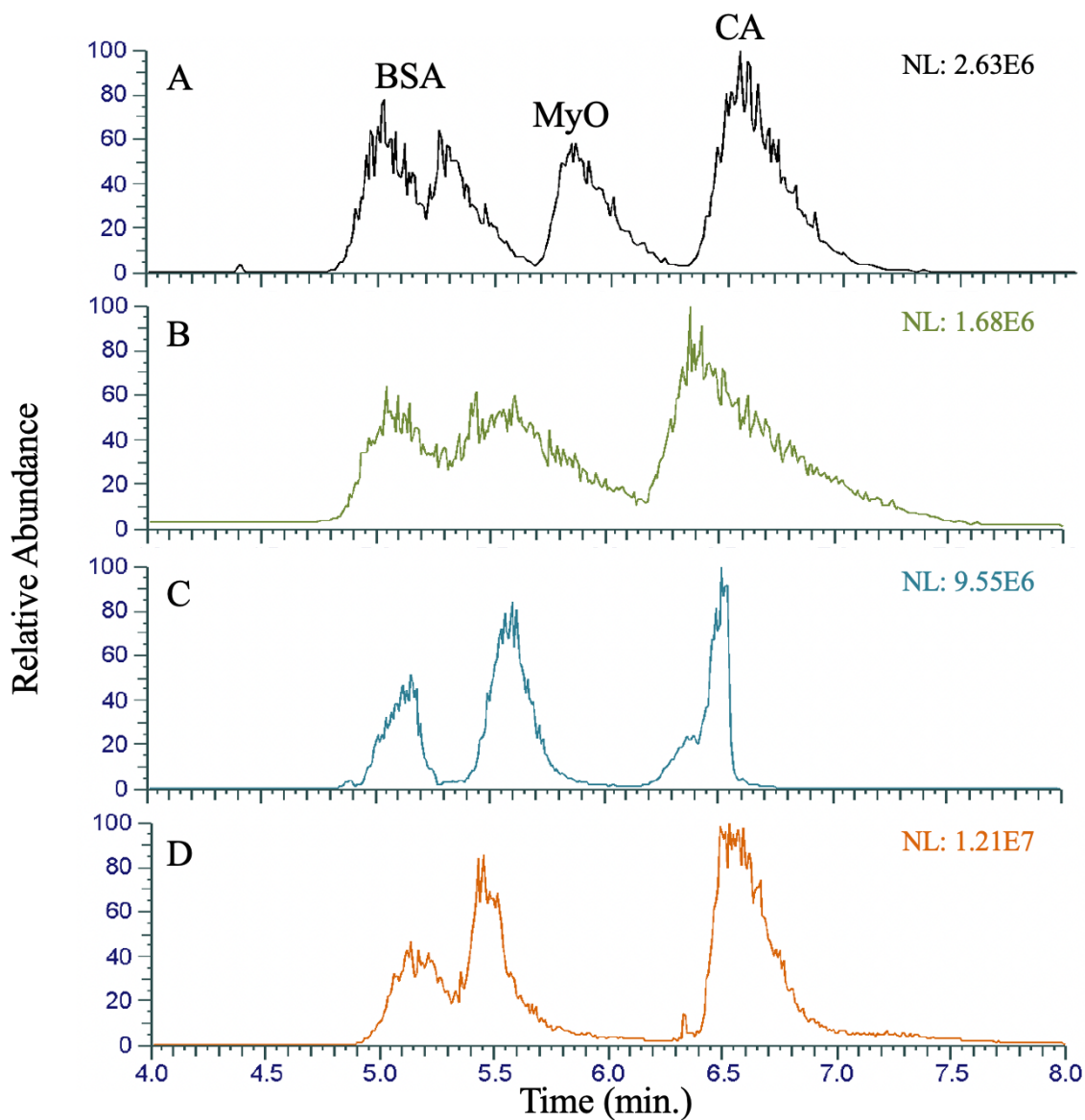
Figure 9. Electropherograms displaying the results of conditions where the BGE was 5% (v/v) AA and (A) 0% (v/v) ACN, (B) 10% (v/v) ACN, (C) 20% (v/v) ACN, and (D) 30% (v/v) ACN. 50 nL of the sample was manually injected on a 50 cm capillary with 10 psi. For each electropherogram, the peak order is the same from left to right: BSA, MyO, and CA. Ubq is unresolved from any peak. "NL" is the normalization level which corresponds to base peak intensity.

Each of these experiments was repeated in triplicate to ensure reliable and reproducible results.

A summary of the signal intensity alongside the resolution between distinguishable proteins is

summarized in Table 1.

Table 1. Table of results for the average intensity and resolution for experiments where organic solvent is introduced to the BGE. "-" indicates that the value was not calculated.

Percent ACN (v/v) added to BGE	Average Intensity	Average Resolution Between BSA and MyO	Average Resolution Between MyO and CA
0	$3.14 \pm 0.77 \times 10^6$	1.0 ± 0.1	1.5 ± 0.1
10	$1.94 \pm 0.94 \times 10^6$	-	-
20	$7.89 \pm 2.61 \times 10^6$	1.4 ± 0.2	3.4 ± 1.4
30	$12.2 \pm 3.26 \times 10^6$	1.1 ± 0.2	3.6 ± 0.9

For the control experiment utilizing no ACN, the average signal intensity is 3.14×10^6. As the concentration of ACN increases, so too do the signal intensities for the runs where concentrations are 20% and 30% (v/v) ACN, being 7.89×10^6 and 1.22×10^7 respectively. This is expected as with a more volatile BGE, an improvement in analyte ionization should be observed. However, the more interesting result of this experiment is the clear improvement in resolution for standard proteins analyzed using an add-in of 20% (v/v) ACN for the BGE. In the control experiment, the resolution between BSA and MyO is on average 1.0 and the resolution between MyO and CA is on average 1.5. For the 20% (v/v) ACN experiments, the resolution is (on average) 1.4 between BSA and MyO and 3.4 between MyO and CA. This means that with the addition of 20% (v/v) ACN, the standard protein peaks were almost entirely resolved from one another. The other percentages of ACN fail to consistently reach conditions where full resolution is achievable. Therefore, the optimized condition for the BGE is 5% (v/v) AA and 20% (v/v) ACN in water. This condition will continue to be used throughout future experiments in this study.

Sheath Buffer Optimization

Previous studies have also shown that the introduction of organic solvent in the CZE-MS sheath buffer has its benefits[45]. Including organic solvents in the sheath liquid is essential. The organic content helps to provide a medium for the stabilization of the electrical connection for the outlet electrode of the CE as well as enhancing electrospray and current stability. A recent literature review notes that typically studies either use MeOH or isopropyl alcohol in a concentrations between 20-80% (v/v)[45]. Currently, the Sun lab already uses a sheath buffer organic solvent concentration of 10% (v/v) MeOH. Using Klampfl and Himmelsbah as a guide for how high the concentration of organic in our sheath buffer can be increased, concentrations of up to 40% (v/v) MeOH were tested. Again, higher concentrations of organic solvent are avoided to limit the potential of hydrophilic proteoform precipitation. Isopropyl alcohol was not chosen for testing to reduce the number of variables when the current system already exists using MeOH. The same standard protein mixture was used as in the BGE experiments (0.7 mg/mL BSA, 0.3 mg/mL CA, 0.2 mg/mL MyO, and 0.05 mg/mL Ubq in 50 mM ABC (pH=8.0)) and the same parameters for assessment were used as well (protein peak resolution and instrument signal intensity). Figure 10 shows the result of four measured concentrations of MeOH (10%, 20%, 30%, and 40% (v/v)) in the sheath liquid and its effect on sample separation.
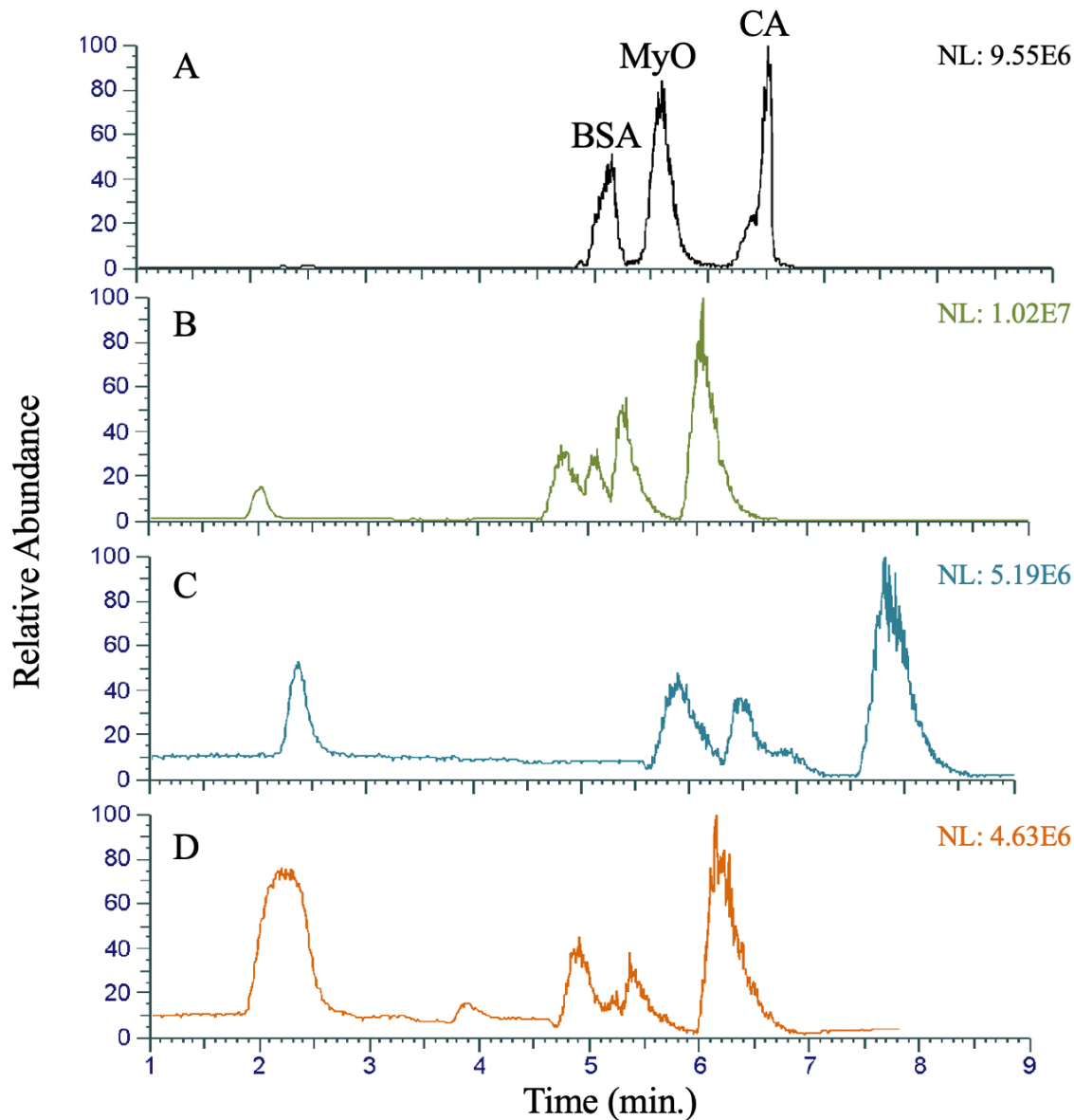
Figure 10. Electropherograms displaying the results of conditions where the sheath buffer was 0.2% (v/v) FA (A) 10% (v/v) MeOH, (B) 20% (v/v) MeOH, (C) 30% (v/v) MeOH, and (D) 40% (v/v). 50 nL of the sample was manually injected on a 50 cm capillary with 10 psi. For electropherogram (A), the peak order is BSA, MyO, and CA from left to right. For electropherograms (B), (C), and (D) the peak order is salt, BSA, MyO, and CA. Ubq is unresolved from any peak. "NL" is the normalization level which corresponds to base peak intensity.

Each of these experiments was repeated in triplicate to ensure reliable and reproducible results.

A summary of the signal intensity alongside the resolution between proteins is summarized in

Table 2.

Table 2. Table of results for the average intensity and resolution for experiments where organic solvent is introduced to the sheath buffer.

Sheath Buffer Condition	Average Intensity	Average Resolution Between BSA and MyO	Average Resolution Between MyO and CA
10% (v/v) MeOH	7.89 ± 2.61 x 10^6	1.4 ± 0.2	3.4 ± 1.4
20% (v/v) MeOH	7.92 ± 2.10 x 10^6	1.2 ± 0.7	2.7 ± 0.5
30% (v/v) MeOH	3.40 ± 1.59 x 10^6	1.3 ± 0.1	2.8 ± 0.1
40% (v/v) MeOH	4.84 ± 0.93 x 10^6	0.9 ± 0.3	2.1 ± 0.2

As discussed previously, the control experiment utilizing 5% (v/v) AA and 20% (v/v) ACN for the BGE and 10% MeOH for the sheath buffer's average intensity is 7.89 x 10^6. As the concentration of MeOH is increased over consecutive experiments the average intensity decreases. This may be the result of the interaction between the sample buffer, 50 mM ABC (pH=8.0), and the organic solvent causing it to precipitate and decrease the ionization of other analyte species. This effect is observed in Figure 10B, C, and D. This same trend is observed when assessing the increased organic's protein peak resolution. As a reminder, the 20% (v/v) ACN for the BGE and 10% MeOH for the sheath buffer's experiments resolution is (on average) 1.4 between BSA and MyO and 3.4 between MyO and CA. Increasing the sheath buffer's concentration of MeOH in this scenario had no improvement on the resolution. In fact, the ABC salt peak continued to grow with each increase in MeOH as interference. Therefore, the optimized condition for sheath buffer is 10% (v/v) MeOH and 0.2% (v/v) FA in water. This condition will continue to be used throughout future experiments in this study.

Automatic Gain Control Target Optimization

The AGC, or automatic gain control, refers to the number of ions collected in a trap and allowed into the mass analyzer at any given time. The AGC parameter for a mass spectrometer is an important one to consider when optimizing the system with a specific goal in mind. The number of ions allowed to enter the mass analyzer at once should be controlled to mitigate space charge effects and improve mass accuracy[50]. The last point is especially important when attempting to analyze very large biomolecules specifically. Large biomolecules like some proteoforms suffer from a low signal-to-noise ratio. The larger the ion the more likely it is to have a higher charge. Thus, the higher the charge state, the larger the charge state distribution with competing isotopic peaks per charge state, which lowers the signal-to-noise ratio. It has been shown that increasing the AGC too high results in a lower number of identifications for protein families, and it may result in some deviations in mass accuracy due to space charging effects[50,51]. Therefore, in theory, lowering the AGC target will have the opposite effect. This will be an interesting parameter to test alongside performing low-resolution MS1 scans for targeting large proteoforms. Reducing the resolution for a highly charged species will allow for a single signal to be produced for each charge state[52].

Thus, the AGC target will be optimized using a complex sample of yeast lysate that contains a wide range of proteoform masses similar to humans. Originally, the AGC target is set to 3 x 10^6 and data is acquired using high-resolution MS1 (480,000) (Figure 11B).
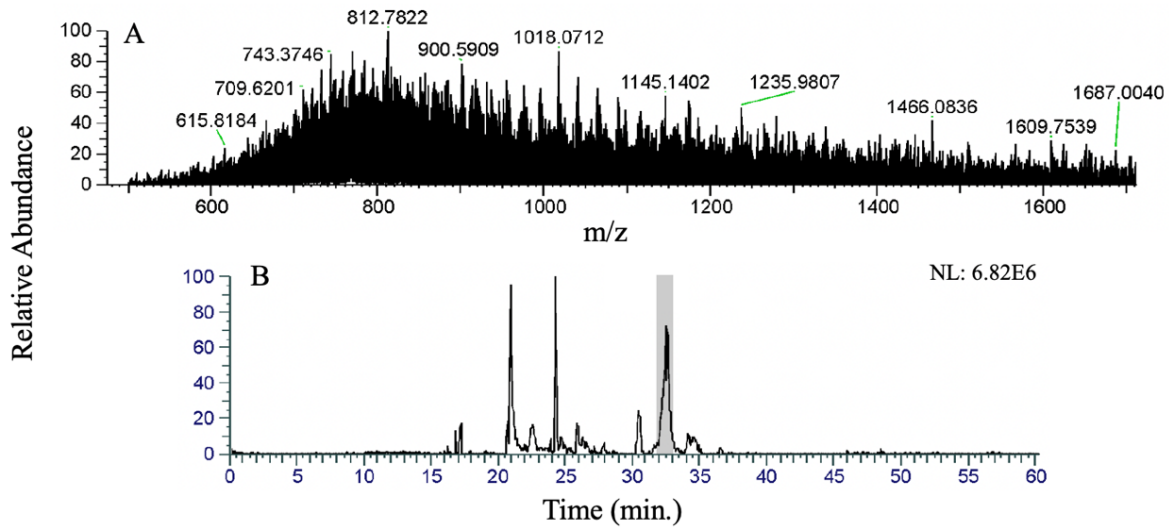
Figure 11. (A) Mass spectrum extracted from the highlighted region of (B) the electropherogram for the separation of yeast lysate. Yeast was lysed in Dulbecco's phosphate buffered saline (DPBS) and buffer exchanged into 50 mM ABC (pH=8.0). 150 ng of sample was pressure injected onto the capillary.

The separation of proteins in Figure 11B looks clear. However, when the MS data is investigated further the resolution for more highly charged species is poor (Figure 11A). To improve this, data acquisition was switched into a low-resolution (7500) mode to make charge states for large and highly charged proteoforms more clear. These results are shown in Figure 12 alongside experiments where the data acquisition was kept in low-resolution mode and the AGC target was lowered to 2 x 10^6 and 1 x 10^6.

Before assessing the spectral data for the results of this experiment, it is still important to consider the intensity of the results. Because fewer ions are being collected for analysis for each scan, the intensity needed to be preserved as much as possible. As shown in Figure 12, the intensity increases as the AGC target lowers. Ions with a lower abundance are detected more clearly.
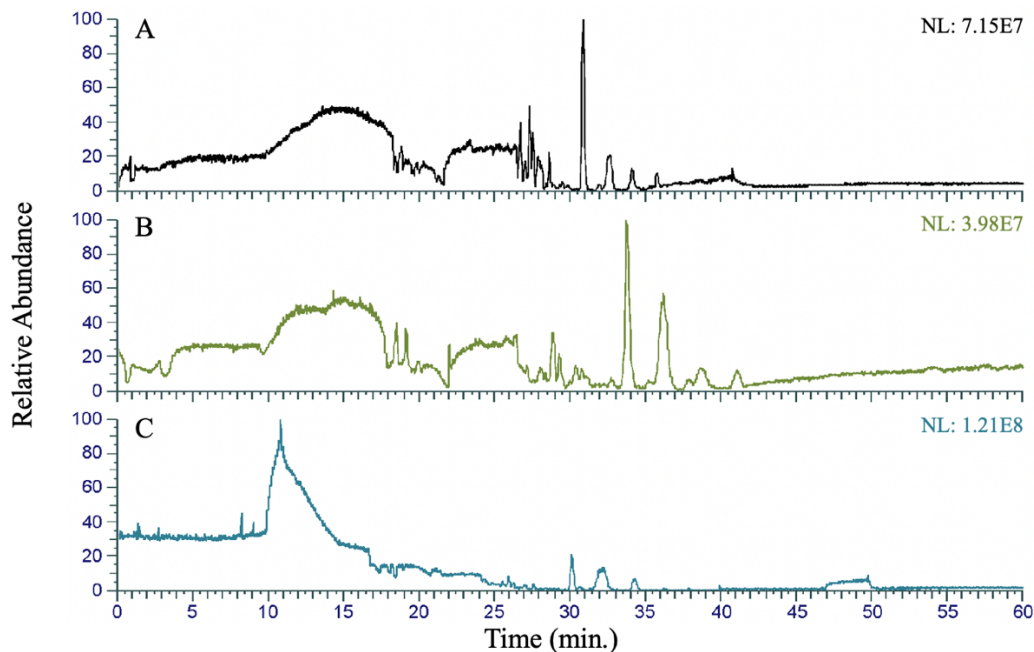
Figure 12. Electropherograms displaying the results of conditions where the instrument was in low-resolution mode and the AGC target was (A) 3×10^6, (B) 2×10^6, and (C) 1×10^6. "NL" is the normalization level which corresponds to base peak intensity.

Figure 13 shows a mass spectrum for a portion of the electropherogram from an experiment where the AGC target is 1×10^6. This highlighted region was chosen specifically because it corresponds to roughly the same area as was chosen in Figure 11.
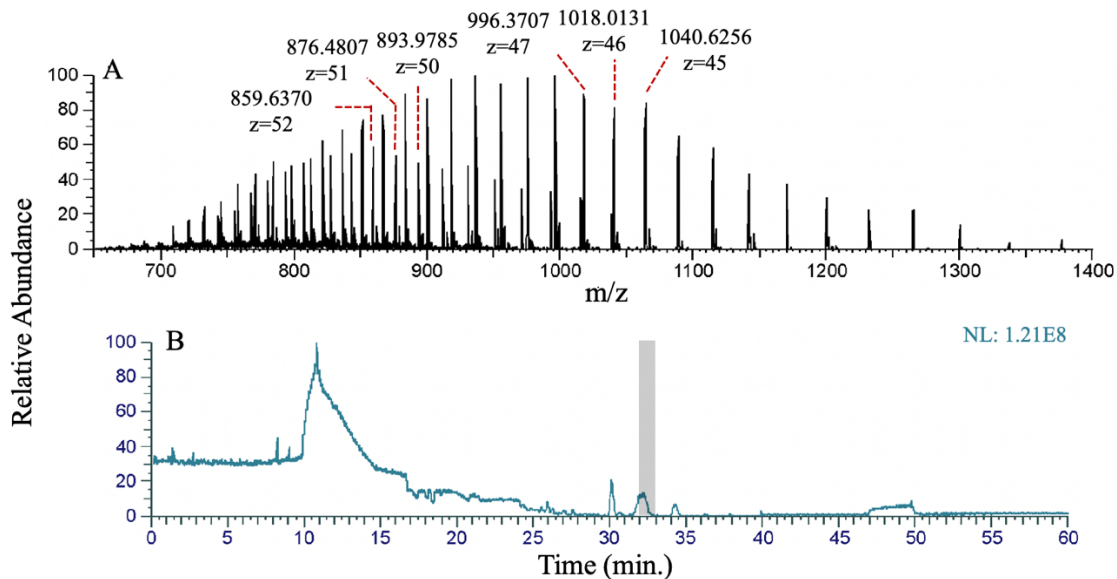
Figure 13. (A) Mass spectrum extracted from the highlighted region of (B) the electropherogram for the separation of yeast lysate under low-resolution mode with the AGC target lowered to 1 x 10^6. The right-most three m/z correspond to a protein that has a mass of approximately 44.7 kDa. The left-most three m/z correspond to a protein that has a mass of approximately 46.8 kDa.

This spectrum clearly shows two proteins that are both over 40 kDa. Without the low-resolution mode and lowered AGC target this was not possible. In order to target large proteoforms while maintaining a high intensity reading, the optimized condition for the AGC target is 1 x 10^6 in low-resolution mode.

Colorectal cancer is a complex disease. Thus, the makeup of its proteome is also complex. When compared to the known proteome of *Saccharomyces cerevisiae*, the human proteome has over 14,000 more known proteins[53,54]. The complexity of the human proteome, in general, make it more difficult to precisely analyze. In the past, it has been difficult to analyze and identify proteoforms over 30 kDa[15,46]. This is why a separation step ahead of CZE will be used for the CRC cell line samples. As the goal is to target large proteoforms, after cell lysis, the cell line samples will be subjected to size exclusion chromatography (SEC). Fundamentally, SEC separates analytes based on their molecular weight. A column for SEC is packed with a porous stationary phase. Larger molecules that cannot enter and interact with the pores of the stationary phase will elute first with the flow of the mobile phase. Smaller analytes that do enter the pores will take longer, and thus separation occurs (Figure 14)[55–57].
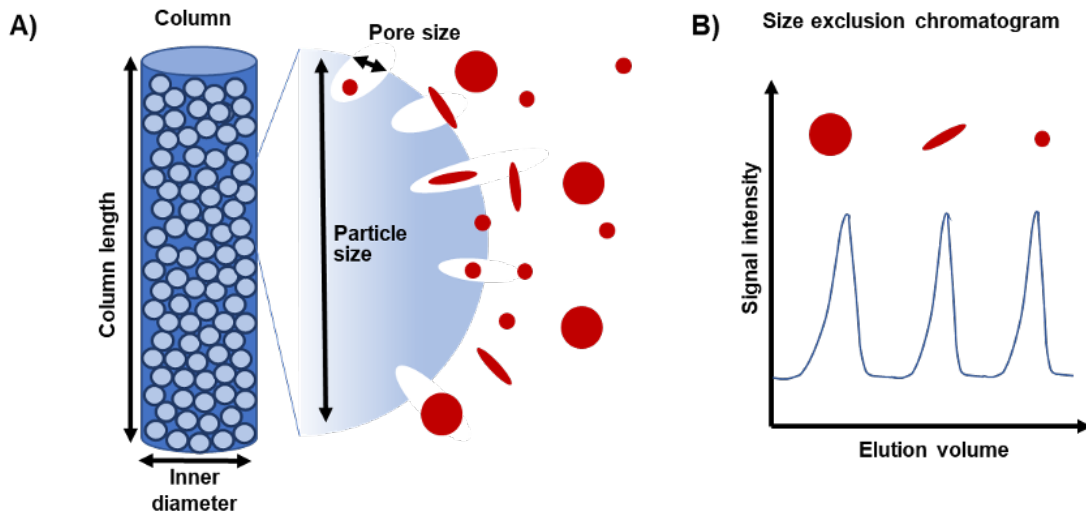

Figure 14. Depiction of (A) an SEC column packed with porous particles and (B) an example chromatogram displaying the elution order of analytes. Figure adapted from reference [53].

This study takes advantage of the separation ability of SEC. Fractions were collected based on the output from the SEC column. As large proteins will elute first, the earliest fractions will be

collected for analysis. The cell lines used in this study are human CRC cell lines SW620 and SW480. These cell lines are derived from metastatic and non-metastatic CRC tumors respectively[46].

Each fractionation was completed in triplicate for each cell line to evaluate consistency. Fraction 1 for each cell line was collected in the 10-12 minute time frame. Each next fraction was collected every two minutes in succession until 20 minutes had elapsed (Figure 15).
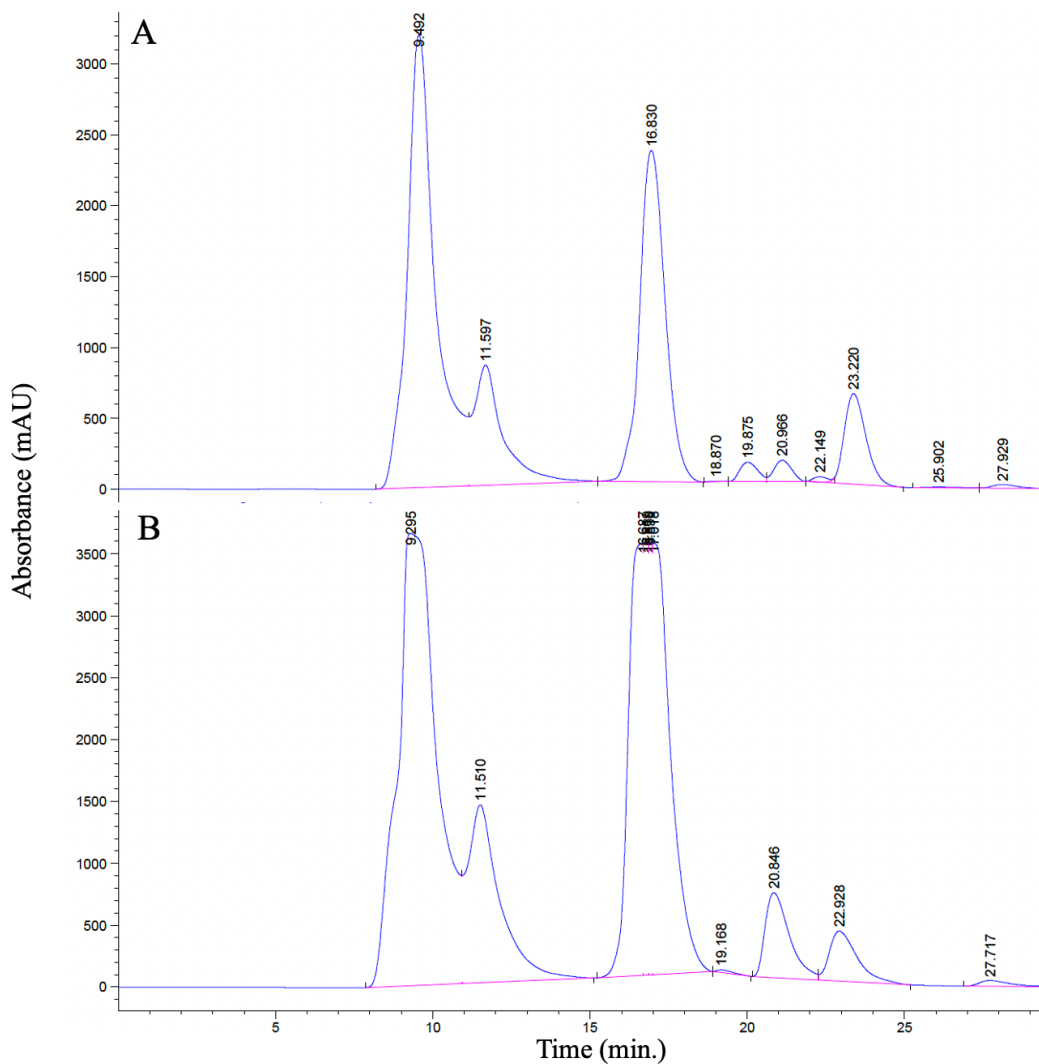

Figure 15. SEC chromatograms for (A) SW480 and (B) SW620. Each separation was performed on an SEC-3, 150 Å pore size, 3 μm particle, 30 cm length column. The flow rate was 0.2 mL/min DPBS and fractions were collected every two minutes from 10-20 min.

The sample was buffer exchanged into 50 mM ABC (pH=8.0) and analyzed using the previously optimized system (Figure 16).
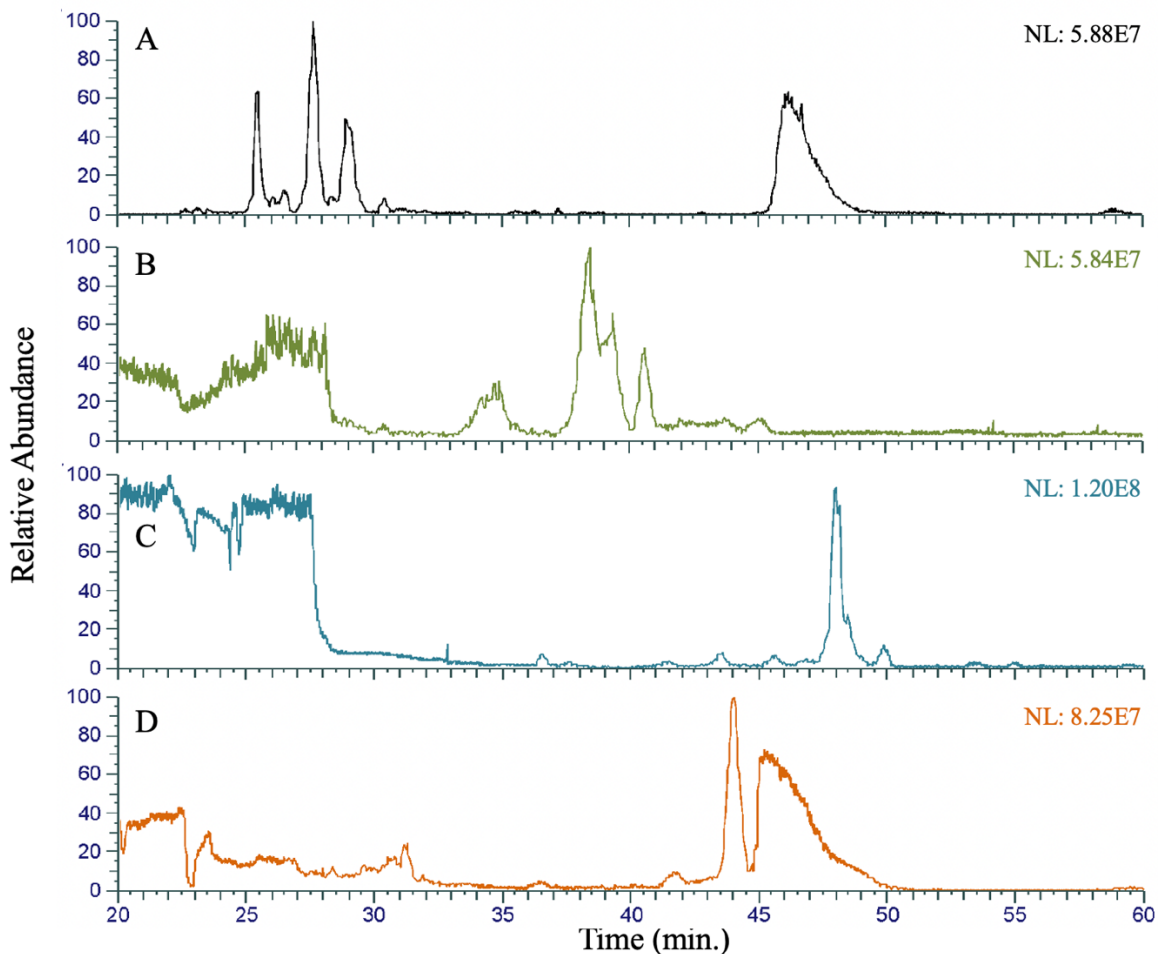
Figure 16. Electropherograms displaying the separation of fraction 1 for CRC cell lines (A,B) SW480 and (C,D) SW620. 50 ng of each sample was pressure injected onto the capillary. "NL" is the normalization level which corresponds to base peak intensity.

Unfortunately, reproducibility was an issue during these experiments. The lack of integrity of the capillary coating over time and consecutive runs is suspected. Another potential issue is the interference of leftover DPBS since it is a non-volatile salt. If any has been left over through the buffer exchange and injected into the mass spectrometer, it will give the effect shown at the beginning of the electropherograms and quickly dirty the instrumentation. Despite the absence of

reproducibility, there were a few proteoforms over 30 kDa that can be detected clearly. These

proteoforms are reported in Table 3.

Table 3. Table of results for the manual mass identification for CRC cell lines SW480 and SW620.

Cell Line	Mass (Da)	m/z	m/z	m/z	Charge State of Most Abundant Peak
SW480	46828.5755	699.92	710.55	721.45	+71
	69280.2059	738.02	745.96	754.06	+93
	38514.3537	756.19	771.29	787.02	+50
	39287.9519	802.80	819.51	836.92	+49
	52295.6525	844.50	858.31	872.59	+62
	46463.8174	1011.09	1033.55	1056.99	+45
	39288.1197	756.56	771.37	786.75	+51
	39522.8113	1098.83	1130.19	1163.52	+35
	43604.5632	1148.50	1179.51	1212.24	+37
	62427.9751	1561.71	1601.73	1643.84	+39

Table 3 (cont'd).

	46828.3262	710.55	721.44	732.68	+65
	39288.1467	836.94	855.08	874.08	+46
	52294.4139	872.57	887.35	902.65	+60
	46463.1359	968.99	989.57	1011.09	+47
SW620	32013.4542	1001.43	1033.69	1068.13	+31
	48873.8174	1063.48	1087.09	1111.78	+45
	43604.2562	1015.05	1039.21	1064.53	+43
	56859.4677	1387.83	1422.49	1458.94	+40

Shown in Figure 17 are examples of a few of the mass-identified proteoforms for the SW480

CRC cell line. To obtain clear spectra for the proteoforms, scans were averaged for

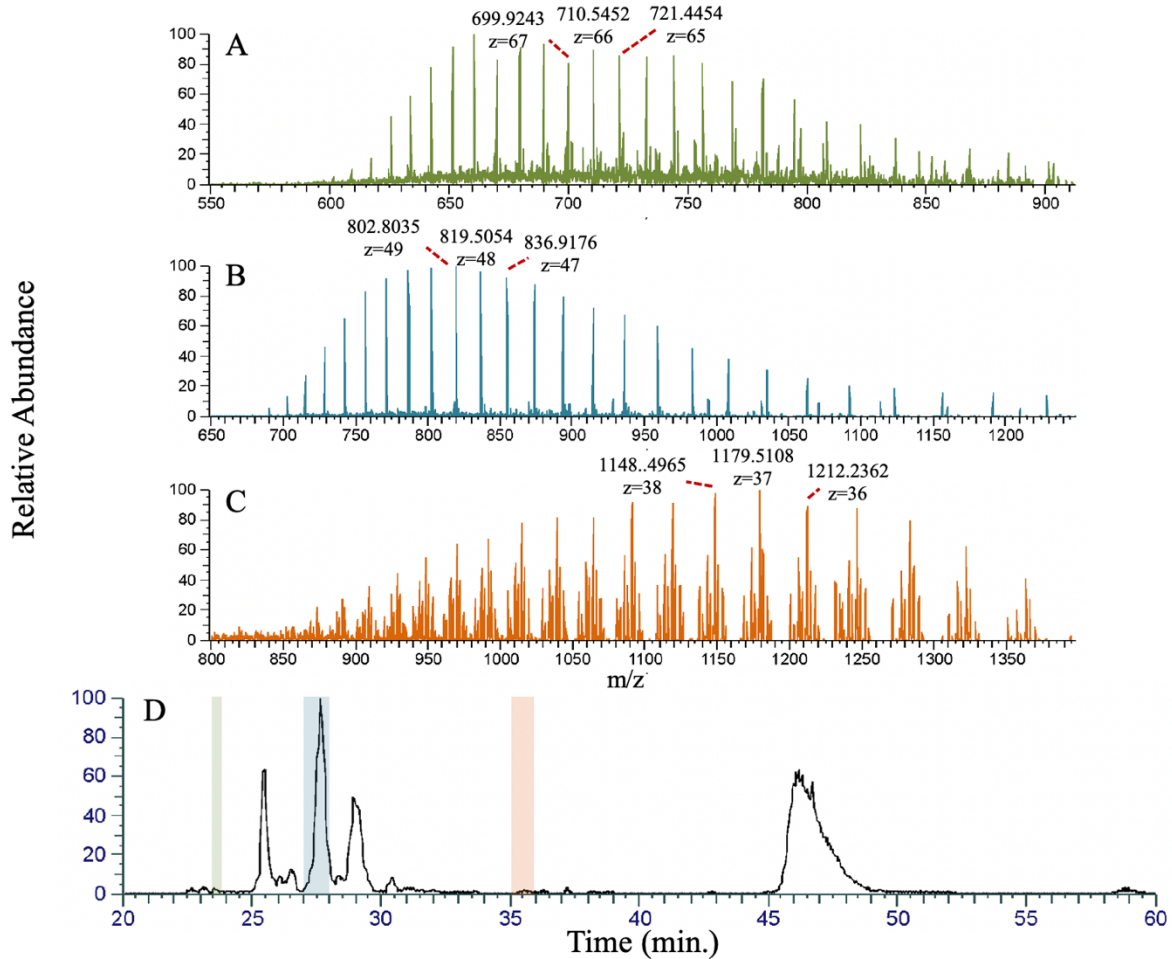approximately one second where proteins were detected.

Figure 17. Mass spectra extracted from the (A) green, (B) blue, and (C) orange highlighted regions of (D) the electropherogram for the separation of SW480 lysate under low-resolution mode with an AGC target of 1 x 10^6, respectively. The proteoform from spectrum (A) is approximately 46.8 kDa while the proteins from spectra (B,C) are 39.3 kDa and 43.6 kDa.

For SW480, 10 different proteoforms over 30 kDa were identified, and for SW620, 8 were

(Table 3). Some of these proteoforms are shared between the two cell lines. It is a good sign that

there are distinguished commonalities between the two cell lines. Yet, there are observable

differences that point to the possibility for proteoform biomarkers in the future.

Another important thing to point out is the visualization of multiple proteoforms for a discovered

protein. In the SW480 spectrum, the 1179.51 m/z, 43.6 kDa protein has multiple other

proteoforms in its immediate family. They are distinguished by having the same isotopic pattern

as the most abundant protein. When zooming in on a specific charge state, there are peaks that have the same charge but lower abundance (Figure 18). These are proteoforms of that highly abundant protein. Likely, they are the product of different PTMs or amino acid variations.
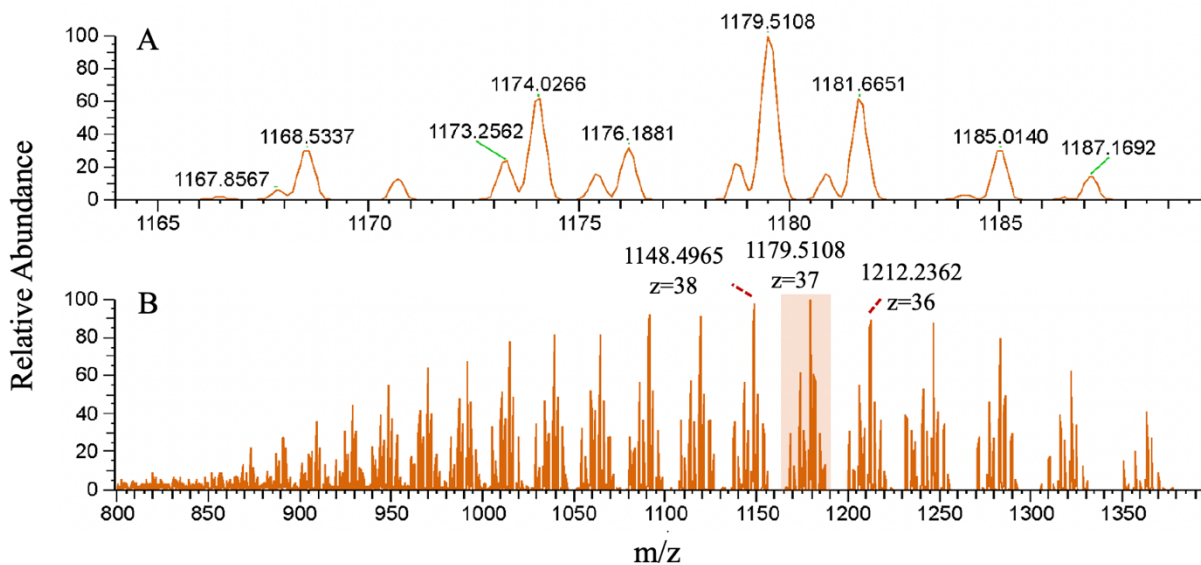

Figure 18. (A) Zoomed in mass spectrum from the orange highlighted region of (B).

FUTURE DIRECTION

MS analysis of proteins over 30 kDa has traditionally suffered from low signal-to-noise ratios due to their high charges and large charge state distributions. This study has provided a workflow to at least mass-identify a few proteoforms by successfully modifying the BGE for CZE and the sheath buffer and AGC target for CZE-MS as well as employing low-resolution MS1 measurement.

However, there is still much room for improvement. One improvement that could be made is to the capillary coating. After several consecutive runs, the linear polyacrylamide coating of the capillary used in our study can be worn down and become less effective. This will result in less reproducibility between runs. Proteins may also adsorb into the capillary wall after initial runs lowering the effectiveness[58]. Alternative coatings to what is currently being used may be necessary. The development of an effective capillary cleanup procedure will also be useful to maintain the effectiveness of CZE separation of large proteoforms[58].

Another suggestion for improvement is the addition of more large proteoform separations. One important option can be related to the use of nanoparticle (NP) protein corona. Protein corona is a layer of protein molecules adsorbed onto the NP surface when NPs are incubated with a complex biological sample (e.g., human plasma). It has been reported that the use of NPs with different surface chemistry can drastically improve the proteome coverage of human plasma via measuring their protein corona using BUP[59,60]. The reason is that the NP surface chemistry can determine the composition of protein corona. We expect that the NP protein corona idea will be an effective protein separation approach by employing various NPs with different surface chemistry to incubate with, i.e., CRC cell lysates, separately.

31

The idea was tested by using carboxyl-modified magnetic NPs and one yeast cell lysate. The NPs and the yeast cell lysate in a DPBS buffer were incubated for 1.25 hours. After that, the protein coronas (proteins bound to the NP surface) were eluted by an 8% sodium dodecyl sulfate (SDS) buffer. The eluted proteins were analyzed by SDS-polyacrylamide gel electrophoresis (SDS-PAGE), Figure 19.
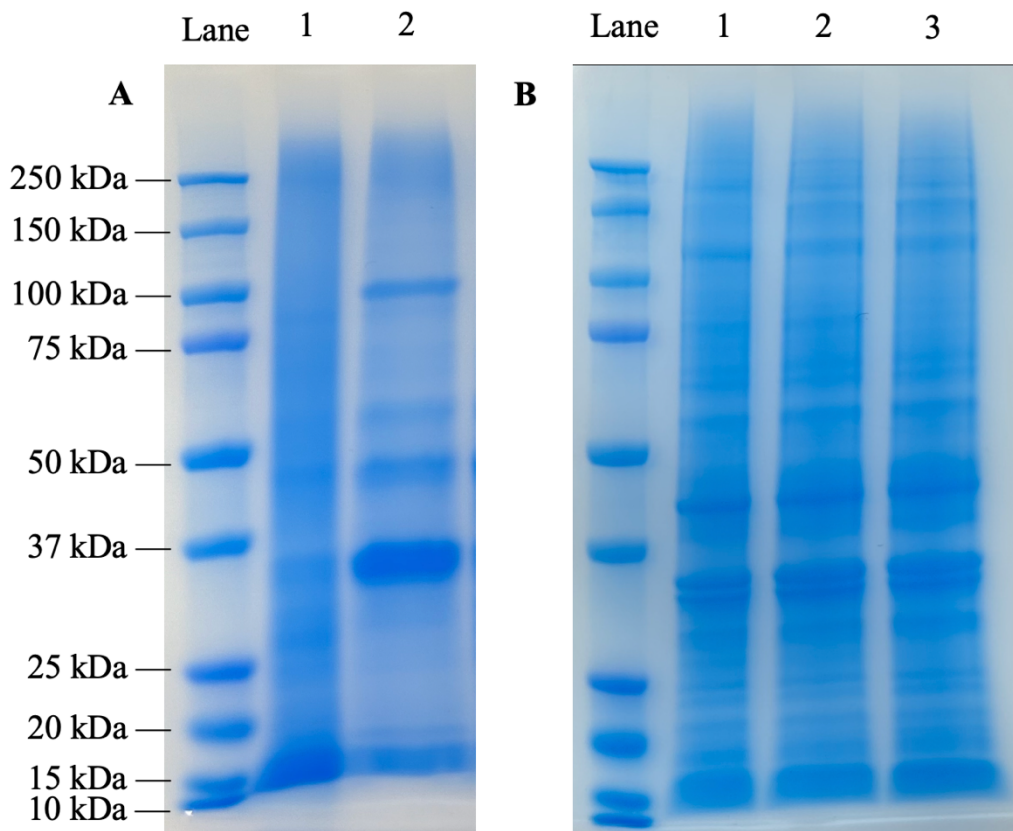

Figure 19. Results of SDS-PAGEs. 20 µg of protein material was loaded per lane. (A) Lane 1: Yeast lysate extracted using DPBS. Lane 2: Extractant from carboxyl-modified magnetic nanoparticles incubated with yeast lysate from DPBS. (B) Lane 1, 2, and 3: Replicates of extractant from carboxyl-modified magnetic nanoparticles incubated with yeast lysate from DPBS.

It is clear that the yeast protein corona sample (Lane 2, Figure 19A) has much more abundant large proteins compared to the yeast cell lysate control (Lane 1, Figure 19A). The phenomenon is most likely due to the sample complexity reduction from the protein corona approach.

Additionally, this approach is highly reproducible evidenced by the consistent SDS-PAGE data in Figure 19B from technical triplicate preparations. The data here suggests that the NP protein corona approach could be a useful strategy to further improve large proteoform characterization in CRC cells.

REFERENCES

[1]Pernot, S.; Terme, M.; Voron, T.; Colussi, O.; Marcheteau, E.; Tartour, E.; Taieb, J. Colorectal Cancer and Immunity: What We Know and Perspectives. *World J. Gastroenterol. WJG* **2014**, *20* (14), 3738–3750. https://doi.org/10.3748/wjg.v20.i14.3738.

[2]Manu, K. A.; Bronte, F.; Giunta, E. F. Editorial: Reviews in Gastrointestinal Cancers. *Front. Oncol.* **2023**, *13*. https://doi.org/10.3389/fonc.2023.1252665.

[3]National Cancer Institute. *Colon Cancer Treatment*. https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq (accessed 2024-04-30).

[4]American Cancer Society. *Colorectal Cancer Stages*. https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/staged.html (accessed 2024-04-30).

[5]Levin, B.; Lieberman, D. A.; McFarland, B.; Andrews, K. S.; Brooks, D.; Bond, J.; Dash, C.; Giardiello, F. M.; Glick, S.; Johnson, D.; Johnson, C. D.; Levin, T. R.; Pickhardt, P. J.; Rex, D. K.; Smith, R. A.; Thorson, A.; Winawer, S. J.; American Cancer Society Colorectal Cancer Advisory Group; US Multi-Society Task Force; American College of Radiology Colon Cancer Committee. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology* **2008**, *134* (5), 1570–1595. https://doi.org/10.1053/j.gastro.2008.02.002.

[6]Lieberman David A. Screening for Colorectal Cancer. *N. Engl. J. Med.* **2009**, *361* (12), 1179–1187. https://doi.org/10.1056/NEJMcp0902176.

[7]Alves Martins, B. A.; de Bulhões, G. F.; Cavalcanti, I. N.; Martins, M. M.; de Oliveira, P. G.; Martins, A. M. A. Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research. *Front. Oncol.* **2019**, *9*. https://doi.org/10.3389/fonc.2019.01284.

[8]National Institute of Diabetes and Digestive and Kidney Diseases. *Colonoscopy*. https://www.niddk.nih.gov/health-information/diagnostic-tests/colonoscopy (accessed 2024-04-30).

[9]Vitrosens. *What Is Fecal Occult Blood and How to Use the Fecal Occult Blood Test Kit (FOBT)?* https://vitrosens.com/what-is-fecal-occult-blood-and-how-to-use-the-fecal-occult-blood-test-kit-fobt/ (accessed 2024-04-30).

[10]Lim, L. C.; Lim, Y. M. Proteome Heterogeneity in Colorectal Cancer. *PROTEOMICS* **2018**, *18* (3–4), 1700169. https://doi.org/10.1002/pmic.201700169.

[11]Pitule, P.; Čedíková, M.; Třeška, V.; Králíčková, M.; Liška, V. Assessing Colorectal Cancer Heterogeneity: One Step Closer to Tailored Medicine. *J. Appl. Biomed.* **2013**, *11* (3), 115–129. https://doi.org/10.2478/v10136-012-0035-6.

[12]Lu, Y.-W.; Zhang, H.-F.; Liang, R.; Xie, Z.-R.; Luo, H.-Y.; Zeng, Y.-J.; Xu, Y.; Wang, L.-M.; Kong, X.-Y.; Wang, K.-H. Colorectal Cancer Genetic Heterogeneity Delineated by Multi-Region Sequencing. *PLOS ONE* **2016**, *11* (3), e0152673. https://doi.org/10.1371/journal.pone.0152673.

[13]Muzny, D. M. et al. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* **2012**, *487* (7407), 330–337. https://doi.org/10.1038/nature11252.

[14]Budinska, E.; Popovici, V.; Tejpar, S.; D'Ario, G.; Lapique, N.; Sikora, K. O.; Di Narzo, A. F.; Yan, P.; Hodgson, J. G.; Weinrich, S.; Bosman, F.; Roth, A.; Delorenzi, M. Gene Expression Patterns Unveil a New Level of Molecular Heterogeneity in Colorectal Cancer. *J. Pathol.* **2013**, *231* (1), 63–76. https://doi.org/10.1002/path.4212.

[15]Yin, X.; Zhang, Y.; Guo, S.; Jin, H.; Wang, W.; Yang, P. Large Scale Systematic Proteomic Quantification from Non-Metastatic to Metastatic Colorectal Cancer. *Sci. Rep.* **2015**, *5* (1), 12120. https://doi.org/10.1038/srep12120.

[16]Walther, A.; Johnstone, E.; Swanton, C.; Midgley, R.; Tomlinson, I.; Kerr, D. Genetic Prognostic and Predictive Markers in Colorectal Cancer. *Nat. Rev. Cancer* **2009**, *9* (7), 489–499. https://doi.org/10.1038/nrc2645.

[17]Oxford English Dictionary. *Biomarker*. https://www.oed.com/dictionary/biomarker_n (accessed 2024-04-29).

[18]Marmol, I.; Sánchez-de-Diego, C.; Pradilla Dieste, A.; Cerrada, E.; Rodriguez Yoldi, M. J. Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *Int. J. Mol. Sci.* **2017**, *18* (1), 197. https://doi.org/10.3390/ijms18010197.

[19]Brennetot, C.; Buhard, O.; Jourdan, F.; Flejou, J.-F.; Duval, A.; Hamelin, R. Mononucleotide Repeats BAT-26 and BAT-25 Accurately Detect MSI-H Tumors and Predict Tumor Content: Implications for Population Screening. *Int. J. Cancer* **2005**, *113* (3), 446–450. https://doi.org/10.1002/ijc.20586.

[20]Faulkner, R. D.; Seedhouse, C. H.; Das-Gupta, E. P.; Russell, N. H. BAT-25 and BAT-26, Two Mononucleotide Microsatellites, Are Not Sensitive Markers of Microsatellite Instability in Acute Myeloid Leukaemia. *Br. J. Haematol.* **2004**, *124* (2), 160–165. https://doi.org/10.1046/j.1365-2141.2003.04750.x.

[21]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. Precise Characterization of KRAS4b Proteoforms in Human Colorectal Cells and Tumors Reveals Mutation/Modification Cross-Talk. *Proc. Natl. Acad. Sci.* **2018**, *115* (16), 4140–4145. https://doi.org/10.1073/pnas.1716122115.

[22]Adams, L. M.; DeHart, C. J.; Drown, B. S.; Anderson, L. C.; Bocik, W.; Boja, E. S.; Hiltke, T. M.; Hendrickson, C. L.; Rodriguez, H.; Caldwell, M.; Vafabakhsh, R.; Kelleher, N. L. Mapping the KRAS Proteoform Landscape in Colorectal Cancer Identifies Truncated KRAS4B That Decreases MAPK Signaling. *J. Biol. Chem.* **2023**, *299* (1), 102768. https://doi.org/10.1016/j.jbc.2022.102768.

[23]Daneshpour, M.; Ghadimi-Daresajini, A. Overview of miR-106a Regulatory Roles: From Cancer to Aging. *Bioeng. Basel Switz.* **2023**, *10* (8), 892. https://doi.org/10.3390/bioengineering10080892.

[24]Pan, Y.-J.; Wei, L.-L.; Wu, X.-J.; Huo, F.-C.; Mou, J.; Pei, D.-S. MiR-106a-5p Inhibits the Cell Migration and Invasion of Renal Cell Carcinoma through Targeting PAK5. *Cell Death Dis.* **2017**, *8* (10), e3155–e3155. https://doi.org/10.1038/cddis.2017.561.

[25]Wang, D. Discrepancy between mRNA and Protein Abundance: Insight from Information Retrieval Process in Computers. *Comput. Biol. Chem.* **2008**, *32* (6), 462–468. https://doi.org/10.1016/j.compbiolchem.2008.07.014.

[26]Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.; Davies, S. R.; Wang, S.; Wang, P.; Kinsinger, C. R.; Rivers, R. C.; Rodriguez, H.; Townsend, R. R.; Ellis, M. J. C.; Carr, S. A.; Tabb, D. L.; Coffey, R. J.; Slebos, R. J. C.; Liebler, D. C. Proteogenomic Characterization of Human Colon and Rectal Cancer. *Nature* **2014**, *513* (7518), 382–387. https://doi.org/10.1038/nature13438.

[27]Hvidsten, T. R.; Lægreid, A.; Kryshtafovych, A.; Andersson, G.; Fidelis, K.; Komorowski, J. A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity. *PLoS ONE* **2009**, *4* (7), e6266. https://doi.org/10.1371/journal.pone.0006266.

[28]Han, X.; Aslanian, A.; Yates, J. R. Mass Spectrometry for Proteomics. *Curr. Opin. Chem. Biol.* **2008**, *12* (5), 483–490. https://doi.org/10.1016/j.cbpa.2008.07.024.

[29]Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003**, *422* (6928), 198–207. https://doi.org/10.1038/nature01511.

[30]Garg, E.; Zubair, M. Mass Spectrometer. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2024.

[31]Banerjee, S.; Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int. J. Anal. Chem.* **2012**, *2012*, e282574. https://doi.org/10.1155/2012/282574.

[32]Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989**, *246* (4926), 64–71. https://doi.org/10.1126/science.2675315.

[33]Geiger, T.; Cox, J.; Mann, M. Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-Ion Fragmentation. *Mol. Cell. Proteomics MCP* **2010**, *9* (10), 2252–2261. https://doi.org/10.1074/mcp.M110.001537.

[34]Zubarev, R. A.; Makarov, A. Orbitrap Mass Spectrometry. *Anal. Chem.* **2013**, *85* (11), 5288–5296. https://doi.org/10.1021/ac4001223.

[35]Smith, L. M.; Agar, J. N.; Chamot-Rooke, J.; Danis, P. O.; Ge, Y.; Loo, J. A.; Paša-Tolić, L.; Tsybin, Y. O.; Kelleher, N. L.; THE CONSORTIUM FOR TOP-DOWN PROTEOMICS. The Human Proteoform Project: Defining the Human Proteome. *Sci. Adv.* **2021**, *7* (46), eabk0734. https://doi.org/10.1126/sciadv.abk0734.

[36]Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; Sun, L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z.; Yu, D.; Shortreed, M. R.; Smith, L. M. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19* (10), e1800361. https://doi.org/10.1002/pmic.201800361.

[37]Bennett, K. L. *A Fusion of Proteomic Practices: The Indisputable Complementarity of "Bottom-Up" and "Top-Down" Approaches*. Proteomics & Metabolomics from Technology Networks. http://www.technologynetworks.com/proteomics/articles/a-fusion-of-proteomic-practices-the-indisputable-complementarity-of-bottom-up-and-top-down-337094 (accessed 2024-04-30).

[38]Heissel, S.; Frederiksen, S. J.; Bunkenborg, J.; Højrup, P. Enhanced Trypsin on a Budget: Stabilization, Purification and High-Temperature Application of Inexpensive Commercial Trypsin for Proteomics Applications. *PLoS ONE* **2019**, *14* (6), e0218374. https://doi.org/10.1371/journal.pone.0218374.

[39]Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445* (4), 683–693. https://doi.org/10.1016/j.bbrc.2014.02.041.

[40]Gregorich, Z. R.; Chang, Y.-H.; Ge, Y. Proteomics in Heart Failure: Top-down or Bottom-Up. *Pflugers Arch.* **2014**, *466* (6), 1199–1209. https://doi.org/10.1007/s00424-014-1471-9.

[41]Harstad, R. K.; Johnson, A. C.; Weisenberger, M. M.; Bowser, M. T. Capillary Electrophoresis. *Anal. Chem.* **2016**, *88* (1), 299–319. https://doi.org/10.1021/acs.analchem.5b04125.

[42]Grossman, P. D.; Colburn, J. C. *Capillary Electrophoresis: Theory and Practice*; Academic Press, 2012.

[43]Voeten, R. L. C.; Ventouri, I. K.; Haselberg, R.; Somsen, G. W. Capillary Electrophoresis: Trends and Recent Advances. *Anal. Chem.* **2018**, *90* (3), 1464–1481. https://doi.org/10.1021/acs.analchem.8b00015.

[44]Beckman Coulter. *Introduction to Capillary Electrophoresis*.

[45]Klampfl, C. W.; Himmelsbach, M. Sheath Liquids in CE-MS: Role, Parameters, and Optimization. In *Capillary Electrophoresis–Mass Spectrometry (CE-MS)*; John Wiley & Sons, Ltd, 2016; pp 41–65. https://doi.org/10.1002/9783527693801.ch3.

[46]McCool, E. N.; Xu, T.; Chen, W.; Beller, N. C.; Nolan, S. M.; Hummon, A. B.; Liu, X.; Sun, L. Deep Top-down Proteomics Revealed Significant Proteoform-Level Differences between Metastatic and Nonmetastatic Colorectal Cancer Cells. *Sci. Adv.* **2022**, *8* (51), eabq6348. https://doi.org/10.1126/sciadv.abq6348.

[47]Han, X.; Wang, Y.; Aslanian, A.; Fonslow, B.; Graczyk, B.; Davis, T. N.; Yates, J. R. I. In-Line Separation by Capillary Electrophoresis Prior to Analysis by Top-Down Mass Spectrometry Enables Sensitive Characterization of Protein Complexes. *J. Proteome Res.* **2014**, *13* (12), 6078–6086. https://doi.org/10.1021/pr500971h.

[48]Staub, A.; Comte, S.; Rudaz, S.; Veuthey, J.-L.; Schappler, J. Use of Organic Solvent to Prevent Protein Adsorption in CE-MS Experiments. *ELECTROPHORESIS* **2010**, *31* (19), 3326–3333. https://doi.org/10.1002/elps.201000245.

[49]Shahinuzzaman, A. D. A.; Chakrabarty, J. K.; Fang, Z.; Smith, D.; Kamal, A. H. M.; Chowdhury, S. M. Improved In-Solution Trypsin Digestion Method for Methanol–Chloroform Precipitated Cellular Proteomics Sample. *J. Sep. Sci.* **2020**, *43* (11), 2125–2132. https://doi.org/10.1002/jssc.201901273.

[50]Kalli, A.; Smith, G. T.; Sweredoski, M. J.; Hess, S. Evaluation and Optimization of Mass Spectrometric Settings during Data-Dependent Acquisition Mode: Focus on LTQ-Orbitrap Mass Analyzers. *J. Proteome Res.* **2013**, *12* (7), 3071–3086. https://doi.org/10.1021/pr3011588.

[51]Kalli, A.; Hess, S. Effect of Mass Spectrometric Parameters on Peptide and Protein Identification Rates for Shotgun Proteomic Experiments on an LTQ-Orbitrap Mass Analyzer. *Proteomics* **2012**, *12* (1), 21–31. https://doi.org/10.1002/pmic.201100464.

[52]Waters. *Accuracy & Resolution in Mass Spectrometry*.
https://www.waters.com/nextgen/us/en/education/primers/the-mass-spectrometry-primer/mass-accuracy-and-resolution.html (accessed 2024-05-01).

[53]UniProt. *UP000002311 Saccharomyces cerevisiae*.
https://www.uniprot.org/proteomes/UP000002311 (accessed 2024-05-02).

[54]UniProt. *UP000005640 Homo sapiens*. https://www.uniprot.org/proteomes/UP000005640 (accessed 2024-05-02).

[55]Goyon, A.; Beck, A.; Colas, O.; Sandra, K.; Guillarme, D.; Fekete, S. Evaluation of Size Exclusion Chromatography Columns Packed with Sub-3 Mm Particles for the Analysis of Biopharmaceutical Proteins. *J. Chromatogr. A* **2017**, *1498*, 80–89. https://doi.org/10.1016/j.chroma.2016.11.056.

[56]Fletouris, D. J. Chapter 10 - Clean-up and Fractionation Methods. In *Food Toxicants Analysis*; Picó, Y., Ed.; Elsevier: Amsterdam, 2007; pp 299–348. https://doi.org/10.1016/B978-044452843-8/50011-0.

[57]Barth, H. G.; Saunders, G. D.; Majors, R. E. The State of the Art and Future Trends of Size-Exclusion Chromatography Packings and Columns. *LC GC N. Am.* **2012**, *30* (7), 544-+.

[58]Sadeghi, S. A.; Chen, W.; Wang, Q.; Wang, Q.; Fang, F.; Liu, X.; Sun, L. Pilot Evaluation of the Long-Term Reproducibility of Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of a Complex Proteome Sample. *J. PROTEOME Res.* **2024**, *23* (4), 1399–1407. https://doi.org/10.1021/acs.jproteome.3c00872.

[59]Blume, J. E.; Manning, W. C.; Troiano, G.; Hornburg, D.; Figa, M.; Hesterberg, L.; Platt, T. L.; Zhao, X.; Cuaresma, R. A.; Everley, P. A.; Ko, M.; Liou, H.; Mahoney, M.; Ferdosi, S.; Elgierari, E. M.; Stolarczyk, C.; Tangeysh, B.; Xia, H.; Benz, R.; Siddiqui, A.; Carr, S. A.; Ma, P.; Langer, R.; Farias, V.; Farokhzad, O. C. Rapid, Deep and Precise Profiling of the Plasma Proteome with Multi-Nanoparticle Protein Corona. *Nat. Commun.* **2020**, *11* (1), 3662. https://doi.org/10.1038/s41467-020-17033-7.

[60]Ashkarran, A. A.; Gharibi, H.; Voke, E.; Landry, M. P.; Saei, A. A.; Mahmoudi, M. Measurements of Heterogeneity in Proteomics Analysis of the Nanoparticle Protein Corona across Core Facilities. *Nat. Commun.* **2022**, *13* (1), 6610. https://doi.org/10.1038/s41467-022-34438-8.