

Spring 2008

The Application of Innovative High-Throughput Techniques to Serum Biomarker Discovery

Izabela Debkiewicz Karbassi
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/biomedicalsciences_etds

 Part of the [Biomedical Engineering and Bioengineering Commons](#), and the [Oncology Commons](#)

Recommended Citation

Karbassi, Izabela D.. "The Application of Innovative High-Throughput Techniques to Serum Biomarker Discovery" (2008). Doctor of Philosophy (PhD), dissertation, Biological Sciences, Old Dominion University, DOI: 10.25777/a00s-qb63
https://digitalcommons.odu.edu/biomedicalsciences_etds/113

This Dissertation is brought to you for free and open access by the College of Sciences at ODU Digital Commons. It has been accepted for inclusion in Theses and Dissertations in Biomedical Sciences by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

**THE APPLICATION OF INNOVATIVE HIGH-THROUGHPUT
TECHNIQUES TO SERUM BIOMARKER DISCOVERY**

by

Izabela Debkiewicz Karbassi
B.A. May 2001, Boston University
M.S. March 2003, University of Rochester School of Medicine and Dentistry


A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

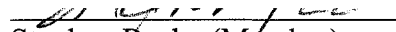
DOCTOR OF PHILOSOPHY

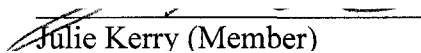
BIOMEDICAL SCIENCE


EASTERN VIRGINIA MEDICAL SCHOOL
AND OLD DOMINION UNIVERSITY
May 2008

Approved by: 

 Richard Drake (Director)


Stephen Beebe (Member)


Julie Kerry (Member)


O. John Semmes (Member)

ABSTRACT

THE APPLICATION OF INNOVATIVE HIGH-THROUGHPUT TECHNIQUES TO SERUM BIOMARKER DISCOVERY

Izabela Debkiewicz Karbassi
Eastern Virginia Medical School and Old Dominion University, 2008
Director: Dr. Richard R. Drake

Time-of-flight mass spectrometry continues to evolve as a promising technique for serum protein expression profiling and biomarker discovery. As seen in our initial SELDI-TOF MS and MALDI-TOF MS profiling study of serum for the assessment of breast cancer risk, many profiling strategies typically employ single chemical affinity beads or surfaces to decrease sample complexity of dynamic fluids like serum. However, most proteins, captured on a particular surface or bead, are not resolved in the lower mass range where mass spectrometers are most effective. To this end we have designed an expression profiling workflow that utilizes immobilized trypsin paramagnetic beads in order to reduce large mass proteins into peptides that are in the ideal mass range for serum expression profiling as well as for direct LIFT-MS/MS sequence determinations. We demonstrate that this bead-based trypsinization is efficient in digesting large serum proteins in short incubation times and is highly reproducible and amenable to an automated platform. Additionally, we show that this workflow may be combined in tandem with many different types of bead fractionation surfaces. Furthermore, by utilizing two different pooled human serum sample cohorts as proof-of-concept experiments, we are able to demonstrate the reproducibility of this method in the profiling of clinical samples and the ease of differential peptide identity determination. Overall, this method is an attractive strategy for high-throughput serum profiling with the goal of detecting and identifying differential peptides.

© Copyright, 2008, by Izabela Debkiewicz Karbassi, All Rights Reserved.

This dissertation is dedicated to my husband, John Arash, my parents, Izabella and Mirosław, and everyone who supported me during this process.

ACKNOWLEDGMENTS

I would like to thank my committee for their valuable insight and guidance throughout my Ph.D. career. I especially like to thank my advisor, Dr. Richard Drake, for his constant encouragement and support and for making my time in his laboratory both educational and enjoyable. Thank you also to my fellow students and laboratory members, you each made this a rewarding experience for me.

TABLE OF CONTENTS

	Page
LIST OF TABLES	viii
LIST OF FIGURES	ix
 Chapter	
I. INTRODUCTION	1
HISTORY OF PROTEOMICS AND EVOLUTION OF MASS SPECTROMETRY	1
PROFILING FOR CANCER BIOMARKERS USING PROTEOMIC TECHNOLOGY	17
BREAST CANCER.....	23
PROSTATE CANCER.....	32
II. DISSERTATION RATIONALE AND SUMMARY OF AIMS	39
III. AIM I: DEVELOPMENT OF PRECIOUS SAMPLE SPARING TECHNIQUES FOR MASS SPECTROMETRY ANALYSIS.	44
INTRODUCTION	44
MATERIALS AND METHODS	47
RESULTS	50
DISCUSSION.....	61
IV. AIM II: INCREASING THE EFFECTIVENESS OF THE MALDI-TOF/TOF FOR ANALYSIS OF LARGE MOLECULAR WEIGHT PROTEINS.....	70
INTRODUCTION	70
MATERIALS AND METHODS	72
RESULTS	76
DISCUSSION	111
V. AIM III: DEVELOPMENT OF INTEGRATED FRACTIONATION PROTOCOLS FOR IN-DEPTH AND AUTOMATED MALDI-TOF/TOF ANALYSIS	121
INTRODUCTION	121
MATERIALS AND METHODS	123
RESULTS	128
DISCUSSION	147
VI. CONCLUSIONS AND FUTURE DIRECTIONS	152
AIM I (CHAPTER III): DEVELOPMENT OF PRECIOUS	

SAMPLE SPARING TECHNIQUES FOR MASS SPECTROMETRY ANALYSIS	152
FUTURE DIRECTIONS OF AIM I.....	153
AIM II (CHAPTER IV): INCREASING THE EFFECTIVENESS OF THE MALDI-TOF/TOF FOR ANALYSIS OF LARGE MOLECULAR WEIGHT PROTEINS	156
AIM III (CHAPTER V): DEVELOPMENT OF INTEGRATED FRACTIONATION PROTOCOLS FOR IN-DEPTH AND AUTOMATED MALDI-TOF/TOF ANALYSIS	158
FUTURE DIRECTIONS OF AIM II AND AIM III	160
CONCLUDING REMARKS	164
REFERENCES	166
VITA.....	179

LIST OF TABLES

Table		Page
1.	Classification and regression tree analysis of 84 serum samples processed on IMAC chips.....	53
2.	Significant peaks (p -value ≤ 0.05) differentiating between cases and controls after SELDI-TOF analysis	53
3.	Cross-validation and External Validation of genetic algorithm models	55
4.	Masses used for the classification of 84 WCX fractionated serum samples using the 5 peak genetic algorithm model	55
5.	Significant peaks as determined by T-test/ANOVA (p -value ≤ 0.05) for the 84 serum sample set	56
6.	Masses used for the classification of 196 WCX fractionated serum samples using the 5 peak genetic algorithm	59
7.	Masses used for the classification of 112 IMAC fractionated serum samples using the 5 peak genetic algorithm model	62
8.	Reproducibility of immobilized trypsin bead method as seen by the coefficient of variance (CV) of twelve representative peaks	93
9.	Peptides from immobilized-trypsin digest of WCX fractionated workflow identified by LIFT-MALDI-TOF/TOF	95
10.	Examples of top peptides seen in the immobilized-trypsin digestion of WAX fractionated serum.....	98
11.	Peptides identified from untrypsinized WCX and WAX fractionated serum....	100
12.	Top 10 peptides seen in the immobilized-trypsin digestion of ConA/WGA fractionated serum after serum albumin and IgG depletion	141
13.	Reproducibility of automated immobilized trypsin bead method as seen by the coefficient of variance (CV) of twelve representative peaks	144
14.	Reproducibility of immobilized trypsin bead method determined by the coefficient of variance (CV) of twelve representative peaks (with and without outliers).....	146

LIST OF FIGURES

Figure	Page
1. Pie chart representing relative composition of proteins within plasma	21
2. MALDI-TOF scheme with magnetic beads for automation	24
3. Schematic depicting the zones of the prostate	34
4. Comparison of scraped and thaw techniques on frozen serum samples	51
5. Cluster plot for the set of 84 WCX fractionated serum samples	57
6. Cluster plot for the set of 196 WCX fractionated serum samples	60
7. Cluster plot for the set of 112 IMAC fractionated serum samples	63
8. Improved trypsinization efficiency of WCX fractionated serum proteins after reduction and alkylation.....	78
9. Effect of ZipTipC18 clean-up on MALDI-TOF spectra of WCX fractionated samples trypsinized with immobilized trypsin beads	80
10. Comparison C8 and C18 HIC-magnetic bead sequential elutions.....	81
11. Comparison of C18 and C8 HIC bead 15% acetonitrile elutions	82
12. Concentration determination for ideal trypsinization using immobilized-trypsin beads.	84
13. Effect of pH on ideal trypsinization using immobilized-trypsin beads	86
14. Workflow designed for immobilized-trypsin beads for peptide profiling using MALDI-TOF/TOF.	87
15. Reflectron mode MALDI-TOF comparison of WCX fractionated samples untrypsinized and trypsinized by either soluble trypsin or immobilized-trypsin	89
16. Linear mode MALDI-TOF comparison of WCX fractionated samples untrypsinized and trypsinized by either soluble trypsin or immobilized-trypsin	90

17.	Example of trypsin contaminant peaks in soluble trypsin digests	91
18.	Reproducibility of immobilized trypsin bead method	92
19.	LIFT analysis of representative peak <i>m/z</i> 2016.24	96
20.	Cluster plot of relative intensity distributions of peaks <i>m/z</i> 2017 and 2383 in the initial run of the SOF cohort	102
21.	Cluster plot of relative intensity distributions of peaks 2017 and 2383 in the repeat run of the SOF cohort.....	104
22.	Box plot of relative intensity distributions of peak <i>m/z</i> 1031 in the SOF cohort	105
23.	Box plot of relative intensity distributions of peak 2017 in the SOF cohort.....	106
24.	Box plot of relative intensity distributions of peak 2383 in the SOF cohort.....	107
25.	Cluster plot of relative intensity distributions of peaks 1031 and 1216 in the PCa vs. BPH cohort.....	109
26.	Box plot of relative intensity distributions of peak 1031 in the PCa vs. BPH cohort after WCX fractionation	110
27.	Box plot of relative intensity distributions of the ApoAIV peaks in the PCa vs. BPH cohort after WCX fractionation	112
28.	Box plot of relative intensity distributions of peak 1216 in the PCa vs. BPH cohort after WAX fractionation	113
29.	Single bead and tandem bead workflows	129
30.	Comparison of spectra from a tandem fractionation scheme and a single fractionation scheme of serum	131
31.	Spectrum of a tandem fractionation scheme	132
32.	Spectra comparison of serum fractionated with MB-ConA/WGA, MB-ConA, and MB-WGA and eluted with Bruker elution buffer	134
33.	Spectra comparison of serum fractionated with MB-ConA and eluted either with Bruker elution buffer or with a competitive sugar	136

34.	Spectra comparison of serum fractionated with MB-ConA or MB-ConA/WCX	137
35.	Spectra comparison of serum fractionated with MB-WGA or MB-WGA/WCX	139
36.	Spectra of serum fractionated with ConA/WGA agarose immobilized lectins...	140
37.	Reproducibility of automated immobilized-trypsin/MB-C18 workflow.....	143
38.	Reproducibility of the automated MB-WCX/immobilized-trypsin/MB-C18 workflow	145

CHAPTER I

INTRODUCTION

1.1 History of Proteomics and Evolution of Mass Spectrometry

The term “proteome” was first coined in 1994 at a meeting in Siena, Italy and was defined as the protein complement to the study of the genome. Thus, the study of the proteome was termed “proteomics” (1). Currently, proteomics is thought of as the study of not only all the proteins of a certain system, but also their structure, isoforms, modifications, interactions with other proteins and almost everything “post-genomic” (2). The early goal of proteomics was the quick identification of all the proteins expressed by a cell or tissue. However, this lofty goal has yet to be achieved for any species. Current research is more varied and focused towards determining systematically the various properties of proteins (1). Many different technologies have been developed, and are constantly evolving, to achieve the goals of proteomic researchers.

Two-dimensional electrophoresis

Initial proteomic approaches relied on protein separation by two-dimensional gel electrophoresis, with identification of protein spots of interest (2). Two-dimensional gel electrophoresis (2DE) was developed independently by Klose (3) and O’Farrell (4) in the 1970s. This is a gel electrophoresis method that separates proteins at high resolution, most commonly by first separating the proteins by their charge through isoelectric focusing in a first dimension, which is then followed by a separation of the proteins by their size in the second dimension via SDS polyacrylamide gel electrophoresis (SDS-

PAGE). The separated proteins are then detected by a stain of choosing (i.e. Coomassie stain, silver stain, or a fluorescent stain) and the staining intensity provides an estimate of the amount of protein present in each spot. It was recognized that the spot patterns generated were relatively reproducible and could be overlaid and compared between samples (5). However, there are many disadvantages to this approach. One issue with 2DE was discovered when upon sequencing of the protein spots it was found that the incidence of co-migration of proteins was more prevalent than originally thought (6). This was a draw-back since quantification of 2DE gels relies on the assumption that there is only one protein present in each spot (1). Another more commonly discussed problem with 2DE is that this is a very time consuming and labor-intensive process, i.e. as opposed to a one-dimensional gel only one sample may be run per 2DE gel. This problem was addressed by 2D-DIGE, or two-dimensional Fluorescence Difference Gel Electrophoresis. This system uses specially designed CyDye™ fluors, which are spectrally resolvable and size and charge-matched, to label samples. There are three fluors: Cy2, Cy3 and Cy5. These fluors have an NHS-ester reactive group, and thus are able to covalently attach to the amino group of lysine in proteins by an amide linkage. The lysine amino acid in proteins carries an intrinsic single positive charge at neutral or acidic pH and the fluors also carry a single positive charge. Therefore, when the fluor is coupled to the lysine it replaces the lysine's single positive charge with its own, thus not altering the pI of the labeled protein significantly from the same unlabelled protein. Cy3 and Cy5 are typically used to label independent samples, while Cy2 may be used either as an internal control between gels or also to label a third independent sample. In this manner up to three samples may be simultaneously separated on a single 2DE gel and up

to three separate conditions may be compared in a single 2DE gel. Therefore, this system is less labor intensive as having to run several separate gels and also helps to remove gel-to-gel variability, another common complaint with the 2DE system (7, 8).

Early protein sequencing methods

During the time that 2DE gels were first being implemented, Edman degradation was the method utilized to sequence the majority of proteins. This method, developed by Pehr Victor Edman in 1949, is a chemical process that removes amino acids from the N-terminus one at a time (9). The automatic version was later introduced in 1967 by Edman and Begg (10). The Edman degradation procedure has three steps: coupling, cleavage and conversion. The coupling reaction consists of phenylisothiocyanate (PITC) modifying the free-amino terminal alpha-amino of a polypeptide to form a phenylthiocarbamyl (PTC) polypeptide. The PTC amino-terminal residue is rapidly cleaved with an anhydrous acid from the polypeptide chain in the cleavage step. This occurs through the formation of a five-membered heterocyclic derivative, anilinothiazolinone (ATZ) that is made by the sulfur atom of the derivatized amino terminus and the carbonyl carbon of the first peptide bond. Thus, the cleavage reaction yields an ATZ amino acid and a shortened polypeptide. The shortened polypeptide has a reactive-terminal alpha-amino group and thus can undergo more cycles of coupling and cleavage. The final conversion step relies on the hydrophobicity of the ATZ amino acid to separate it and extract it from the hydrophilic polypeptide by a nonpolar solvent. The unstable ATZ derivative amino acid is then converted to a more stable phenylthiohydantoin (PTH) derivative via treatment with an aqueous acid. Since this

procedure removes the PTH without destroying the remaining peptide, a sequential degradation of the peptide can be achieved. Each amino acid of the peptide chain may then be identified through one cycle of the Edman degradation technique followed by one cycle of high-performance liquid chromatography to analyze the PTH amino acid (11).

The Edman degradation method was not the first and only foray into protein sequencing. There were several techniques that either came before, or were contemporaries, of the Edman degradation technique. A method that preceded Edman degradation, and which the Edman degradation method built upon, was the stepwise degradation through the use of phenylisocyanate (PIC). This method was developed in 1930 by Abderhalden and Brockmann and was based on the ability of PIC to couple to amino groups and produce an intermediate that is rearranged under acidic conditions, thus cleaving the derivatized terminal amino acid from the parent peptide (12). Edman improved on this method by changing the coupling agent to PITC, which proved to be a more readily cyclized intermediate and thus a more easily cleaved amino-terminal acid. However, it was in 1954 that the first complete description of the chemical structure of a protein was successfully performed. This was done by Frederick Sanger, who was studying the pancreatic hormone, insulin, which is a low-molecular-weight protein, composed of fifty-one amino acids. Sanger was able to determine the composition of insulin by first breaking the two chains of insulin into peptides. The insulin peptides were sequenced using a DNP (dinitrophenyl)-labeling method which covalently modifies the end amino acid in a peptide. The DNP group behaves as a chemical marker that stays attached to the amino group after the peptides are hydrolyzed into their constituent amino acids. The hydrolyzed, and partially hydrolyzed, peptides are then separated using two-

dimensional paper chromatography, also called partition chromatography, and the N-terminus is identified by its color. By aligning the peptides a contiguous sequence may be determined (11). Still, such methods as Edman degradation are time consuming and require large amounts of sample. Thus, as mass spectrometry based sequencing methods have emerged over the last 15 years, they have rapidly evolved to replace methods like Edman degradation as the technique of choice for protein analysis.

The Revolution of “Soft Ionization” Mass spectrometry

Mass spectrometry is an instrumental approach that allows for the mass measurement of ions generated from molecules and is capable of forming, separating and detecting ions based on their mass-to-charge ratio (m/z). Mass spectrometers are made up of several modular sections: The ionization source, the mass analyzer, the detector and the data recorder/processor. The ionization source converts and transfers molecules into gas-phase ions, while the mass analyzer is the device that separates gas-phase ions, usually by electric or magnetic fields. The major types of mass analyzers are quadrupole (uses oscillating electrical fields to selectively stabilize or destabilize ions), ion-trap (is typically coupled with the quadrupole mass analyzer, but now allows the ions to be trapped and sequentially ejected), and time-of-flight (TOF) (uses an electric field to accelerate ions down a flight tube). The ions from the mass analyzer go on to strike the detector. Intensity (abundance) and the m/z values of the ions are based on the magnitude of the current produced at the detector as a function of time. This is collected by the data recorder and typically displayed as m/z on the x-axis and ion abundances on the y-axis (11).

To determine the m/z of a molecule in a mass spectrometer, the analyte is first ionized and then transferred into a high vacuum system. Traditionally, the ionization of molecules into the gas phase was accomplished by electron impact (EI) or chemical ionization (CI) (11). However, peptides and proteins are large molecules and are consequently difficult to ionize by this manner since it may destroy the molecule through extensive thermal decomposition. In the early 1980s “soft ionization” techniques were first discovered and helped revolutionize ionization of peptides and proteins and thus the detection and sequencing capability of mass spectrometry. “Soft ionization” techniques are accordingly named because they allow for ionization of large, nonvolatile, polar compounds such as proteins and peptides at high sensitivity, but without excessive fragmentation (1, 5, 11). One of the first “soft ionization” techniques was fast atom bombardment (FAB) developed in the early 1980s (13-15). In FAB, the analyte is dissolved in a nonvolatile liquid matrix and placed under vacuum. The sample is then “bombarded” with fast neutral atoms in order to eject analyte ions into the gas phase. Yet, it wasn’t until the late 1980s that two “soft ionization” methods, electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI), were introduced commercially and thus made mass spectroscopy more routinely available to biological researchers (11).

ESI was developed by Fenn et al (16) and is based on the application of a high-voltage potential to a liquid as it passes through a small capillary. The ions are then desorbed into the gas phase after the evaporation of the droplet as it enters the capillary. A unique characteristic of ESI is that both singly and multiply charged ions can be formed from a single precursor. The composition and pH of the electrospray solvent, and

the chemical nature of the analyte, determine the extent of the multiple charging of an analyte. Typically peptides (<2000 daltons) yield singly or doubly charged ions, while proteins (>2000 daltons) give rise to multiply charged species. Multiple charged states are both a benefit and a drawback to the ESI technique. ESI is typically coupled with triple-quadrupole or ion-trap instruments (though recently hybrid quadrupole TOF spectrometers have become available). Thus, the benefit to multiple charging is that even simple quadrupole instruments, and other types of mass analyzers with limited m/z range, may be used to detect masses that exceed the m/z maximum of the instrument. The drawback is that the multiple charging may be very complex with overlapping ions, especially in the analysis of a mixture. Currently, though, all commercial ESI mass spectrometers are equipped with a deconvolution algorithm, which processes the charge state and isotopic envelope in order to provide a statistically averaged molecular mass (11).

The MALDI platform was first developed by Karas and Hillenkamp (17) and is based on directing a pulsed laser light onto a matrix-embedded, crystallized sample (many sample sets may be spotted with matrix on AnchorChip™ plates thus adding to this method's relative high-throughput and automation capabilities). The interaction of the laser pulse with the sample results in ionization, typically protonation, of both matrix and analyte molecules by a transfer of energy from the matrix to the embedded analyte, instead of by direct laser ionization (11). Unlike ESI, MALDI typically produces singly charged ions, which are accelerated by an electric field into the analyzer (typically a TOF for MALDI), which is a chamber under vacuum. The ions drift through the analyzer with the kinetic energy obtained from the potential energy of the electric field and are

separated based on their m/z . The relation of m/z is proportional to the square of the flight time (t) and this information may be used to determine the related mass (their mass is thus based on the “time of flight” it takes to reach the detector). In addition to this standard “linear mode”, the ions may also be deflected with an electrostatic reflector which works like an ion mirror. This technique is termed “reflectron mode” and it is based on the inversion of the ion trajectory in an oppositely polarized electric field. The reflectron mode thus provides a longer flight path allowing the masses of the ions reaching the detector to be determined with higher precision. A caveat to the MALDI platform is that the relative peak intensities may be influenced by ion suppression effects and in this way may mask peaks or at the least the spectrum may not correctly reflect the concentration of the detected peptides/proteins (18, 19). However, this may be overcome by coupling the MALDI platform to a front-end fractionation step such as a High Performance Liquid Chromatography (HPLC), termed LC-MALDI, or other chromatography-based fractionation method.

Protein Identification using Mass Spectrometry

There are two main methods that are used for protein sequencing without the need for de novo sequencing. The first is called peptide mass fingerprinting (PMF) or peptide mapping. This is the most popular method for the identification of spots from 2DE gels, since it essentially requires a pure target protein. Typically, PMF is carried out on MALDI-TOF mass spectrometers (5). In this method, an isolated protein (i.e. from an excised 2DE spot) is first digested with a protease (typically trypsin). The spectrum from the resulting peptide fragment masses is then compared against masses calculated from

the same proteolytic digestion of each entry in a sequence database to obtain the identification of the target protein (1).

The other method is termed tandem mass spectrometry (MS/MS). Tandem mass spectrometry occurs when a specific ion (termed parent ion) is selected from a mixture on the basis of its m/z ratio and then fragmented within the instrument. MS/MS has allowed the identification of proteins without the need for purification or separating proteins by SDS-PAGE. The specificity of MS/MS-based protein identification is often times much higher than that of PMF, because a peptide sequence (thus the MS/MS spectrum of a peptide) can uniquely identify a protein. Since peptide ions fragment in a sequence-dependent manner, the MS/MS spectrum is in theory the amino acid sequence of that peptide. In certain cases where PMF does not provide sufficient enough information for protein identification, a peptide from the PMF spectrum may be selected and subjected to MS/MS for improved protein identification (1, 19). In essence both ESI and MALDI methodologies have the capability of MS/MS.

In regards to the ESI platform, the method of MS/MS was greatly improved with the introduction of nanospray-ESI. This technique sprays peptide mixtures into the mass spectrometer at low flow rates through very narrow capillary columns. The capillary column serves as the ionization source and the slow flow rate allows generation of fragment ion spectra of several of the observed precursor ions. This led to peptides being detected at sensitivities not previously achieved with ESI (11, 20). For ESI, the fragments from parent ions are generated by Collision Induced Dissociation (CID). CID involves the activation of selected ions through energetic collisions with a neutral target gas. This converts translational energy into internal energy and places the targeted ions

into an activated or excited state, which is followed by the unimolecular dissociation of the activated ions to yield fragment products (11). In the beginning the selection of precursor ions was performed by the operator, but most recently software is available for computer-controlled ion selection thereby automating the MS/MS process (1).

Additionally, the ESI source is typically coupled to a Liquid Chromatography (LC) system for the improved identification of peptides in mixtures. However, even with automation, ESI is relatively low throughput as it is time consuming and only one sample may be run at a time. Furthermore, column contamination is sometimes observed and overly contaminated instruments are difficult to clean due to the instruments high sensitivity for certain compounds (11).

MALDI, on the other hand, typically uses a tandem TOF, or TOF/TOF, method termed LIFT. The LIFT mode of the MALDI instrument (named so because it "lifts" the potential energy of ions) works by fragmenting a parent ion, re-accelerating the parent and fragment ions and focusing them on the detector thereby generating a MS/MS spectrum. The instrument also contains an additional device that suppresses precursor ions. Termed "post lift metastable suppressor", it is located between the LIFT device and the reflector, where it deflects any remaining intact precursor ions and also prevents unwanted fragment ion formation after post-acceleration. The MALDI platform may fragment the parent ion by either LID (Laser-Induced Dissociation) or CID. In terms of LIFT, LID is fragmentation induced by laser irradiation and it is typically used for protein identification, while the high-energy CID is typically used for de novo sequencing (due to its ability of differentiating between leucine and isoleucine) or for

glycan analysis (19, 21). The LIFT-MALDI-TOF/TOF instrument may be fully automated or may be operator controlled as needed (19).

As ESI and MALDI increased in popularity the size of the sequence library available has also increased thereby improving the likelihood of a peptide identity match. Additionally, algorithms that match MS/MS spectra to sequence databases (22, 23) have much improved protein identification by mass spectrometry. MS/MS spectra may also be used to search translated ESTs, or expressed sequence tags, and other sequence databases containing incomplete sequences. ESTs libraries were a product of mass DNA sequencing of cDNAs derived from large pools of mRNA in the early 1990s. A decade later an ultimate normalized sequence library was available which encompassed the complete human genomic sequence, and normalized it to account for the dynamic range of transcript numbers expressed in cells, thus including the low-abundant species as well (24, 25).

There are many search engines available to search these vast libraries. One such search engine is SEQUEST. SEQUEST, a commercially available product, is the prototypical algorithmic tool for scanning MS/MS spectra against comprehensive protein databases. This algorithm finds all peptides in the database that match the input mass and then theoretically calculates the expected fragment ion masses against the observed MS/MS spectrum. SEQUEST has been modified several times to include modifications in searches (22), and to allow searching of DNA databases (26), MALDI fragmentation data (27) and high-energy CID data (28). Another search engine is Mascot, which is a free web search service. It allows for uninterrupted MS/MS ion searches of data from various mass spectrometry instruments. It works by uploading fragment ion masses and

intensities and then selecting the most intense peaks. Mascot looks for the group that most clearly differentiates the top score of the matched protein and reports its results using a probability-based MOWSE (molecular-weight search) score and level of significance. Several variable modifications, as well as constant modifications, may be selected for each search. Mascot may also be used for PMF searches (29). In actuality Mascot is an extension of the original MOWSE search engine, in which the molecular weight of the protein was taken into account based on a normalized distribution frequency value calculated for different proteases. The MOWSE scoring system is based on the principle that larger peptides carry more scoring weight and thus compensates for the nonrandom distribution of fragment molecular weights in proteins of different sizes (11, 30).

Trypsin digestion in Mass Spectrometry

As mentioned above both PMF and MS/MS are greatly facilitated by digesting samples with a known protease. This is true for both in-gel digestions (i.e. from one-dimension or two-dimensional gels) or for in-solution digestion. The knowledge of the enzyme that created the peptide in question greatly aids in protein identification through a sequence database query. In general proteases like trypsin, which produce small peptides, are beneficial for mass spectrometry because the peptides fall within the optimal m/z range of most mass spectrometry instruments (11).

One of the most commonly used and best-characterized proteases in proteomics is trypsin. Trypsin is a serine protease that cleaves at the carboxy side of lysine and arginine, except if either is followed by a proline. Typically, commercially available

trypsin is modified to render it resistant to proteolytic digestion, since proteolysis can generate a version of trypsin that has chymotrypsin specificity. Due to this, extra precautions are made and most trypsin products available for proteomics are treated with L-(tosylamido-2-phenyl) ethyl chloromethyl ketone (TPCK) to prevent chymotryptic activity (11).

Traditional tryptic digestions can be tedious and require long incubation times, since using high concentrations of enzyme will create interfering auto-digestion peaks and this may also lead to ion suppression. This has led many researchers on a quest to develop a more efficient trypsin digestion method that will allow for more high-throughput proteomic analysis. In this regard, it has been well documented that immobilizing enzymes can yield reactions that are faster, more efficient and have high-throughput (31, 32). This is due in part to the increased stability of the immobilized enzyme and also to the ability of using a higher enzyme-to-substrate ratio. There have been many approaches to immobilize trypsin onto solid supports to increase its catalytic ability, thus minimizing the time needed for digestion and streamlining the trypsinization process. Innovative approaches such as trypsin adsorbed directly onto a metal MALDI plate (33, 34), linked to copolymer MALDI sample array chips (35) or immobilized onto different monolithic HPLC columns (32, 36) have been described (monolithic columns consist of one piece of continuous, porous material that is sealed against the wall of a tube, so that mobile phase can't bypass any significant length of this porous bed but instead must permeate through it (37)). Currently, trypsin is also commercially available bound to agarose beads and immobilized as individual spin columns.

Quantification in Mass Spectrometry

Another important question in proteomics, in addition to the knowledge of a certain peptide's or protein's identity, is whether, in a particular system being studied, there are any differentially expressed proteins in the sea of proteins with unchanged expression. Therefore, quantification is another important issue in mass spectrometry. However, peptides analyzed in a mass spectrometer will produce different intensities based on chemical composition, the matrix in which they are present and other poorly understood variables, thereby hampering quantification (1). Quantification may be thought of as relative or absolute quantification. Relative quantification looks at the amounts, or concentrations, of proteins between two conditions that are being compared. This is the basis for profiling in proteomics (i.e. profiling normal versus cancer). Profiling typically occurs on the MALDI platform, since, as mentioned above, this platform has the ability to read more than a hundred samples on a single target plate, thus using the same laser settings, matrix preparations and other conditions for all sample comparisons. For relative quantification using a MALDI instrument it is very important that there is a homogenous distribution of analyte in the cocrystallite composed of matrix and analyte. Hot-spot formation, the observation that at several points of the sample no analyte signals can be detected, while at other points strong signals can be seen, must be avoided. The phenomenon of hot-spot formation is hard to predict and is dependent on such properties as hydrophobicity, polarity, and H-bond-formation potential of the analyte, the matrix and the solvent used for the sample preparation. Hot-spots cause varying ion response on different positions of the sample spots, which leads to poor spot-

to-spot and shot-to-shot reproducibility and is therefore one of the main reasons hampering quantitative MALDI-MS (38).

Absolute quantification refers to the amount, or the concentration, of a protein of interest in a particular system. This type of quantification can be performed on either MALDI-TOF or ESI instruments using internal standards, or after measurement of a calibration curve with known amounts of the particular analyte (38). One type of internal standard techniques that is growing in popularity is the labeling of all proteins in a solution mixture with stable isotopes. These methods are used for the quantification of peptides after MS/MS fragmentation. One such method is termed Isotope-Coded Affinity Tags (ICAT) and it involves labeling the cysteine residues in one sample with d0-ICAT (polyether mass encoded linker with eight hydrogens) reagent and the cysteine residues in a second sample with d8-ICAT (polyether mass encoded linker with eight deuteriums) reagent. Deuterium is a stable isotope of hydrogen that has a neutron and thus is twice as heavy as hydrogen. This will give the peptides labeled with d8-ICAT reagent a mass difference compared to the peptides labeled with d0-ICAT reagent. After labeling, the samples are combined and digested. The biotinylated ICAT-labeled peptides are enriched on an avidin affinity column and analyzed by LC-MS/MS (1, 39). Another labeling strategy is termed isotope coded protein label (ICPL), which is based on stable isotope labeling of free amino groups in intact proteins. This labeling strategy is similar to ICAT as it also relies on heavy (isotope-encoded) and light (isotope-free) mass tags. Schmidt et al demonstrated that this approach may be multiplexed by adding different weighted deuterium atoms (i.e. 7, 3 or 0 deuterium atoms) (40). A slightly different approach to ICAT and ICPL is termed isobaric Tags for Relative and Absolute

Quantification (iTRAQ). The iTRAQ method labels peptides on lysine residues and on the N-terminus with cleavable multiplex isobaric tags to produce MS/MS signature ions with the relative peak area corresponding to the proportion of labeled peptides. This technique allows up to 4 samples to be labeled, mixed and analyzed at the same time due to the availability of 4 mass tags (114, 115, 116, 117) (41, 42). Recently, Applied Biosystems, the manufacturer of iTRAQ, released an 8 tag version (8-plex with a mass tag range of 113-121) of this platform, therefore allowing the comparison of up to 8 samples.

SELDI platform

Besides ESI and MALDI, another platform has been described termed Surface Enhanced Laser Desorption/ Ionization (SELDI) time-of-flight mass spectrometry. SELDI is essentially the MALDI process but with an incorporated surface capture chemistry on the spot plate surface, and refined to individual spots on a chip platform. SELDI-TOF MS technology has helped fuel large-scale clinical proteomic profiling, with its ability to separate and analyze complex mixtures of proteins in a relatively high-throughput manner (43-45). The SELDI platform uses chips that contain specific surface-chemistries (with several different chemistries available i.e. ion-exchange, hydrophobic, normal-phase or metal chelate functional groups) for the affinity capture of proteins from biological samples. The captured proteins are then analyzed by TOF mass spectrometry to yield m/z and relative intensities of each ion (46). This technique garnered much excitement from the research community. However, soon papers emerged demonstrating the lack of analytical reproducibility of this method from

different institutions and diminished robustness of discovered biomarkers upon validation (47-49). Some of these problems have been addressed and shown to be based on study design bias, chance, overgeneralization of results, and sample processing issues and not actual problems with the instrument (48). Additionally, it has also been demonstrated that when careful study design and sample handling is applied along with instrument calibration, automation of sample preparation and supervised bioinformatics data analysis, serum expression profiling can be reproducible and portable across multiple laboratories (47, 50). Still, one of the main concerns with the SELDI platform is that, in contrast to MALDI-TOF-MS/MS, SELDI-TOF MS has the disadvantage that the peaks deemed differential can not be subjected to tandem mass spectrometry. Instead, alternative, and more time-consuming, processes must be utilized to identify the identity of any peptide or protein of interest (46).

1.2 Profiling for Cancer Biomarkers using Proteomic Technology

Cancer occurs when cells within the body divide aberrantly. These cells may then become metastatic by dislodging from the primary tumor and the disease may spread throughout the body via direct organ invasion, the lymphatic system, and/or the circulatory system (51). At least one in three people will develop cancer, of which one in four men and one in five women will die from this disease (52).

The method by which cancer develops is a multifactoral process and includes both endogenous factors, such as genetic predisposition, and exogenous factors, such as exposure to environmental carcinogens and infectious agents. Another important factor for the development of cancer is age. There is an age-associated, organ-specific tumor

incidence. Most cancers may be divided into three groups: 1) embryonic (i.e. neuroblastomas, retinoblastomas, and Wilms' tumors); 2) juvenile/young adulthood (i.e. certain leukemias and testicular cancer); and 3) those that have increasing incidence with age (i.e. prostate, breast, colon, and bladder cancers). There are several explanations as to why certain cancers are associated with aging. One such possible factor is the continuous exposure through-out life to low levels of exogenous carcinogens which would allow genetic alterations to accumulate over time. There may also be age-associated changes in some cells (whether caused by exposure to carcinogens or not), such as a decline in DNA repair capacity, that may lead to mutations that are favorable to tumor formation. Finally, with age there are alterations to the human body that may create a more permissive setting for cancer development, such as changes in the immune and hormonal environment. Typically, cells that divide are at a higher risk of acquiring mutations than cells that do not divide. Thus, cancer is generally rarer in tissues that do not divide, such as nerve tissue, but more common in breast, prostate, skin and colon, which divide frequently (52).

The first large scale technique applied to search for biomarkers in the cancer field involved the use of DNA microarrays for mRNA expression profiling (53, 54). However, mRNA levels do not necessarily correlate with corresponding protein abundance. Additionally, proteins are subjected to post-translational modifications such as phosphorylation, acetylation, glycosylation and protein cleavages. These post-translational modifications are not detected at the mRNA level (2). Thus, proteomics for biomarker discovery gained popularity as a complement to the genomic information gathered from past microarray data.

The use of blood for proteomic profiling

Various tissues and biological fluids have been used in proteomic studies for biomarker discovery. However, plasma and serum have been especially appealing as sources for cancer biomarkers. This is because blood collection is minimally invasive, economical to perform, and has the ability to be portable to remote locations (55). In addition, the “leaky” nature of the newly formed blood vessels of a developing tumor and the increased hydrostatic pressure within tumor potentially leads to the escape of tumor molecules into the circulation (56).

Cancer is as much a product of its microenvironment as the microenvironment is the product of the cancer. This is because the pre-cancerous cells interact with surrounding epithelial and stromal cells, vascular channels and with the immune system, which all may be involved in providing a favorable environment for the tumor to flourish. Conversely, tumors also affect their surroundings by partaking in abnormal cell growth, angiogenesis and cellular invasion, which are characterized by the release of proteases that digest normal tissue and blood proteins. Thus, these events may give rise to a unique cascade of events that will produce distinctive biomarkers (57). It is believed that full-length cellular or tissue protein may be too large to enter the blood vessel wall passively and thus biomarkers shed into the circulation from the tumor microenvironment are predominately peptides and/or cleavage products. This led to the coining of a new “omics” term, peptidomics. A controversial facet to peptidomics is the knowledge that peptides are generated both *in vivo* and *ex vivo*. *Ex vivo* peptide production occurs in serum by undefined collections of proteinases present in the blood that act on degradative products of the clotting cascade (58). This is a point of contention, because many believe

that ex vivo generation of peptides may falsely bias data, especially if the sample sets are handled by different individuals or are exposed to other variable conditions. It has even been demonstrated that the plasma type or type of serum separator used in a study is a source for profound variability in spectra (59, 60). However, there are also researchers that argue that the peptides generated ex vivo provide valuable insight into the nature of the proteinases that generated them. The rationale is that if the proteinases are altered between disease states, and/or hail from the tumor microenvironment, then the peptide patterns determined will reflect the activity of resident proteases in a given sample. One concept that receives universal agreement is that much care must be taken to attain uniformity in the collection and processing of blood samples (47, 56, 58-62).

Fractionation Techniques

The use of blood for biomarker discovery is hampered by the complexity and dynamic concentration range of this fluid. The potential biomarkers generated by a tumor are very dilute in the blood stream. Since, early-stage tumors might arise within a tissue volume of less than 0.1 mL then the dilution factor of the tumor-generated biomarkers would be 50,000 (assuming that the biomarkers attributed to this tumor are uniformly dispersed in the 5,000 mL total blood volume) (56). Additionally, the concentration range of serum/plasma proteins spans about twelve orders of magnitude (63), in which twenty-two proteins constitute about 99% of the protein content of plasma/serum, with the remaining 1% considered to be at low-abundance levels (Figure 1) (64). There are thought to be many proteins that are not detected in the convoluted serum proteome because they are

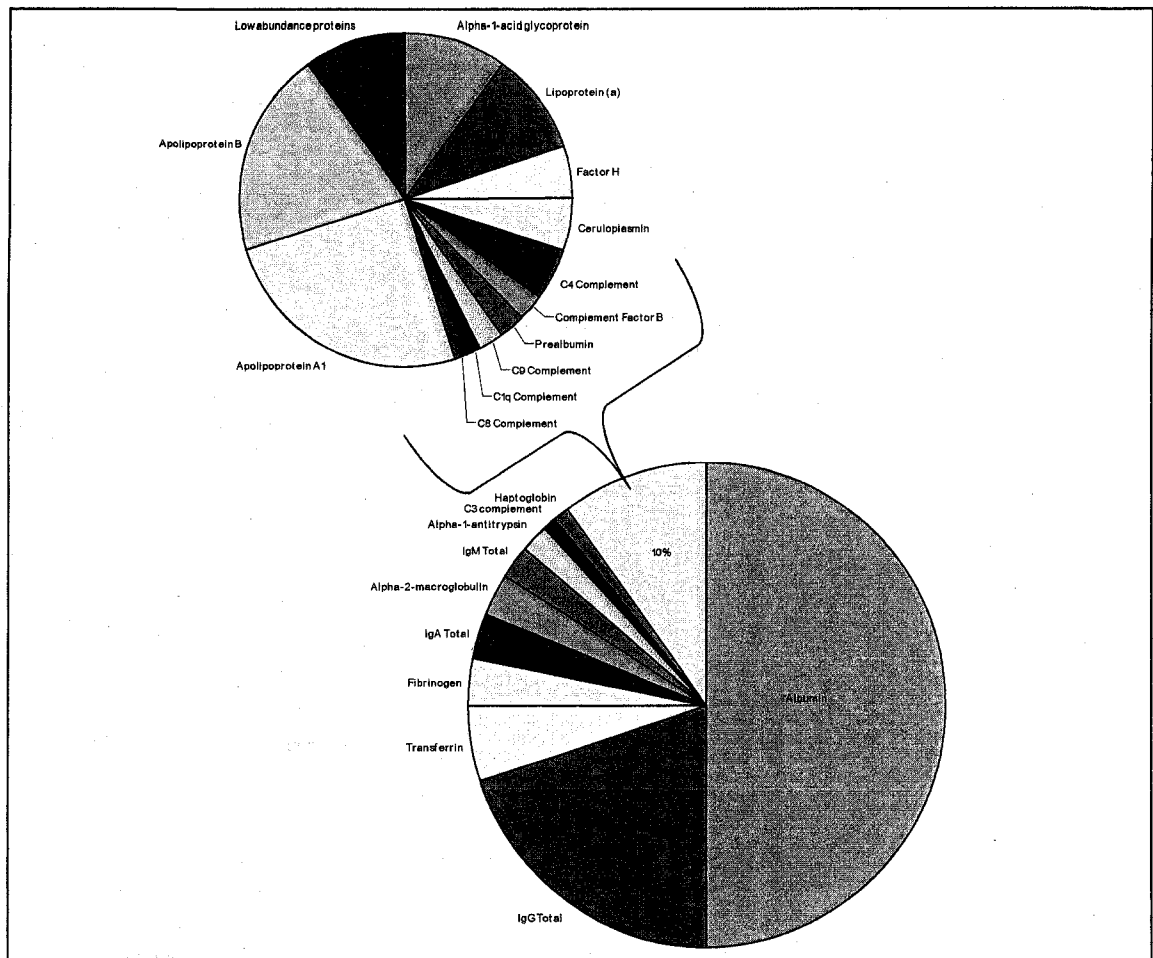


Figure 1. Pie chart representing relative composition of proteins within plasma. Twenty-two proteins make up ~ 99% of plasma (adapted from Tirumalai et al, 2003).

overshadowed by high-abundant proteins, such as serum albumin.

The current methods for evaluating the blood proteome, which includes two-dimensional gel electrophoresis-based techniques and mass spectrometry-based techniques, are only capable of examining about three orders of magnitude (64, 65). Therefore, it appears to be highly unlikely that current proteomic approaches may be able to identify molecules in the concentration range of common tumor markers (ng/mL range) without first reducing the complexity of the plasma/serum proteome (63).

One avenue for simplifying the proteome is the targeted capture of specific proteins. This capture may be for the purpose of depletion, for example the depletion of top-abundant proteins like albumin or immunoglobulins, so that the lower abundant proteins may be more readily accessible for analysis. However, the targeted capture may also be to study the captured proteins themselves. For example, one theory exists that biomarkers may be enriched in the circulatory system by accumulating on high-concentration resident proteins (i.e. albumin), thus being protected from clearance by the kidneys (56). This has led to the study of the “albuminome” where albumin is captured (i.e. targeted capture via albumin antibodies) and the proteins bound to albumin are analyzed via mass spectroscopy (64, 66).

The most common approach to reducing sample complexity is through the use of affinity-based chromatography columns (commonly used on the front-end of LC-MS/MS instrument). However, these columns are not ideal for large sample numbers and are not suited for automation. An alternative chromatography-based fractionation method that is more suited for high throughput analysis is SELDI-TOF MS, which as discussed previously, uses chips coated with different surface chemistries to fractionate samples.

However, because the surface area of the chips is small, this leads to a diminished binding capacity and thus competition among proteins. This greatly influences the spectrum of peptides and proteins detected (67-70). Chromatography columns or magnetic particles with higher surface area and binding capacities provide fractionation with less influence from competition over binding sites (70).

MALDI-TOF MS analysis typically takes advantage of these higher surface area affinity-coated magnetic beads, which may be interfaced with robotic instruments (i.e. ClinProt robot from Bruker Daltonics) that utilize magnets to automate the front-end manipulation of samples. There are various magnetic bead types available that may be combined with the ClinProt robot and the MALDI-TOF/TOF instrument from Bruker Daltonics to produce an automated method that is high-throughput, reproducible, limits operator error and consumes small amounts of the patient's sample (Figure 2) (43, 62, 71, 72). This leads not only to more significant results due to the increase in sample numbers, but may also provide an ideal technique that translates effectively into clinical, diagnostic laboratories.

1.3 Breast Cancer

Breast cancer is the most common cancer among women, and is the second leading cause of cancer mortality in women after lung cancer. According to the American Cancer Society, approximately 178,480 women in the United States were estimated to be diagnosed with invasive breast cancer and around 40,460 women were expected to die from the disease in 2007. Though predominately a female disease, breast

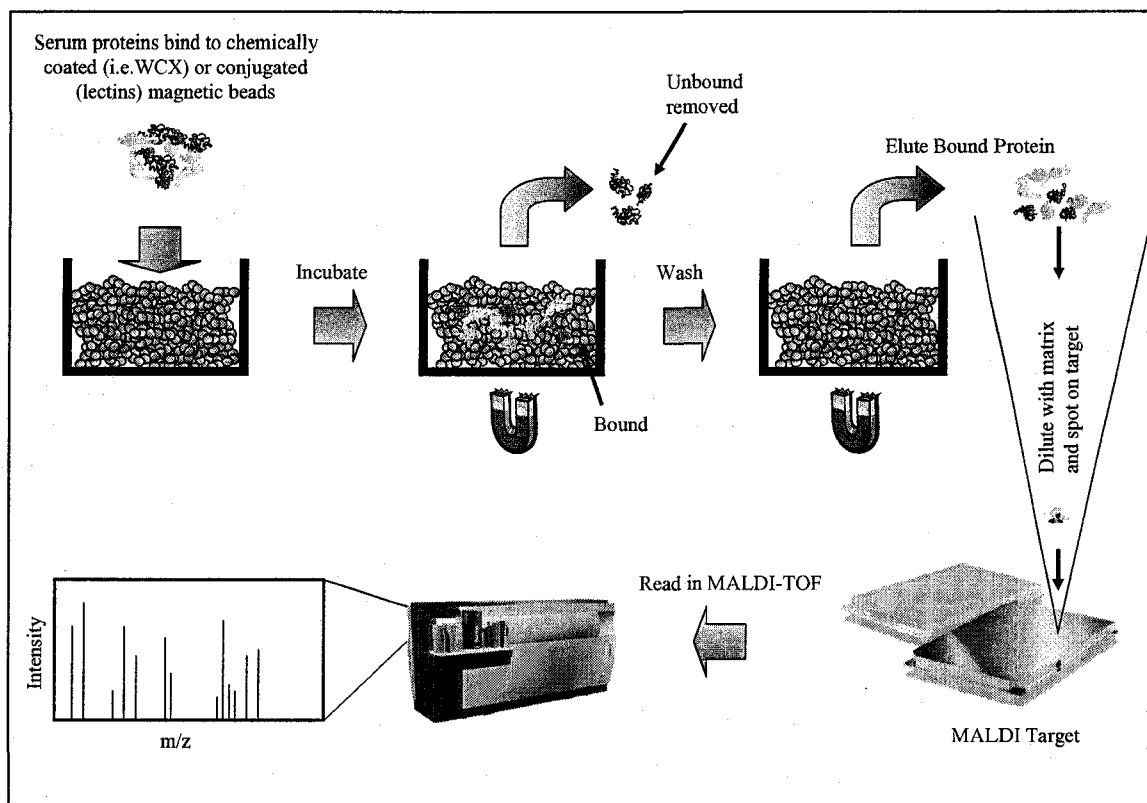


Figure 2. MALDI-TOF scheme with magnetic beads for automation. Serum proteins are first bound to magnetic beads that may be either chemically altered (such as cation/anion exchange) or conjugated to a protein (i.e. lectin or antibody). After incubation the unbound fraction is removed and the beads are washed to remove non-specifically bound proteins. The bound proteins are removed, diluted with matrix (i.e. CHCA), and spotted on a target plate. This entire process may be performed manually or robotically using a platform like the ClinProt robot from Bruker. The spotted samples are analyzed using the MALDI-TOF instrument. Adapted from Semmes et al, 2006.

cancer also affects men. In 2007 2,030 men were projected to be diagnosed with invasive breast cancer and 480 men were estimated to die from the disease (73).

The breast is composed of several lobes separated by septa of connective tissue. Each lobe consists of several lobules. These lobules are made of connective tissues and contain clusters of alveoli (secreting cells of the gland) that surround small ducts called ductules. The ductules stem into ducts and then these ducts from the various lobules come together to form a single lactiferous (milk-carrying), duct for each lobe (15-20 for each breast). Each of these main ducts terminates in a tiny opening on the surface of the nipple. A comparatively large amount of adipose tissue is deposited around the surface of the gland and between the lobes. Breast size is determined mainly by this fat surrounding the glandular tissue and not by the glandular tissue itself. The glandular and connective tissues are supported by suspensory ligaments that help anchor the breast to the underlying fascia of the pectoral muscles (74).

Estrogen and progesterone control breast development during puberty, with estrogen stimulating growth of the ducts of the mammary glands, while progesterone stimulates development of the alveoli (74). Initially, terminal ductal lobular units are formed. These are the most actively growing terminal ductal structures and are believed to contain stem cells that potentially give rise to breast cancer. Terminal ductal lobular units grow until they either regress to terminal ducts, or differentiate to alveolar buds and later to lobules during pregnancy's high branching period. The differentiated lobular structures rarely give rise to malignant tumors, thus partly explaining the protective effects of pregnancy against breast cancer (75, 76).

There are several pathological types of breast cancers demonstrating the heterogeneity of this disease. Ductal carcinoma in situ (DCIS) is the most common type of noninvasive breast cancer, and is considered a precursor of breast cancer, potentially leading to invasive disease. DCIS is confined to the ducts and does not spread into the tissue of the breast. This type of breast cancer is best detected with a mammography and almost all women with cancer at this stage are cured. Lobular carcinoma in situ (LCIS), begins in the milk-producing glands, but does not infiltrate beyond the wall of the lobules. Having LCIS increases a woman's risk for being diagnosed with more invasive breast cancer later in life and thus women with LCIS are closely monitored. The most common invasive breast cancer, accounting for 80% of invasive breast cancers, is invasive (infiltrating) ductal carcinoma (IDC). IDC starts in a milk duct, metastasizes beyond the confines of the duct into the breast tissue from where it can spread to other parts of the body. A less common invasive breast cancer is invasive (infiltrating) lobular carcinoma (ILC) and it accounts for about 10% of invasive breast cancers. This type of breast cancer starts in the milk glands, or lobules, and can metastasize to other parts of the body. Finally, the rarest, but also the breast cancer with the worst prognosis, is inflammatory breast cancer. Women presenting with this type of breast cancer have reddened-swollen breasts, due to the presence of cancer cells in the lymphatics of the skin, in addition to the presence of cancer cells in the ducts and lobules (77).

Pain is typically the most frequent breast complaint that brings a patient to a doctor's office, yet this is an uncommon risk factor since breast cancer, especially in its early stages, is usually painless. Thus, in the past, the primary symptom of breast cancer was a palpable mass and was typically first detected by the patient. However, today there

is an increased use of mammography, especially in screening programs, which has resulted in many cancers being found at a preclinical stage (78).

Risk assessment models for breast cancer

There are many factors that can put you at risk for developing breast cancer, some more significantly than others. Typically these risk factor are thought of as a culmination of risk up to a certain point in time, instead of one specific risk factor dictating whether a woman develops breast cancer or not. Some common risk factors are discussed below.

As mentioned previously, age is a common risk factor for breast cancer, with breast cancer being more common in older women. Race is another risk factor, as African American and Hispanic women are more prone to present with advanced breast cancer than Caucasian women. However, though biological differences do play a part, it is also thought that this may also be based on certain socioeconomic conditions, such as access to the same quality health care and screening (79). Another risk factor is early menarche and/or late menopause as it appears that the number of ovulatory menstrual cycles a woman experiences correlates with breast cancer risk. This is supported by the finding that oophorectomy before the age of menopause lowers the risk of breast cancer by two thirds (80, 81). Other risk factors are benign breast lesions (i.e. atypical hyperplasia is associated with the most risk, resulting in a four to five fold increase in breast cancer, while non-proliferative lesions do not pose any risk for further breast cancer development), the aforementioned LCIS and DCIS, and prior history of invasive breast cancers (81).

A protective factor against breast cancer is pregnancy. Pregnancy at a young age, especially before the age of twenty, is associated with a striking reduction on breast cancer risk. However, being over thirty years old at first live birth or nulliparity (no pregnancy) are associated with a greater than two fold risk for developing breast cancer. Additionally, the “protective” effect of pregnancy is only seen in the birth of a viable fetus (80-82). This may be because the shedding of the placenta after delivery cuts off a major source of estrogen. This drastic drop in estrogen then stimulates the secretion of prolactin, which in turn stimulates alveoli to secrete milk. Additionally, the suckling movements of the baby stimulate the secretion of oxytocin, which stimulates the alveoli to eject milk into the ducts (74). The actions of these various hormones and also the differentiation of the branched breast structure are thought to add to the protective effect of pregnancy and breast feeding.

Finally, another risk factor is familial or hereditary risk. Familial breast cancer risk occurs when one or more first- or second- degree relatives have breast cancer. On the other hand, hereditary breast cancer is a subset of familial breast cancers in which the incidence of breast cancer is related to an autosomal dominant susceptibility trait (81). The majority of hereditary cancers can be attributed to mutations in the *BRCA1* gene (breast cancer predisposition gene 1), which was discovered in 1990 by Hall et al on chromosome 17q21 (83) and the *BRCA2* gene which was mapped to chromosome 13q12-13 in 1994 (84). Subsequently, others found that germ line *BRCA1* mutations substantially increase the risk not only of breast cancer but also of ovarian cancer, while *BRCA2* mutations increase the likelihood of breast cancer development in females as well as in males (78). *BRCA1* and *BRCA2* are nuclear proteins that function in DNA repair

pathways. Loss of BRCA1 and/or BRCA2 function leads to the inability to repair damaged DNA. When damage occurs to critical checkpoint genes, such as *p53*, checkpoints such as p21 cannot be activated and cells proliferate (85).

Based on this knowledge risk assessment models that determine the probability of developing breast cancer help health care providers determine an individual's best options for cancer screening, follow-up, and the use of risk management therapies. The most popular statistical model used is the Gail model which takes into account such risk factors as age, race, age at menarche, age at first live birth, the number of first-degree relatives with breast cancer, the number of previous breast biopsy examinations, and presence of atypical hyperplasia (81). However, this model has been shown to perform poorly on an individual basis and incorporating more epidemiological risk factors only modestly improves its discriminatory accuracy (86, 87). There are other models available that mainly look at heredity, i.e. models such BRCAPro (88) are designed to predict who is a *BRCA1* or *BRCA2* gene mutation carrier. However, the use of such models is restricted to a small subset of the population as it is predicted that at most 5-10% of diagnosed breast cancers have *BRCA1* or *BRCA2* mutations (89).

Intervention for people that are deemed at a high risk for breast cancer is very physically and emotionally draining. Current risk-reducing options include lifestyle modifications, chemoprevention with tamoxifen, prophylactic surgery and ovarian suppression. Prophylactic mastectomies provide the most breast cancer risk reduction (decreases the woman's breast cancer risk to 10% of the original risk). However, this radical option is very unattractive and very traumatic to women even with the current improvements in reconstructive options (89). This is also the case for prophylactic

oophorectomies, which if performed on women in their thirties, can reduce their breast cancer risk by 60% (90). A less invasive therapy is the use of tamoxifen, a selective estrogen receptor modulator (SERM), which competitively binds estrogen receptors (ER) and is therefore prescribed to patients with ER+ breast cancers. However, as with most hormonally responsive cancers, they may develop hormone resistance (or independence) and thus will no longer respond to this type of treatment. Additionally, tamoxifen carries with it certain serious potential side effects such as development of endometrial cancer, stroke, and pulmonary embolism (which is more age dependent and seen typically in older women) (89). Since current therapies for risk-reduction carry heavy health and well-being burdens there is a need for complementary analysis tools, such as breast cancer associated biomarkers, that will aid in improving the current breast cancer risk assessment models.

Biomarkers in Breast Cancer

Recently, the American Society of Clinical Oncology released their 2007 recommendations for the use of tumor markers in breast cancer (91). CA 15-3 and CA 27.29, which are well-characterized assays for screening MUC-1 antigen in peripheral blood, were approved for monitoring of patients with metastatic disease during therapy, but in conjunction with physical examinations and diagnostic imaging. MUC-1 is a type of mucin protein. Mucins are high molecular weight glycoproteins that provide a protective layer on epithelial surfaces and are involved in cell-cell interactions, signaling, and metastasis (92). Rising levels of MUC-1 as seen by CA 15-3 and CA 27.29, in the absence of measurable disease, would indicate treatment failure. The same is true for

another approved biomarker, carcinoembryonic antigen or CEA, though it is less sensitive than the MUC-1 test for detecting metastatic disease. Current established biomarkers that are used to initially screen breast cancers to aid in the determination of treatment options are the ER, progesterone receptor (PgR) and HER2/*neu* (c-erb-B2). ER and PgR content are associated with favorable prognosis and would benefit from hormonal treatments such as tamoxifen or hormone ablation therapy. HER2 is a member of the epidermal growth factor receptor family (93) and is amplified and overexpressed in 15-30% of newly diagnosed breast cancers and is linked with more aggressive breast cancers (94). Additionally, circulating HER2 extracellular domain (ECD) levels, which can be detected in plasma or serum and are elevated in about 30% of metastatic breast cancer cases (91, 95, 96), have been proposed for the monitoring of patient response to certain therapies (91). An assay recently approved for the prediction of breast cancer recurrence in patients diagnosed with node-negative, ER positive breast cancer and treated with tamoxifen is the *Oncotype DX* assay (from Genomic Health Inc, Redwood City, CA). *Oncotype DX* is a reverse transcriptase (RT)-PCR assay that measures the expression of 21 genes (16 cancer related genes and five reference genes). This information is then processed by an empirically derived algorithm to categorize patients into 3 risk groups of distant recurrence: low, intermediate and high. Thus, this assay is used to identify patients that would obtain the most therapeutic benefit from adjuvant tamoxifen and thus may not require adjuvant chemotherapy. However, patients with high-recurrence score would be best treated with chemotherapy rather than tamoxifen (91, 97). Finally, there are many biomarkers and novel techniques that are reported, but yet to be successfully validated and recommended for use on patients. One example is

the use of circulating tumor cells (CTCs) as markers for breast cancer. CTCs are cells within blood that possess antigenic or genetic characteristics of a specific tumor type. Thus, the presence of CTCs in a breast cancer patient may predict the presence of an aggressive primary tumor or potentially micrometastasis. CTCs may be detected by positive cell selection using immunocapture (using immunomagnetic beads conjugated with an antibody specific for a cell surface, epithelial or cancer related antigen) and immunocytochemistry or by gene expression analysis for the presence of cytokeratins and tumor antigens. A reverse methodology is to first remove the leukocytes and then interrogate the remaining cells by immunocytochemistry or RT-PCR. Recently the US Food and Drug Administration approved a test for CTCs called CellSearch Assay (Veridex, Warren, NJ). However, this assay still needs to undergo additional validation to confirm clinical value of this test for use in patients (91).

1.4 Prostate Cancer

Prostate cancer (PCa) remains the most common malignancy and second-leading cause of cancer deaths among males in the United States, with an estimated 218,890 new cases in 2007, accounting for 29% of new male cancers (98). Early prostate cancers demonstrate few signs and symptoms and the presence of symptoms such as hematuria, obstructive voiding symptoms, and bone pain generally indicate advanced prostate cancer.

Disorders of the prostate can be divided into three main categories: benign prostatic hyperplasia (BPH), prostate cancer (prostatic adenocarcinoma), and prostatitis (bacterial infections of the prostate). The prostate, a small glandular organ in men that is

located in front of the rectum and beneath the urinary bladder, has three distinct zones: transitional, central and peripheral. These zones have different histology, embryonic origins and give rise to different pathologic entities. The central zone extends from the bladder base and encircles the ejaculatory ducts. It contains roughly 25% of prostatic glandular elements and only 1-5% of prostate cancers stem from there. The transitional zone surrounds the proximal urethra and, in youth, contains 5-10% of the prostatic glandular tissue. BPH arises from the transitional zone, as well as 20% of prostate cancers. The peripheral zone comes from the mesoderm and accounts for the majority of the glandular tissue. This zone gives rise to about 70% of prostate cancers and is also the site of most prostatic infections (Figure 3) (99).

Risk factors for developing prostate cancer include a family history of prostate cancer (which increases with the number of first-degree relatives affected), age (since 70-80% of patients who have prostate cancer are 65 years old or older), and race (i.e. African-American men have the highest incidence of prostate cancer in the United States). Interestingly, African American men have higher levels of endogenous androgen than Caucasians, and it is these higher levels that are thought to play a role in the development of prostate cancer. The prostate is a hormone responsive glandular organ, similar to the breast, and therefore a typical treatment for advanced stage prostate cancer is androgen ablation. Androgen ablation may be performed through surgical or medical (chemical) castration. However, the use of androgen ablation, termed androgen deprivation therapy (ADT), leads to androgen-independent or hormone-refractory prostate cancer. Thus, most patients with metastatic prostate cancer will respond initially to this therapy, but eventually these patients will develop progressive disease despite

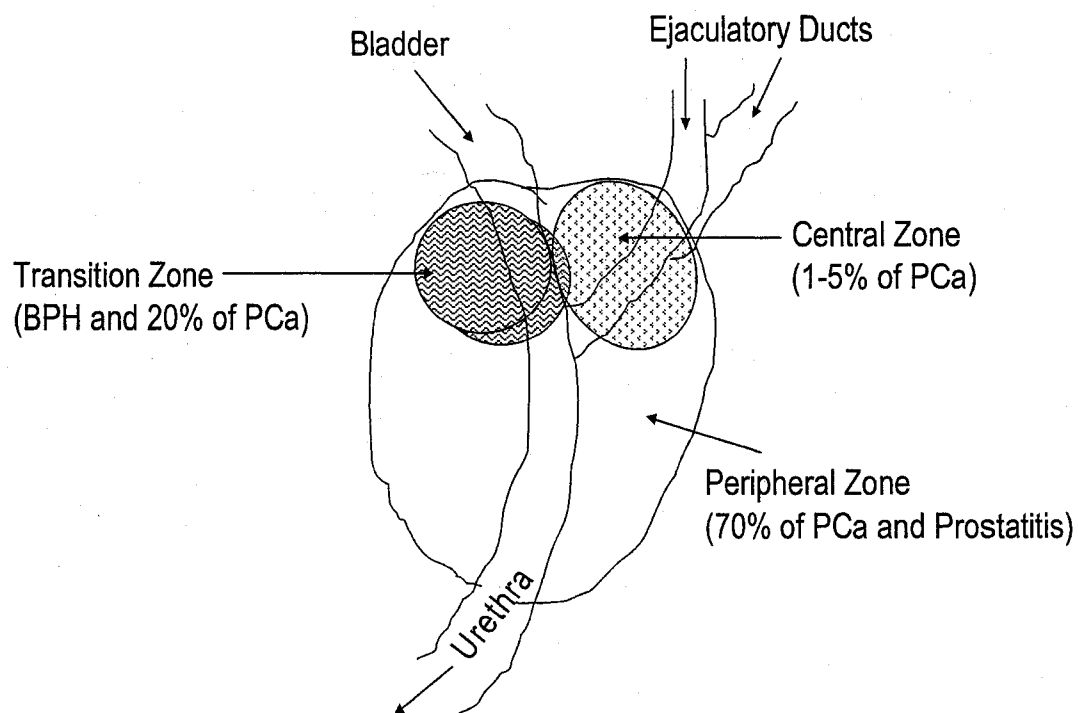


Figure 3. Schematic depicting the zones of the prostate. The central zone extends from the bladder base and encircles the ejaculatory ducts accounts for only 1-5% of prostate cancers cases. The transitional zone surrounds the proximal urethra and gives rise to BPH as well as 20% of prostate cancers. The peripheral zone, which contains the majority of the glandular tissue, gives rise to about 70% of prostate cancers and is also the site of most prostatic infections.

continued androgen suppression. Unfortunately, advanced prostate cancer patients on ADT have an average survival of twenty-four months and the five year survival is only 20% (100). Additionally, ADT therapy has been linked with an increased risk of heart disease and diabetes (101). However, if the disease is caught early on it is far more manageable. Treatment for localized prostate cancer is either watchful waiting or surgery. Current surgical methods, such as laproscopic radical prostatectomy, have improved in terms of becoming less invasive and the five year survival rates of men after radical prostatectomy are around 80% (99).

PSA as a Biomarker for Prostate Cancer

Currently, one of the main pretreatment diagnostic screens used in the detection of prostate cancer is the presence of the biomarker Prostate Specific Antigen (PSA). PSA, a 240 amino acid serine protease, is a member of the human kallikrein gene family (102). The functional role of PSA is not entirely understood, but it is known that it is secreted in high concentrations in the seminal fluid where it acts to liquefy the semen. PSA is found in much lower concentrations in the serum (103). The elevation of this marker in blood, usually in conjunction with an abnormal digital rectal exam (DRE) will lead to a biopsy of the prostate for a definitive diagnosis (104).

There is still a debate over the exact value that should serve as the threshold between normal and abnormal serum levels of PSA. Traditionally, the normal serum value for PSA was accepted to be 4.0 ng/mL or less. However, more recently, it has been suggested that men sixty years old and younger should not have a PSA value of 2.5 ng/mL or higher (104, 105). This change was brought about by the Prostate Cancer

Prevention Trial (PCPT), which was a randomized clinical trial to test the hypothesis that blockade of 5-alpha reductase activity with finasteride could prevent prostate cancer. The criteria for patient enrollment in this study were PSA levels of less than 3 ng/mL, normal results of a DRE and being at least fifty-five years of age. The patients were then randomly assigned to finasteride or placebo for seven years. At the end of the study all participants, regardless of PSA value, underwent an end-of-study biopsy. When the prevalence of prostate cancer among men in the placebo group was evaluated it was found that even in this very-low-risk group of men, the incidence of prostate cancer on sextant biopsy was 15.2%. Ultimately, it was determined that there is no single PSA value that will provide assurance that a man does not have prostate cancer (106). However, this study also found that as the PSA value increases, the likelihood that there is a detectable prostate cancer and more specifically a high-grade prostate cancer significantly increases (107).

Ability of PSA to differentiate between BPH and Prostate Cancer

Unfortunately, PSA is not a specific biomarker for prostate cancer since its serum level increases with BPH, and can also be affected by many other factors such as inflammation, prostatitis and even ejaculation. It has been estimated that two out of three men with abnormal results on routine PSA screening will not have prostate cancer (108). It is therefore critical to be able to distinguish between prostate cancer and BPH, since BPH is highly prevalent amongst older men. Histologically, 50% of men in the fifth decade of life demonstrate evidence of BPH at the time of their autopsies. In addition, it has been estimated that 18% of men in their forties, 29% of men in their fifties, 40% of

men in their sixties, and 56% of men in their seventies have signs of BPH, such as decreased force of stream, nocturia, straining, urinary frequency, and urinary urgency. This data is based on the study of 7,588 men from nine Asian countries and these rates have been shown to be similar in Australia, America, and Europe (99, 105).

A step forward in improving the specificity of the PSA test for prostate cancer screening came about in the form of free PSA (fPSA). In serum, PSA exists in bound and unbound forms, with the bound form being more prevalent. Generally, the bound form consists of PSA being complexed to the anti-proteases, alpha-1-antichymotrypsin or alpha-2-macroglobulin. However, it is the fPSA that has been found to be lower in patients with cancer and seems to be less affected by benign hyperplasia than total PSA. The risk of cancer is high for those who have a free/total PSA (f/tPSA) of less than 15%, whereas BPH is more likely when f/tPSA is more than 25%. Unfortunately, for most patients, f/tPSA falls between these two values and is mainly used to evaluate the need for repeat biopsies when negative (105). Another method of improving the accuracy of PSA may be to not rely on an isolated PSA value, but rather study PSA trends. For example, many studies suggest that if a patient's serum PSA increases more than 0.75 ng/mL per year, then there is an elevated risk for prostate cancer regardless of the absolute serum PSA value (99). A new diagnostic tool that is currently being evaluated is a test for the inactive precursor form of PSA (proPSA). The native form of PSA is designated as (-7), which is proPSA without 7 amino acids, while the proPSA form includes the truncated products of PSA, (-5), (-4), (-2) and (-1) forms (109, 110). ProPSA has been shown to increase the sensitivity of PSA and specificity of PSA in differentiating between PCa and BPH. The (-2)proPSA has been found to have the most

significant correlation and is seen elevated in serum of PCa patients compared with serum of BPH patients. Further validation of this assay needs to be performed before its clinical utility is determined (111-113).

The road to prostate cancer diagnosis is still very much littered with unnecessary biopsies and this leads to needless anxiety, in addition to expensive follow-up testing and procedures that carry further health risks (55). It is becoming very clear that PSA alone will not accurately predict the development and progression of this complex disease. Hence, the pursuit of new biomarkers that will complement and improve the current diagnostic tools continues.

CHAPTER II

DISSERTATION RATIONALE AND SUMMARY OF AIMS

The goal of this research project is to develop techniques to address the requirements of serum protein expression profiling of cancer cohorts for the purpose of early detection and the prediction of cancer risk. To effectively profile serum for cancer biomarkers one must preserve the integrity of the proteome, have the capacity to reproducibly process many samples simultaneously for statistical validity, allow for several fractionation techniques to simplify the serum proteome and also possess the ability to determine the identity of differential peaks.

MALDI-TOF MS is currently the platform of choice for expression profiling, due to its high-throughput nature and its capability to identify peaks of interest directly with the LIFT-MS/MS platform. Additionally, the MALDI instrument may be integrated with front-end automated fractionation processing of samples. However, the effective range for most MS instruments, including the MALDI platform, is in the low molecular weight range (less than 20kDa) and thus higher molecular weight proteins are typically excluded from high-throughput profiling studies. Hence, in addition to profiling these low-molecular weight endogenous peptides and proteins, one may increase the mass range of the MALDI platform by trypsinizing fractionated serum and profiling the resulting tryptic peptides. **Thus, the specific hypothesis of this dissertation is that development of integrated fractionation and digestion techniques will allow for more effective detection and identification of differential cancer biomarkers secreted or shed into the blood from the growing tumor or from the interaction of the cancer with the**

surrounding microenvironment. Different front-end fractionation schemes will be considered for their ability to increase the scope of the proteome under investigation and for their compatibility with the developed digestion protocol. Additionally, we will also investigate sample sparing techniques that will allow continued use of precious, limited samples, while preserving the quality of the proteome under investigation.

During the course of this project we will use two model serum cancer cohorts to gauge the validity of the techniques laid out in this thesis dissertation. The first serum cohort is a group of patients that will or will not develop breast cancer within the next 5 years. Biomarkers from such a study would be ideal for use in conjunction with current breast cancer risk assessment models such as the Gail method. The second serum cohort is a group of men that either have been diagnosed with either BPH or PCa, but have PSA levels in the non-discriminatory zone of 2-10 ng/mL. Results from this type of sample set would be instrumental in assessing the possibility for the development of early detection screens that would compliment the current PSA diagnostic tool in the differentiation between BPH and PCa afflicted males.

The hypothesis of this dissertation was evaluated by addressing the following specific aims:

Aim I: Development of precious sample sparing techniques for mass spectrometry analysis. This aim entails:

A. Validating a “scrape” technique for the use of sparing precious samples unnecessary freeze-thaw cycles.

B. Application of the scrape technique to a large cohort of valuable serum samples that are in limited quantity and need to be preserved for further experiments. The cases in this sample cohort are of women who developed invasive breast cancer during the first 10 years of follow-up and who had stored serum available that had been drawn between 1 and 5 years prior to the diagnosis of breast cancer. The controls are matched to the cases on age and length of follow-up, who were not diagnosed with breast cancer during the first 10 years of follow-up.

C. Utilization of SELDI-TOF MS and MALDI-TOF MS to investigate this cohort for the purpose of differentiating between women that will develop invasive breast cancer 1-5 years into the future from women that will not develop breast cancer in that same time frame. IMAC (immobilized metal affinity capture) chips will be used prior to SELDI-TOF/MS and magnetic bead (MB)-IMAC and MB-WCX (weak cation exchange) will be used prior to MALDI-TOF MS.

D. Generating algorithm models, using Biomarker patterns software for SELDI-TOF data analysis or ClinProTools software for MALDI-TOF data analysis, for identification of differential peaks and assessing their ability to segregate the two groups.

E. Validation of models using larger serum sample sets that are processed as described in B and C. The validation sample sets are both unblinded and blinded thus allowing for the alteration of models and the determination of the ability of these models to correctly classify samples groups.

Aim II. Increasing the effectiveness of the MALDI-TOF/TOF for analysis of large molecular weight proteins. This aim entails:

A. Integrating a trypsin digestion step following standard chemical affinity fractionation of serum samples. The effect of buffer pH and protein concentration on digestion efficiency and sample clean-up prior to MALDI-TOF analysis will be evaluated. The protocol will take advantage of paramagnetic bead technology and will strive to make each aspect of the workflow compatible with future automation.

B. Comparing the efficiency of the trypsin magnetic bead digestion of serum proteins with a standard soluble trypsin protocol.

C. Assessing reproducibility of the trypsin bead digestion workflow, using MB-WCX as a representative initial serum fractionation step.

D. Investigating a MB-WAX (magnetic bead weak anionic exchange) front-end fractionation step for its adaptability to the trypsinization workflow.

E. Performing and comparing sequence identifications of selected m/z peaks from the WCX and WAX fractionation/ trypsin bead digestion workflows using the LIFT-MALDI-TOF/TOF. Additionally, tryptic peptide workflows will be compared against endogenous peptide results in terms of profiling and ability of peptide identification efficiency.

F. Performing proof-of-concept trypsin bead workflows on clinical samples in order to determine the ability of this method to detect and identify differential peptides in a reproducible manner. The two serum cohorts used in this effort will either investigate the ability to predict occurrence of breast cancer in the future (using the samples from Aim I) or will attempt to differentiate between BPH and PCa patients.

Aim III. Development of integrated fractionation protocols for in-depth and automated MALDI-TOF/TOF analysis. This aim entails:

A. Investigating whether workflows designed in Aim II (MB-WCX and MB-WAX initial fractionation followed by immobilized-trypsin digestion and MB-C18 capture) may be compatible if performed in tandem for the purpose of sample preservation and for further mining of the serum proteome.

B. Determining if lectins immobilized on magnetic or agarose bead supports may be incorporated in the immobilized-trypsin bead workflow for the purpose of automated protocols. These protocols may be used to profile the serum proteome for differences in glycan moieties between captured glycoproteins.

C. Designing schemes for the ClinProt robot to determine if the protocols developed in this dissertation may be automated in an effective and reproducible manner.

CHAPTER III

AIM I: DEVELOPMENT OF PRECIOUS SAMPLE SPARING TECHNIQUES FOR MASS SPECTROMETRY ANALYSIS

3.1 Introduction

Breast cancer is the most common malignancy in women and the second most common cause of cancer-related death according to the 2007 American Cancer Society report (73, 77). When treated early breast cancer is a manageable disease i.e. women presenting with localized disease have a greater likelihood of remaining disease-free after five years than women presenting with regional disease or metastatic disease (114). Thus, detecting a malignancy before its clinical appearance is the goal of cancer diagnosis and treatment (77). In response to this, risk assessment models that determine the probability of developing breast cancer have been developed to help health care providers determine an individual's best options for cancer screening, follow-up, and the use of risk management therapies such as chemoprevention with tamoxifen or surgical intervention.

The most popular statistical model used is the Gail model which takes into account such risk factors as age, race, age at menarche, age at first live birth, the number of first-degree relatives with breast cancer, the number of previous breast biopsy examinations, and presence of atypical hyperplasia (81). However, this model has been shown to perform poorly on an individual basis and incorporating more epidemiological risk factors only modestly improves its discriminatory accuracy (86, 87). Therefore,

there is a need for complementary analysis tools that will aid in improving the current breast cancer risk assessment models.

To this end, using proteomic technology, such as the surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) and matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) instrument, allows for the profiling of patient serum in a reproducible manner (18, 50, 115) with the potential to identify serum biomarkers that may differentiate between women who in the future will develop breast cancer and those who will not.

Serum samples are collected and processed in specialized tubes, which allow for the clotting of the blood and the removal of blood cells and resulting fibrin clots from the fluid portion of the blood. If the serum samples are destined for a particular study then they will typically be aliquoted from these larger sample collection/serum storage tubes into smaller tubes and stored until the experiment is about to be performed. In this way the aliquots may be used for the study on an “as needed” basis, thus minimizing freeze-thaw cycles. However, more often than not, studies are performed on samples that were collected and frozen for storage. This may be due to lack of freezer space or that the samples may be needed for another retrospective study down the line, beyond the goals of the original study design plan.

Excess freeze-thaw cycles have been shown to effect the dynamic alterations of the serum proteome in the mass range of most mass spectrometry instruments (18). Additionally, adding extra freeze-thaw cycles allows for more sample handling and may introduce variability into the sample set. It has been found that the length of time that samples are left on ice or on the bench-top greatly influenced the peptide profile by mass

spectrometry (18, 116). One reason for this may be the fact that there are still proteases in the serum that may lead to the production of ex-vivo peptides. Ex vivo generation of peptides may falsely bias data, especially if the sample sets are handled by different individuals or are exposed to other variable conditions.

In Aim 1 we thus present a sample cohort utilized for the goal of predicting a women's chance of developing breast cancer within the next five years. This sample cohort was originally destined for a women's osteoporosis study in the San Francisco Bay area. During the course of this large study many women developed breast cancer, thus the serum cohorts were selected to compare samples from women who were going to develop breast cancer versus age-matched women who will not. Since, these samples are precious and repeated freeze-thaw will damage the quality of the proteome for future studies, we scraped the serum samples, while they were still frozen, with a sterile, blunt needle. This technique negated a freeze-thaw cycle for both the stock samples and aliquots to be analyzed. The scrape method is shown to be comparable in spectra quality to spectra generated from samples that are completely thawed before processing for mass spectrometry analysis. Additionally, SELDI-TOF MS and MALDI-TOF MS data of scraped samples, after weak cationic exchange (WCX) and immobilized metal affinity chromatography (IMAC) fractionation to reduce serum proteome complexity, is presented with the goal of finding discriminatory peaks that will aid in the prediction of breast cancer risk in the future. The discriminatory biomarkers for breast cancer risk can be developed into simple blood tests that can be combined with current statistical methods to improve risk assessment for breast cancer.

3.2 Materials and Methods

Sample Selection and Processing

Serum samples were selected by Jeffrey Tice, M.D., from participants in the Study of Osteoporotic Fractures (SOF), which is a population based cohort study of risk factors for the development of osteoporotic fractures in postmenopausal women. Dr. Tice is a coordinator of the SOF study and an epidemiologist and breast cancer clinician from the University of California San Francisco. The original osteoporosis study had 9704 women enrolled from four geographic regions (Baltimore, MD; Pittsburgh, PA; Minneapolis, MN; Portland, OR) between September 1986 and October 1988. African American women were excluded at baseline because of their low risk of hip fracture. For our breast cancer study, “cases” of women were selected at random who developed invasive breast cancer during the first 10 years of follow-up and who had stored serum drawn between 1 and 5 years prior to the diagnosis of breast cancer. “Controls” were chosen randomly which were matched to cases on age and length of follow-up, from the remaining participants in the SOF cohort who were not diagnosed with breast cancer during the first 10 years of follow-up. For the pilot study, 42 cases and 42 controls were selected. For the validation study, an additional 104 cases and 104 controls were selected. All serum samples selected came from post-menopausal women who were not using hormone replacement therapy. Institutional Review Boards at the four clinical sites and the coordinating center approved the study protocol and all participants signed informed consent at enrollment. The selected serum cases and controls were scraped without thawing and processed with IMAC chip surfaces for SELDI-TOF MS analysis and with MB-WCX (weak cationic exchange magnetic beads) and MB-IMAC

(immobilized metal affinity chromatography magnetic beads) on the Bruker ClinProt robot for MALDI-TOF MS analysis.

Scraping vs. Thawing serum processing techniques

A traditional “thaw” technique was compared to a “scraping” technique. Initially, 100 μ L was scraped from the top of each frozen sample (samples had been frozen at -80°C) with a steel nail. The same frozen samples were then allowed to thaw to completion and 100 μ L of each sample was collected to compare against the “scrape” technique. For the large cohort processing, the steel nail was replaced with a sterile, blunted, thick steel needle.

SELDI-TOF MS and data analysis

Twenty microliters of each serum samples was diluted in a 1M urea, 0.125% CHAPS and phosphate-buffer saline buffer. These diluted samples were then robotically processed onto eight-spot copper activated immobilized metal affinity chromatography (IMAC-Cu or IMAC3) chip arrays (CIPHERGEN Biosystems, Fremont, CA) with a Biomek 2000 liquid handling system (Beckman Coulter, Fullerton, CA). The IMAC3 chip arrays were air dried and overlaid with a saturated matrix solution containing sinapinic acid in 50% (v/v) acetonitrile and 0.5% (vol/vol) trifluoroacetic acid.

The IMAC3 chips with sample and matrix were run on a SELDI ProteinChip System (PBS-II, CIPHERGEN Biosystems). The mass spectrometer was externally calibrated using a mixture of known peptides. Blinded serum samples were randomized in blocks of seven to ensure that each chip included at least three cases and three controls.

Each sample was run in duplicate on separate chips and the results were averaged. Additionally, a standard serum sample (QC) was applied to one of the eight spots on each chip for quality control. Spectra generated were subjected to pattern recognition and sample classification analysis performed with the Biomarker Pattern Software (CIPHERGEN Biosystems).

MALDI-TOF MS and data processing

Samples were randomized (QC samples were included) and processed using the ClinProt robot (Bruker Daltonics, Bremen, Germany) using both MB-WCX and MB-IMAC independently as per manufacturer's instructions (10 μ L magnetic beads incubated with 20 μ L serum sample). The eluted samples were mixed 1:10 with R-cyano-4-hydroxycinnamic acid (CHCA) matrix solution (0.008 g CHCA prepared in 2mLs acetone and 2mL ethanol) and 0.8 μ L was spotted in duplicate robotically on an AnchorChip plate. The fractionated endogenous peptide profiles were generated from an average of four hundred laser shots in the linear mode by the MALDI-TOF Ultraflex I mass spectrometer (Bruker Daltonics) and analyzed with ClinProTools 2.0 (Bruker Daltonics). ClinProt software baseline subtracted and normalized the spectra (using total ion current). A k-nearest neighbor genetic algorithm contained in this software suite was used to generate prediction models to classify the groups analyzed. Twenty percent of the samples were left out of the model generation process and used to cross-validate the model within the software.

3.3 Results

“Scrape” vs. traditional “Thaw” method

The SOF study cohort is made up of 9,704 participants, thus making aliquoting these samples into multiple tubes not feasible at the main storage site. Additionally, each of these participants only has one available sample drawn at a given date. In order to utilize this expansive in participants, yet limited in individual sample quantity, serum cohort that would give us the potential of predicting whether a woman will develop breast cancer within the next five years, we had to evaluate a method to process the samples without thawing whole stock/storage vials. The method we developed in lieu of a freeze-thaw cycle was a simple scraping of the frozen sample from the top of the vial. We determined the validity of this scrape method by comparing a traditional total thaw technique to a scraping technique. Initially, 100 μL was scraped from the top of each frozen sample with a steel nail (later this was adapted to the utilization of a blunt sterile needle). The same frozen samples were then allowed to thaw to completion and 100 μL of each sample was collected to compare against the “scrape” technique. Samples were processed on IMAC chips and analyzed with SELDI-TOF MS. Spectra patterns and intensities were compared between the thawed and scraped samples and standard deviations were calculated for ten representative peaks. Since no difference in intensities and standard deviations was observed between scraped or thawed samples (Figure 4), we deemed that this scrape technique was compelling enough to continue and process the remainder of the cohort samples.

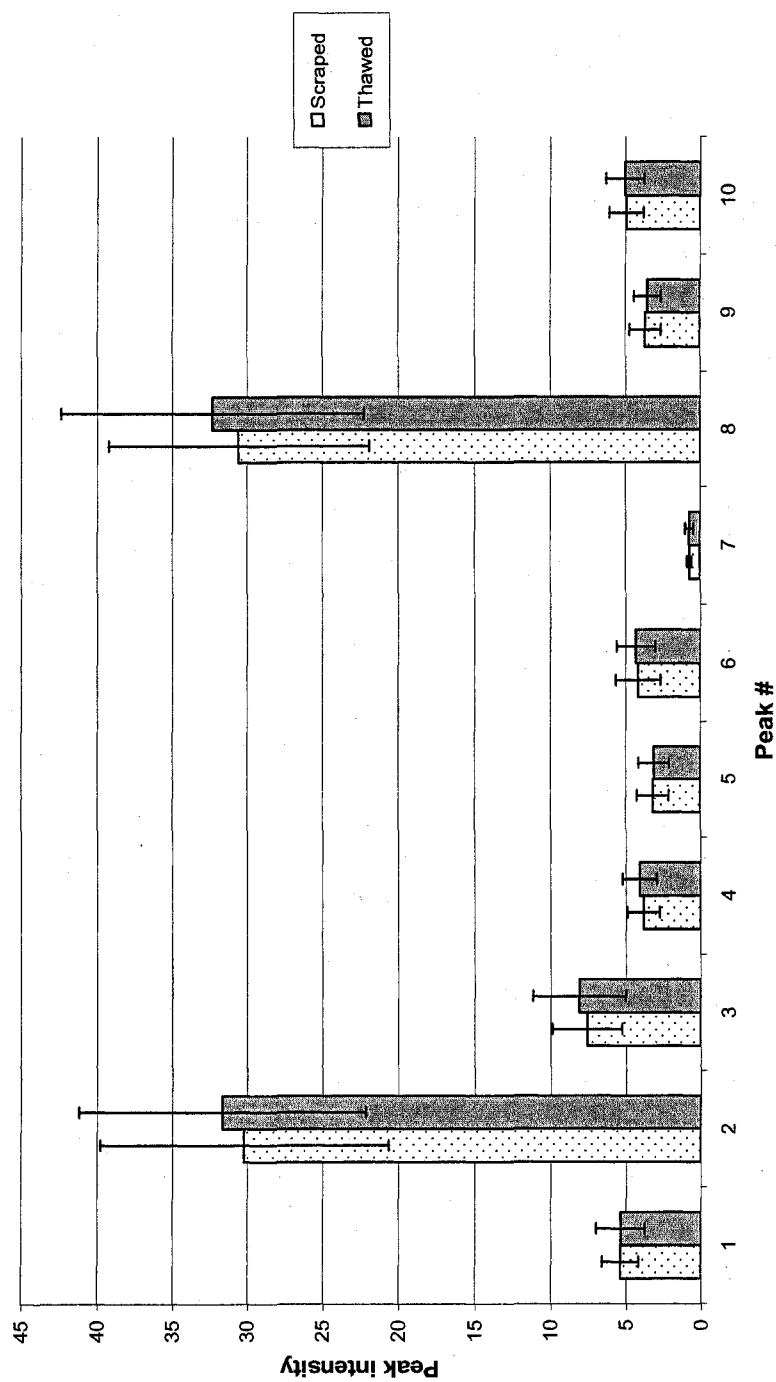


Figure 4. Comparison of scraped and thaw techniques on frozen serum samples. Intensities and standard deviations of ten representative peaks are shown, with no significant differences seen between the techniques. Spectra were generated using the SELDI-TOF MS platform after processing on IMAC chips.

Analysis of samples processed with scrape method using SELDI-TOF MS

We compared 84 samples (42 cases and 42 controls) using IMAC chips and the SELDI-TOF MS platform as described in Materials and Methods. In a classification and regression tree analysis using Biomarker patterns software (Ciphergen Biosystems), we generated a tree with a recognition capability of 85.7% of cases and 78.6% of controls of the test set. After cross-validation of the generated tree we could correctly classify 31 of the 42 cases (74% cases correctly classified) and 30 of the 42 controls (71% of controls correctly classified) (Table 1). Of the 11 peaks that were predictor variables used in the generation of the classification and regression trees, only the tree containing 4 terminal nodes was deemed optimal and thus was used for the analysis. Seven peaks were used in the generation of this tree with only 3 peaks used as splitting factors in this tree analysis. These peaks were m/z 7850.989, 9303.888, and 9190.488, with peak m/z 7850.989 being the most significant in differentiating between the two groups (p-value of 0.039). A list of peaks with significant p-values is provided in Table 2. Overall, sensitivity (correctly classified cases) and specificity (correctly classified controls) were similar and moderately impressive in predicting future breast cancer risk. Unfortunately, this SELDI-TOF MS data was re-analyzed by an independent institution with blinded-serum samples and was not found to have statistically relevant sensitivity and specificity. A peak probability contrast (PPC) procedure was utilized for this analysis (117).

Table 1. Classification and regression tree analysis of 84 serum samples processed on IMAC chips.

Class	N Cases	N Misclassified	N Correctly classified	% Correctly classified
Case	42	11	31	73.8%
Control	42	12	30	71.3%

Table 2. Significant peaks (p-value ≤ 0.05) differentiating between cases and controls after SELDI-TOF MS analysis.

<i>m/z</i>	p-value
3992.462	0.003
4184.852	0.003
7850.941	0.039
8157.624	0.050
9190.488	0.005
9303.888	0.009
9439.381	0.031

Analysis of samples processed with scrape method using MALDI-TOF MS

The 84 samples were also processed using the ClinProt robot with MB-WCX (found previously by our lab to yield the most robust peaks within the MALDI-TOF MS mass range) and robotically spotted on an AnchorChip plate with CHCA matrix. The resulting spectra were imported into ClinProTools 2.0 software and genetic algorithm models were generated. These models were then used to externally validate a set of 112 samples (56 controls and 56 cases) run in duplicate. Though increasing the number of peaks in the genetic algorithm model beyond 5 peaks improved the internal cross-validation, it resulted in lower sensitivity and specificity of the external validation (Table 3). Additionally, the models that had less than 5 peaks also had decreased sensitivity and specificity compared to the 5 peak model. Thus, the genetic algorithm model containing 5 peaks was deemed the most ideal and models containing more than 5 peaks were probably over-fitted to the sample set of 84. The genetic algorithm model containing 5 peaks had a 100% recognition capability of the test set and yielded an overall 63.64% cross-validation with 59.87% correctly classified cases and 67.42% correctly classified controls. Additionally, of the 112 (224 total when in duplicate) samples used for external validation of this model, 60.7% were classified correctly as cases and 61.6% were correctly classified as controls using this 5 peak genetic algorithm model. Table 4 shows the peaks utilized in the genetic algorithm model and Table 5 shows the top significant peaks as determined by T-test/ANOVA. Figure 5 shows the cluster plot for the set of 84 samples using the two peaks with the most significant p-values as determined by T-test/ANOVA. Additionally, a set of 96 blinded samples was run through the genetic

Table 3. Cross-validation and External Validation of genetic algorithm models

3a. 5 peak genetic algorithm model		
Class	Correct Classified (cross-validation)	Correct Classified (external validation)
Cases	59.9 %	60.7 %
Controls	67.4 %	61.6 %

3b. 6 peak genetic algorithm model		
Class	Correct Classified (cross-validation)	Correct Classified (external validation)
Cases	64.5 %	62.5 %
Controls	67.4 %	45.5 %

3c. 7 peak genetic algorithm model		
Class	Correct Classified (cross-validation)	Correct Classified (external validation)
Cases	73.68 %	53.6 %
Controls	68.54 %	42.9 %

Table 4. Masses used for the classification of 84 WCX fractionated serum samples using the 5 peak genetic algorithm model

Mass (m/z)	Weight (importance given in model)
2139.38	0.25
2675.72	0.09
3542.28	0.04
1781.11	0.13
1501.3	0.1

Table 5. Significant peaks as determined by T-test/ANOVA (p-value \leq 0.05) for the 84 serum sample set

Mass (m/z)	P-value
2178.28	0.043
2511.78	0.043
7342.55	0.043
2975.96	0.043
3706.07	0.043
3276.18	0.046
1716.00	0.046
5700.62	0.046
4742.36	0.046
3366.19	0.046
6524.28	0.046
7117.81	0.046
1911.81	0.046
5262.98	0.046
3839.32	0.046
3492.29	0.046
5247.62	0.046
6906.83	0.046
3352.33	0.046
4350.35	0.046
2578.2	0.046
1487.85	0.046
3120.22	0.046
4370.59	0.046
1377.35	0.046
7055.13	0.046
1138.63	0.046

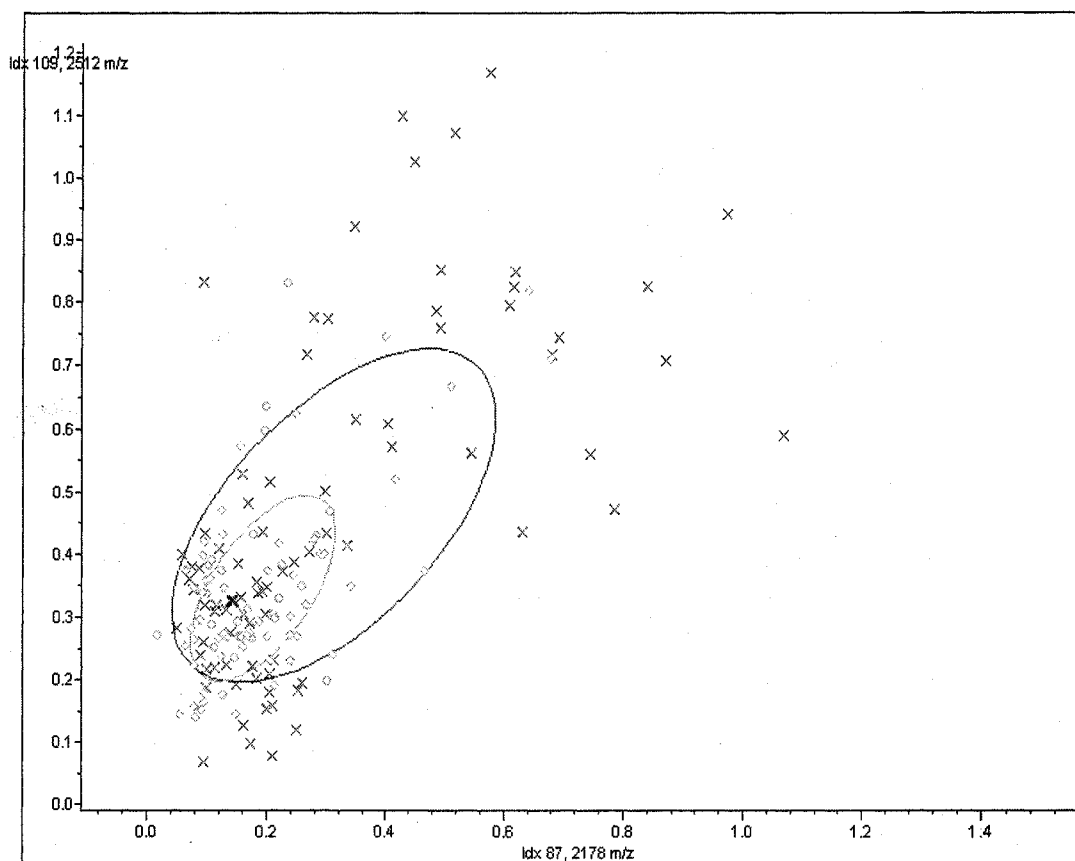


Figure 5. Cluster plot for the set of 84 WCX fractionated serum samples. The cluster plot was generated by the ClinProTools 2.0 software. The intensities of 2 peaks with the most significant p-values (as determined by T-test/ANOVA) are plotted on 2 axes. The more clustered the points are in relation to their group and the more separated the clusters are from each other then the more significantly distinct the 2 groups are in relation to each other. In this cluster plot the intensities of peak m/z 2178.28 are found along the “x” axis, while the intensities of peak m/z 2511.78 along the “y” axis. The control sample peaks are designated by “o” and the case sample peaks are designated by “x”.

algorithm model generated from the 84 samples. Of the 96 blinded samples, 52.1% were classified correctly with 51.0% of the cases correctly classified and 53.1% of the controls correctly classified.

In order to evaluate if the classification of the unknowns may be improved, an additional model was generated using the 84 samples with the 112 samples (n=196 run in duplicate). The set of 96 blinded samples was run through the various genetic algorithm models generated from the 196 samples and the 5 peak model was found to have the best classification ability. The genetic algorithm model containing 5 peaks had a 100% recognition capability of the test set and the cross-validation yielded an overall 57.34% correct classification with 58.2% sensitivity and 56.4% specificity. Table 6 shows the peaks utilized in the genetic algorithm model (the top significant peaks as determined by T-test/ANOVA was m/z 7341.66 with a p-value of 0.04). Figure 6 shows the cluster plot for the set of 196 samples using the two peaks with the most significant p-values as determined by T-test/ANOVA. Of the 96 blinded samples, 52.1% were classified correctly with 58.3% of the cases correctly classified and 45.8% of the controls correctly classified. Thus, the sensitivity was slightly improved in the classification of the unknowns as compared to the genetic algorithm model generated from the 84 samples; however the specificity was slightly decreased.

The 112 samples were also analyzed using MB-IMAC beads as a comparison to the IMAC chip SELDI data. A genetic algorithm model containing 5 peaks was found to have a 100% recognition capability of the test set and the best cross-validation with 60.8% correctly classified cases and 61.8% correctly classified controls. However, the classification results of the 96 independently run blinded samples was disappointing. Of

Table 6. Masses used for the classification of 196 WCX fractionated serum samples using the 5 peak genetic algorithm model

Mass (m/z)	Weight (importance given in model)
2757.08	0.05
6056.85	0.03
1733.82	0.06
4144.51	0.01
4642.61	0.04

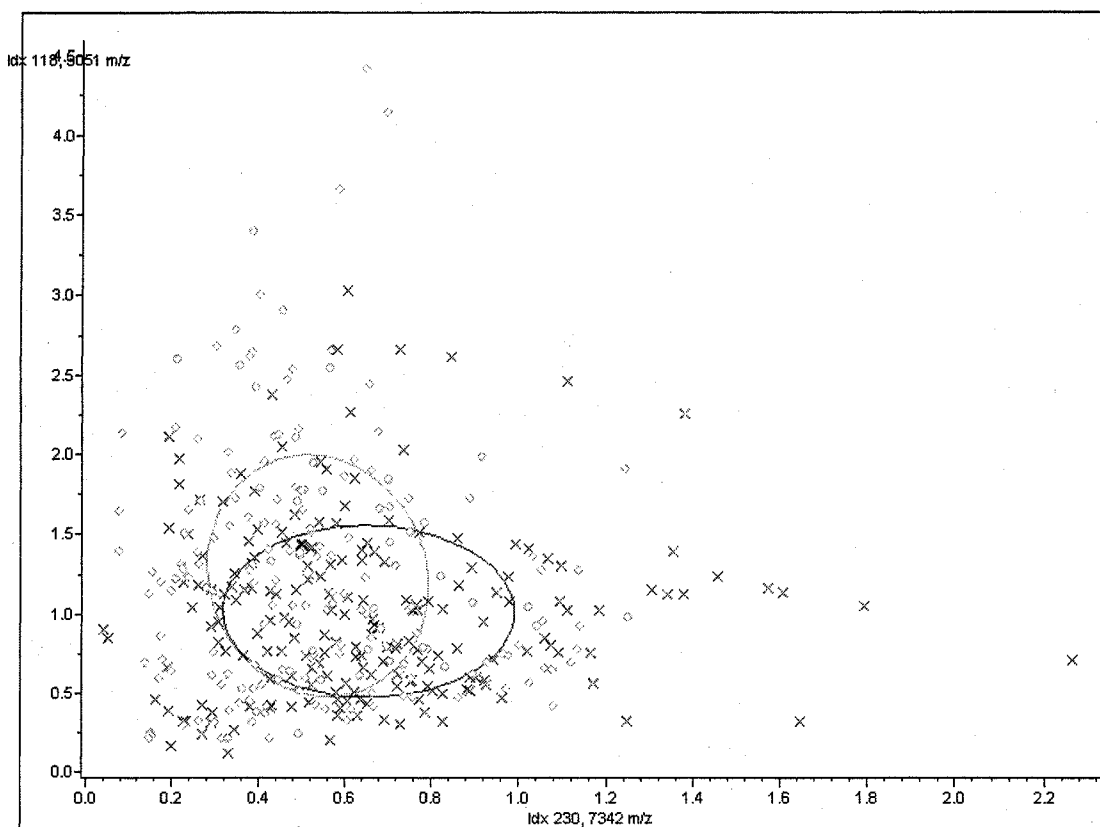


Figure 6. Cluster plot for the set of 196 WCX fractionated serum samples. The cluster plot was generated by the ClinProTools 2.0 software. The intensities of 2 peaks with the most significant p-values (as determined by T-test/ANOVA) are plotted on 2 axes. The more clustered the points are in relation to their group and the more separated the clusters are from each other then the more significantly distinct the 2 groups are in relation to each other. In this cluster plot the intensities of peak m/z 7341.66 are found along the “x” axis, while the intensities of peak m/z 3051.22 along the “y” axis. The control sample peaks are designated by “o” and the case sample peaks are designated by “x”.

the 96 blinded samples, 43.3% were classified correctly with 40.6% of the cases correctly classified and 45.8% of the controls correctly classified. The cross-validation of the IMAC sample set appeared to have a comparable sensitivity and specificity as the cross-validation of the models from WCX samples; however this was not true for the classification of the blinded sample sets. Table 7 shows the peaks utilized in the 5 peak genetic algorithm model used for unknown classification (there were no significant peaks, p -value <0.05 , as determined by T-test/ANOVA). There appeared to be no similarities between the peaks used for the tree analysis of the SELDI-TOF MS data and the peaks used for the genetic algorithm analysis of the MALDI-TOF MS data. Figure 7 shows the cluster plot for the set of 112 samples using the two peaks with the most significant p -values as determined by T-test/ANOVA.

3.4 Discussion

We have demonstrated a profiling scheme on an interesting and novel sample set, which may prove helpful in finding biomarkers to predict a woman's risk of developing breast cancer. The cases in this sample cohort were randomly chosen from a group of women who developed invasive breast cancer during the first 10 years of follow-up and who had stored serum available that had been drawn between 1 and 5 years prior to the diagnosis of breast cancer. The controls were randomly selected, matched to cases on age and length of follow-up, from the remaining participants in the SOF cohort who were not diagnosed with breast cancer during the first 10 years of follow-up. Thus, this serum cohort contains proteomic information of women that will develop invasive breast cancer

Table 7. Masses used for the classification of 112 IMAC fractionated serum samples using the 5 peak genetic algorithm model.

Mass (m/z)	Weight (importance given in model)
5916.51	0.13
2549.66	0.14
2606.95	0.12
5336.53	0.07
1548.85	0.01

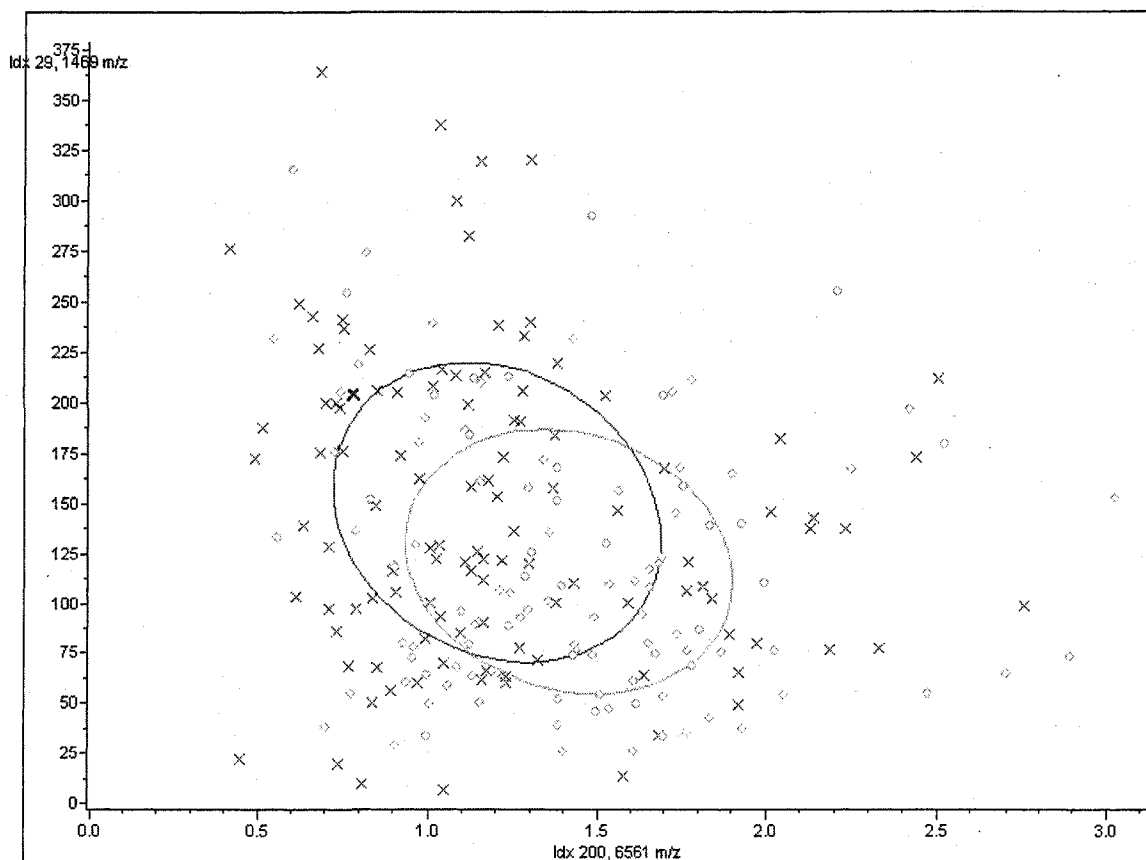


Figure 7. Cluster plot for the set of 112 IMAC fractionated serum samples. The cluster plot was generated by the ClinProTools 2.0 software. The intensities of 2 peaks with the most significant p-values (as determined by T-test/ANOVA) are plotted on 2 axes. The more clustered the points are in relation to their group and the more separated the clusters are from each other then the more significantly distinct the 2 groups are in relation to each other. In this cluster plot the intensities of peak m/z 6560.83 are found along the “x” axis, while the intensities of peak m/z 1468.88 along the “y” axis. The control sample peaks are designated by “o” and the case sample peaks are designated by “x”.

1-5 years into the future and how they may differ from women that will not develop breast cancer in that same time frame.

The main success from Aim 1 is the validation of the scrape technique for the use of sparing precious samples unnecessary freeze-thaw cycles. The results showed that there was no significant difference in quality of data between scraping a frozen sample and thawing the whole sample. A main concern for the scrape technique would be the unequal distribution of proteins throughout the frozen sample, with a fractionation of proteins based on biochemical properties and/or with a predominance of ice crystals at the top of the frozen sample. The unequal distribution of proteins throughout the frozen sample and the predominance of ice crystals on the top of the sample are especially concerning for samples that are allowed to freeze slowly. Generally, slow freezing causes large, non-uniform ice crystals to form and freeze concentrations to occur, where solutes and protein molecules are pushed into non-frozen regions, producing a large increase of solute and protein concentrations. The faster the freezing process the more nucleation is promoted and the greater the number of ice crystals of smaller size that will result, thus making the frozen sample more uniform (118). However, we seemed to have avoided the problem of unequal distribution of proteins and solutes within the sample. If the top of the tube to be scraped mainly contained "sample" that was composed of ice crystals then we would have seen a very dilute amount of peptide/proteins in our SELDI-TOF MS spectra, if we saw anything at all. Additionally, as stated before there seemed to be no significant changes between the spectra from the samples that were scraped first and the same samples that were subsequently thawed and mixed. This technique also illustrates the importance of storage and sample handling during the collection of

samples. MALDI-TOF MS (18, 115) and SELDI-TOF MS are reproducible (50) as far as instrumentation performance, however much thought into the standardization of protocols must occur for this to be correct for the true comparison of sample groups. McLerran et al found that initial discrimination, as seen by SELDI-TOF MS, between serum from patients that have been diagnosed with benign prostatic hyperplasia and those diagnosed prostate cancer may have been due to storage time variability between the two groups, thus leading to a general bias in the sample analysis process (119).

In our study the initial modeling results from the scraped serum sample cohort were promising; with MALDI-TOF MS ClinProTools 2.0 software generating models with 100% recognition capability of the test set groups, cases and controls. SELDI-TOF MS had an overall 82.15% recognition capability of the test set groups. However, the cross-validation was less than ideal with the best SELDI-TOF MS regression tree algorithm having an overall 74.6% recognition capability between groups. The best MALDI-TOF MS genetic algorithm model was generated using the WCX fractionation scheme with the initial 84 sample set and yielded an overall 71.1% recognition capability between groups. This 7 peak genetic algorithm model proved to be over-fitted with the 84 sample set and yielded low external validation sensitivity and specificity. The genetic algorithm model that performed the best in the external validation, with an overall recognition capability of 61.2% between cases and controls, was the 5 peak model generated from the 84 serum sample set. However, this model performed poorly when used again to identify case and control status of 96 blinded samples (overall recognition capability of 52.1%). Additionally, the SELDI-TOF MS data set was unable to correctly classify the majority of the blinded sample set as determined by an independent lab.

Interestingly, this epidemiology group also found that that the MALDI-TOF MS data was more reproducible between duplicate samples than the SELDI-TOF MS data. This analysis was performed using the peak finding and alignment algorithm from the PPC procedure. Briefly, for each pair of replicates, the peaks were aligned and deemed discordant or concordant peaks. If both spectra possess a peak at a specific point, then that site has a concordant peak. However, if only one spectrum has a peak at a particular sites and the other does not, then that site has a discordant peak. Thus, MALDI-TOF MS data was found to have a lower percentage of discordant peaks than SELDI-TOF MS data.

Overall, this is a very difficult sample set since we are asking the peptide/protein profiles in the serum to predict the future, not the current state of the women at blood draw. Breast cancer is a clinically heterogeneous disease with histological type, grade, tumor size, lymph node involvement, ER and HER-2 receptor status all influencing prognosis and response to available therapies (120). Thus, to provide better risk assessment models we may need to interrogate the sample sets with more specific guidelines. In this study age was controlled for as all patients were age-matched and postmenopausal, both characteristics which are risk factors for breast cancer. Additionally, two other main variables were already taken out of the equation. African-American women were excluded during the osteoporotic fracture study and thus were unavailable for our cohort. Additionally, women that were on hormone replacement therapies (HRT) were excluded from the selection process for the breast cancer prediction study. Race is a known risk factor for developing breast cancer, as African American and Hispanic women are more prone to present with advanced breast cancer than Caucasian

women. This is a complex issue involving both biological differences and certain socioeconomic conditions, such as access to the same quality health care and screening (79). HRT, especially in the long term, has also been recently implicated as a risk factor for breast cancer (121). Additionally, women that are postmenopausal and on HRT would more than likely have distinct peptide/protein secretion patterns compared to women that are postmenopausal and do not have stimulation of estrogen signaling pathways through HRT. Other variables that may be minimized to increase sensitivity and specificity of detecting a particular woman's risk of developing breast cancer is stratifying samples based on their *BRCA* and HER2 status. As discussed in Chapter I, mutations in the *BRCA* genes are known risk factors for breast cancer development and the presence of HER2 receptor fragments in serum has been shown to directly correlate to the aggressive nature of the breast cancer and its subsequent treatment. By stratifying samples according to these and other variables (which will be further discussed in the future directions section of Chapter VI) we may begin to focus more on biomarkers related to breast cancer risk then on biological variables compounded between various patients.

One facet hampering the discovery of breast cancer biomarkers is the overall complexity and large dynamic range of the blood proteome and the relatively low abundance of these cancer biomarkers in the blood as compared to other proteins. For example, early-stage tumors might arise within a tissue volume of less than 0.1mL, thus making the dilution factor of the tumor-generated biomarkers about 50,000 (assuming that the biomarkers attributed to this tumor are uniformly dispersed in the 5,000mL total blood volume) (56) . In this study there may not even be an early-stage tumor in the

classical sense, but rather a condition within the breast tissue that will promote the growth of said tumor. One approach that other researchers have taken is to work with proximal fluids that are regionally closer to the tumor such as nipple aspirate fluid (NAF) and proximal or tumor breast tissue (91, 122) in hopes of later being able to detect, or correlate, the discovered biomarkers in the blood for patient screening. However, for our study, in which a valuable and interesting cohort of serum is already available, and for individuals that are interested in discovery of biomarkers in the same fluid type as will later be used for patient screening, reducing the complexity of the serum proteome using various fractionation techniques is a pre-requisite. There are several fractionation techniques available to further dissect the proteome such as the depletion of the top most abundant proteins, lectin-capture strategies for the targeted capture of glycoproteins (since glycan changes in proteins have been linked to cancer disease states (71, 123, 124)) and tandem fractionation techniques. These extensive fractionation techniques are more readily adaptable to the MALDI platform, which typically employs front-end fractionation with paramagnetic beads conjugated to various chromatography chemistries (either chemical or biological) and would allow in-tandem use of these bead types in an automated manner. However, as discussed in Chapter I, the SELDI-TOF MS utilizes flat chips coated with different surface chemistries that have smaller surface areas than the paramagnetic particles and thus lead to less efficient fractionation.

Another problem facing researchers profiling for breast cancer biomarkers is that differential spectra patterns are not complete for validation purposes without the knowledge of the identity of the peptides/proteins behind the peaks. As discussed in Chapter I, the linear TOF mode typically utilized with the SELDI and MALDI platforms

(as it was in this Aim), does not yield information about the identity of peptides/proteins. Thus, profiling in the range of reflectron mode, which has the capability on the MALDI-TOF/TOF for tandem-MS identification of peaks, may prove useful to ascertain additional, complementary information of a specific sample set. A caveat to this complementary technique is that endogenous peptides/proteins found in the range of the reflectron mode are sparse and not robust enough for useful, informative profiling. Thus, this problem will be addressed in the next Aim.

CHAPTER IV

AIM II. INCREASING THE EFFECTIVENESS OF THE MALDI-TOF/TOF FOR ANALYSIS OF LARGE MOLECULAR WEIGHT PROTEINS

4.1 Introduction

Biomarker discovery is an ever evolving research area spurred by advances in technology and improvements in clinical study design and bioinformatics strategies (125). Typically, in order to reduce the sample complexity of high protein concentration fluids like serum and plasma, chemical affinity capture using beads or chip surfaces has been employed along with time-of-flight mass spectrometry to generate comparative spectral peak profiles. These approaches, also termed expression profiling, can be automated for relatively high throughput and generally consume small amounts of clinical sample (44, 45). Additionally, expression profiling can be reproducible and portable across multiple laboratories, especially when rigorous study design and sample handling are combined with carefully controlled instrument calibration, automated sample preparation, and supervised bioinformatic data analysis (47, 50, 62, 126). Nevertheless, the difficulty in determining the protein identities of potential biomarker peaks, and a concern that the sensitivity and dynamic range of prevalent proteins in serum or plasma prohibits identification of proteins associated with disease continues to hamper these expression profiling approaches (48, 49, 127). Recently, however, the development of TOF/TOF technology has brought with it the capability of protein identification (4,000 m/z or less) through the generation of fragment ions and subsequent homology searching (19).

Most MALDI-TOF based expression profiling studies only examine endogenous low molecular mass constituents (1-20 kDa) of serum or plasma. Yet most proteins captured on a particular surface or bead are not effectively resolved in the MALDI instruments used due to their larger sizes (>20 kDa). In this regards, we report a workflow of magnetic bead-based chromatography surfaces and immobilized trypsin to generate peptide profiles reflective of the broader range of proteins captured in front-end purification and fractionation strategies applied to complex clinical fluids like serum or plasma. This is essentially a “bottom-up” approach (43), but tailored for the MALDI-TOF as the generation of tryptic peptides increases the breadth of proteins detected and provides peak masses ideal for LIFT-MALDI-TOF/TOF sequencing identification.

Using pooled human serum samples, two different workflow combinations of chromatography beads with the immobilized trypsin beads are described. We found that the bead-based trypsinization method was highly reproducible and efficient in digesting large serum protein fractions at short incubation times, and that the resulting peptides were readily able to be identified by LIFT-MALDI-TOF/TOF. Representative lists of proteins present in a pooled healthy serum sample are presented. Additionally, as a proof-of-concept for clinical application, the method was used in two serum profiling studies. One such study utilizes the aforementioned SOF sample cohort from Aim I. A separate serum study has the goal of detecting differences between individuals diagnosed with benign prostatic hyperplasia (BPH) and individuals diagnosed with prostate cancer (PCa). Similar to the SOF samples, which are designed to assist in the prediction of a woman’s breast cancer risk years into the future, the question posed by the BPH/PCa sample set is also very difficult to answer. As discussed in more detail in the first

chapter, prostate specific antigen (PSA) is the current biomarker used in the diagnosis of PCa. Unfortunately, it is not a specific biomarker for prostate cancer since its serum level is affected by many other factors such as inflammation, ejaculation and BPH. It has been estimated that two out of three men with abnormal results on routine PSA screening will not have prostate cancer (108). Thus, a diagnosis of BPH or PCa is often mistaken for the other, resulting in men being exposed to unnecessary medical intervention and anxiety. The PSA levels of the men in this study fall between 2 and 10 ng/mL, which is a particularly grey area for doctors, as these levels are considered elevated by today's standards, but are not high enough to for a confident PCa diagnosis. Using these two sample cohorts we demonstrate the reproducibility of the trypsin bead method with clinical samples and showcase typical workflow strategies that may be applied for the purpose of biomarker identification.

4.2 Materials and Methods

Serum Samples

A pooled human serum sample collected from over 360 donors (50) was used for method development procedures.

For the proof-of-concept tryptic analysis of clinical serum samples, pools of each group were generated. For the prostate cancer study, BPH and PCa confirmed-diagnosis patients with elevated levels of PSA (range 2-10 ng/mL) were pooled in the following manner: 10 pools for BPH and 10 pools for PCA with each pool containing 6 samples. For the SOF samples, the samples were pooled in the following manner: 12 pools for control and 12 pools for cancer with each pool containing 8 samples.

Magnetic bead-based fractionation

Initial fractionation of serum was done with either MB-WCX (weak cationic exchange) or MB-WAX (weak anionic exchange) paramagnetic beads essentially as described by the manufacturer's protocols (Bruker Daltonics, Bremen, Germany). Briefly, 20 μ L of serum was mixed with 40 μ L of binding solution (the WAX protocol utilized a pH 5 binding solution) supplied with the beads and 20 μ L MB-WCX or MB-WAX beads (note that the WAX beads were equilibrated with activation solution prior to this step) for 15 minutes (mixing every 5 minutes). A magnetic bead separator was used to concentrate the beads and for the wash/rinse processes. Unbound serum proteins were removed and the beads were washed 3 times with 100 μ L of MB-WCX or MB-WAX wash solution. Bound serum proteins were eluted with 10 μ L of MB-WCX or MB-WAX elution solution supplied by the manufacturer. Finally, 8 μ L of HPLC water and 1 μ L of MB-WCX stabilization solution were added to the WCX eluate (this step was added during method development and it is stated in the Results when it was incorporated into the protocol), and 11 μ L of MB-WCX elution was added to the WAX eluate, to give a final sample pH of 7.5-8.5.

For reduction and alkylation, 8 μ g of the fractionated samples were reduced with 8 mM DTT in 25 mM ammonium bicarbonate (pH 7.8) at 56 $^{\circ}$ C for one hour (24 μ L total volume). The reduced samples were then alkylated with 17 mM iodoacetamide in 20 mM ammonium bicarbonate total solution (29 μ L total volume).

Liquid and magnetic bead-based trypsinization

Sequencing grade trypsin (Roche, Basel, Switzerland) was re-suspended in 50 mM ammonium bicarbonate / 4% acetonitrile (ACN) to a final concentration of 40 ng/ μ L. For each reaction in the comparative soluble trypsin study, 200 ng of trypsin was added to the reduced and alkylated samples yielding a 40:1 serum protein to trypsin ratio (other ratios were utilized and are indicated in the figure legends) and incubated for 30 minutes and overnight at 37⁰C. Paramagnetic immobilized trypsin, EnzyBeadsTM Trypsine (Agro-Bio, La Ferte Saint Aubin, France), were initially washed with 25 mM ammonium bicarbonate (pH 7.8). Twenty microliters of reduced/alkylated samples were trypsinized with 25 μ L beads as described by the manufacturer for 30 min. at 37 ⁰C. This is the equivalent of 3 units of enzyme activity per reaction, with one unit defined as the amount of EnzyBeads Trypsine required to hydrolyze 1 μ mole of chromogenic substrate in one minute at 25⁰C. Digested peptides were removed from the beads that were held in place by a magnetic separator.

Sample clean-up and concentration

Initially, ZipTipC18 cartridges (Millipore, Billerica, MA) were used to clean-up and concentrate the digested sample. The 10 μ L of trypsinized sample was acidified with 1 μ L of 1% TFA and allowed to bind to the C18 cartridge. The C18 cartridge was washed with 0.1% TFA and the sample was eluted in 5 μ L of 50% ACN. Later, this method was adapted so that tryptic peptides were re-captured and concentrated with Hydrophobic Interaction Chromatography (HIC)-C18 paramagnetic beads (Bruker Daltonics, Bremen, Germany) as follows. Twenty microliters of the tryptic digest was

incubated with 10 μ L HIC-C18 beads and 40 μ L HIC-C18 binding buffer (or were noted with HIC-C8 beads and binding buffer). Bound peptides were washed with the manufacturer's wash solution and eluted in 10 μ L of 50% ACN as per the manufacturer's specifications.

MALDI-TOF/TOF

Two microliters of the tryptic peptide sample after HIC-C18 clean-up was mixed with 4 μ L of CHCA matrix solution (4 mL ethanol, 2 mL acetone, 0.008 g CHCA and 0.1% TFA) and 1 μ L of the mixture was manually spotted (or robotically spotted by the ClinProt robot where indicated) onto an AnchorChip plate using a dried droplet spotting technique. Also, where noted, a reverse thin-layer spotting technique was used where 1 μ L of the tryptic peptide sample was overlaid with 2 μ L of CHCA matrix. The spotting techniques found ideal for untrypsinized samples were as follows: For untrypsinized WCX fractionated samples, the samples were mixed 1:15 with matrix and for untrypsinized WAX fractionated samples, the samples were spotted using the thin-layer method (1 μ L sample overlaid with 2 μ L matrix). Additionally, where noted, a matrix formulation (ACN, Acetone, 0.1% TFA, CHCA) was used for LIFT-MALDI-TOF/TOF analysis.

UltraFlexI and UltraFlex III MALDI-TOF/TOF instruments (Bruker Daltonics) were used to analyze peptides in linear and reflectron modes. The resulting spectra were processed using FlexAnalysis and ClinProTools 2.0 software (Bruker Daltonics). The ClinProt software baseline subtracted and normalized the spectra using total ion current and an m/z starting point of 800. A k-nearest neighbor genetic algorithm contained in

this software suite was used to generate prediction models to classify the groups analyzed. Twenty percent of the samples were left out of the model generation process and used to cross-validate the model within the software. Peaks of interest were further analyzed on a separate platform using the LIFT function of a MALDI-TOF/TOF Ultraflex III instrument. The BioTools software and the MASCOT search engine (www.matrixscience.com) were used to compare the TOF/TOF spectra against primary sequence databases (SwissProt) to determine peptide sequence identities (unless otherwise noted the search criteria is as follows: carbamidomethyl and oxidation modifications; 100 ppm mass tolerance MS; 0.5 Da MS/MS tolerance).

4.3 Results

Integrating trypsin digestion into a bead-based affinity fractionation workflow

The initial goal of this study was to integrate a trypsin digestion step following standard chemical affinity fractionation of serum samples, the latter being a common (off-line) first step in many serum/plasma proteomic profiling studies (50, 62, 125, 128, 129). We hypothesized that inclusion of the trypsin digestion would facilitate more direct protein identifications of high mass novel proteins by generating peptides in an optimal mass range for detection and sequence identification by MALDI-TOF instruments ($\leq 4000 m/z$). This approach could also broaden the dynamic concentration and mass range of detected proteins.

The common approach to digestion of complex protein samples like serum is to perform an in-solution digest with added, soluble trypsin. A shortcoming of trypsin in-solution protocols is long reaction times (4 -16 hours) and the autocatalytic activity of the

trypsin may create contaminating trypsin cleavage products. One solution to this has been to use trypsin immobilized on a solid support, which acts to stabilize the trypsin and greatly increase the concentration of trypsin that can interact with substrate leading to a more rapid digestion (31). For these studies, we utilized a newly developed trypsin product immobilized on paramagnetic beads, EnzyBeadsTM Trypsine, which can quickly and efficiently remove the trypsin from the digestion reaction by placing the reaction tube against a magnet.

We first assessed the importance of initially reducing and alkylating the sample before the trypsinization step. We used the reducing agent DTT (dithiothreitol) and the alkylating agent iodoacetamide. DTT reduces disulfide bonds and maintains monothiols in a reduced state. After reduction, the sulfhydryls are then reacted with iodoacetamide to prevent reformation of disulfide linkages in a random manner (130). Figure 8 clearly shows that for these secreted proteins, which have disulfide bonds, reduction and alkylation improves the digestion efficiency. Thus, the reduction/alkylation step was included in the trypsin digestion protocol.

Another observation that was made during the development phase of the trypsin digestion workflow pertained to the MALDI-TOF spectra. It was noted that the signal intensity was weak and that the matrix/sample spots on the target plate were not uniform (i.e. there was a propensity towards “hot-spot” formation), regardless of the spotting technique (i.e. dried droplet or reverse thin-layer technique). It is known that certain reagents negatively effect matrix crystallization and hinder ionization of the sample (11). Thus, we first investigated whether ZipTipsC18 would improve the spectra and the uniformity of the spot on the MALDI AnchorChip plate. These ZipTips have C18 (18

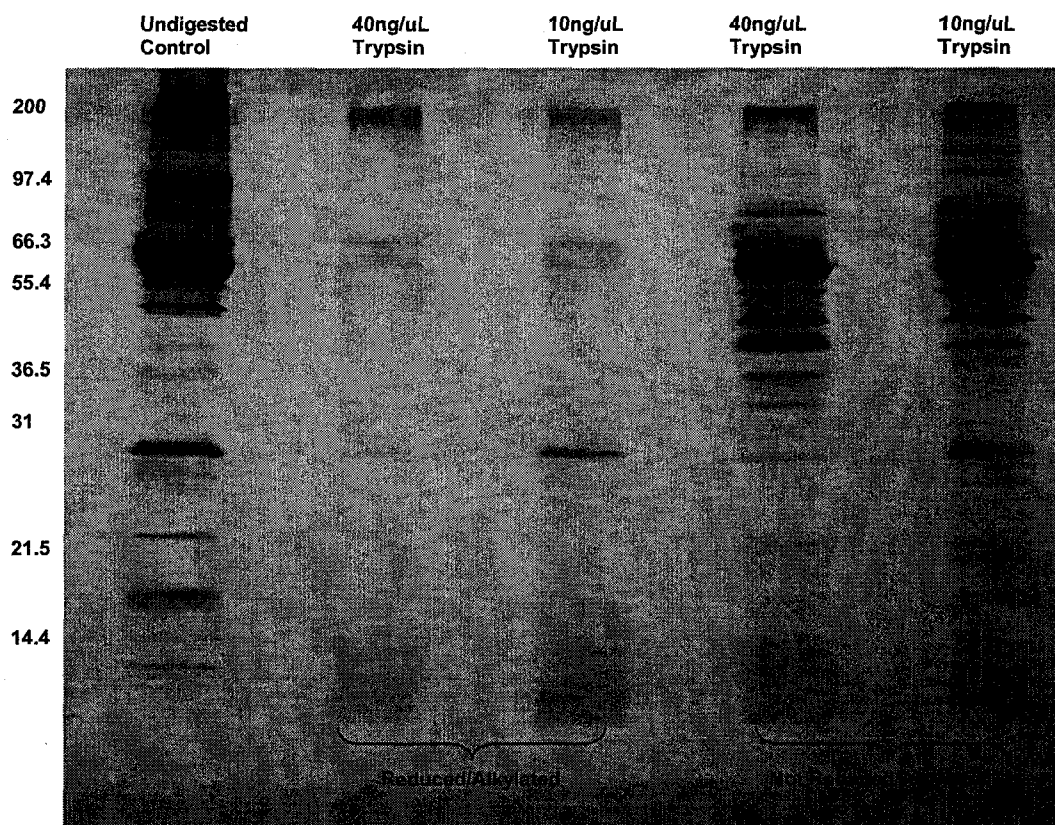


Figure 8. Improved trypsinization efficiency of WCX fractionated serum proteins after reduction and alkylation. A Biorad Criterion Tris-HCl 8-16% is shown. The lanes labeled 40 ng/ μ L of trypsin utilized a 1:30 soluble trypsin to protein ratio, while the lanes labeled 10 ng/ μ L of trypsin utilized a 1:120 soluble trypsin to protein ratio. Three micrograms of protein from each condition was loaded on the gel. The gel was silver stained using a Biorad reagent.

carbons) derivatized surfaces that bind peptides in aqueous solutions through hydrophobic interactions, allowing small interfering molecules (i.e. salts, buffers, and chaotropes) to be washed off. The peptides are then eluted with various organic solvents that are compatible with the MALDI-TOF technique (11). In Figure 9 the improvement in spectra quality in a sample with clean-up as compared to a sample without clean-up is shown. Hence, a final clean-up and concentration step was added into the trypsin digest workflow.

Since, ZipTips are labor-intensive and would later be difficult to accommodate to an automated workflow, we investigated whether magnetic beads with immobilized carbons could substitute. We found no significant difference between the ZipTipC18 and the C18 treated magnetic beads (Bruker Daltonics). However, there are other hydrophobic interaction chromatography (HIC) magnetic beads available as well. In fact, various papers published by Bruker utilize the C8 magnetic bead-type for capture of peptides (131). Thus, we compared which bead type yielded the best results in terms of the capture and concentration of our trypsinized peptides. Simultaneously, we also assessed whether we would gain more peptide information through the sequential elution of the HIC-magnetic beads. Figure 10 shows the sequential elutions off of the C18 HIC-magnetic bead versus the sequential elutions off of the C8 HIC-magnetic bead. When comparing the initial elution (which yielded the most robust spectra), using the FlexAnalysis software, for both bead types, the spectra of tryptic peptides after C18 purification yielded 41 peaks between m/z 1000 and 4000, while the spectra of tryptic peptides after C8 purification yielded only 20 peaks between m/z 1000 and 4000. This mass range is most ideal for sequence identification using LIFT-MALDI-TOF/TOF and

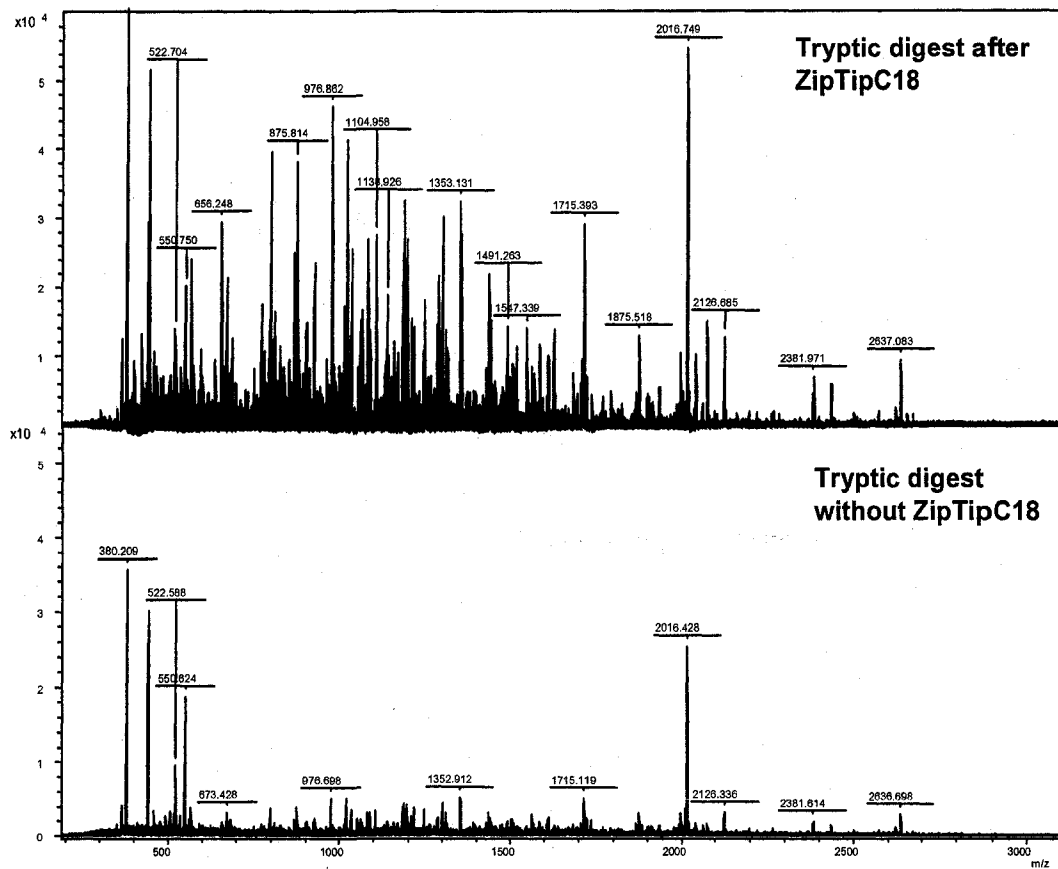


Figure 9. Effect of ZipTipC18 clean-up on MALDI-TOF spectra of WCX fractionated samples trypsinized with immobilized trypsin beads. The top panel showcases a sample that was concentrated and cleaned-up using ZipTipC18, while the bottom panel shows the same tryptic digest not processed with ZipTipC18. Samples were processed using MB-WCX fractionation and immobilized-bead trypsinization, spotted on an AnchorChip plate using a reverse thin-layer method and analyzed on the MALDI-TOF Ultraflex I in reflectron mode as stated in Materials and Methods.

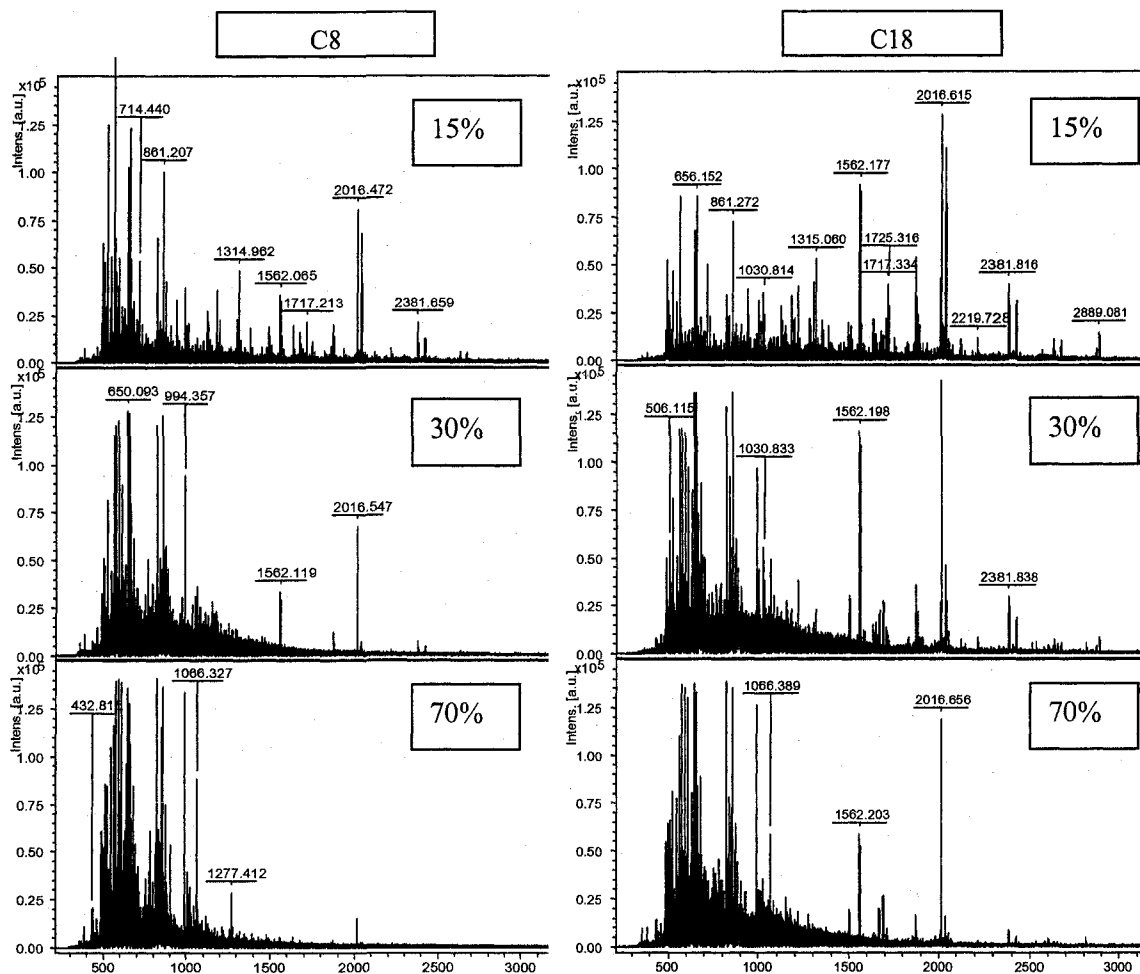


Figure 10. Comparison C8 and C18 HIC-magnetic bead sequential elutions. The panel on the left shows sequential elutions (15, 30 and 70% of ACN) off of the MB-C8 bead of tryptic peptides after MB-WCX fractionation and immobilized-trypsin bead digestion. The panel on the right shows sequential elutions off of the MB-C18 bead of tryptic peptides derived by the same protocol. Samples were analyzed on the MALDI-TOF Ultraflex III in reflectron mode after being spotted (1:5 ratio of sample to CHCA matrix) by the ClinProt robot on an AnchorChip plate as described in Material and Methods.

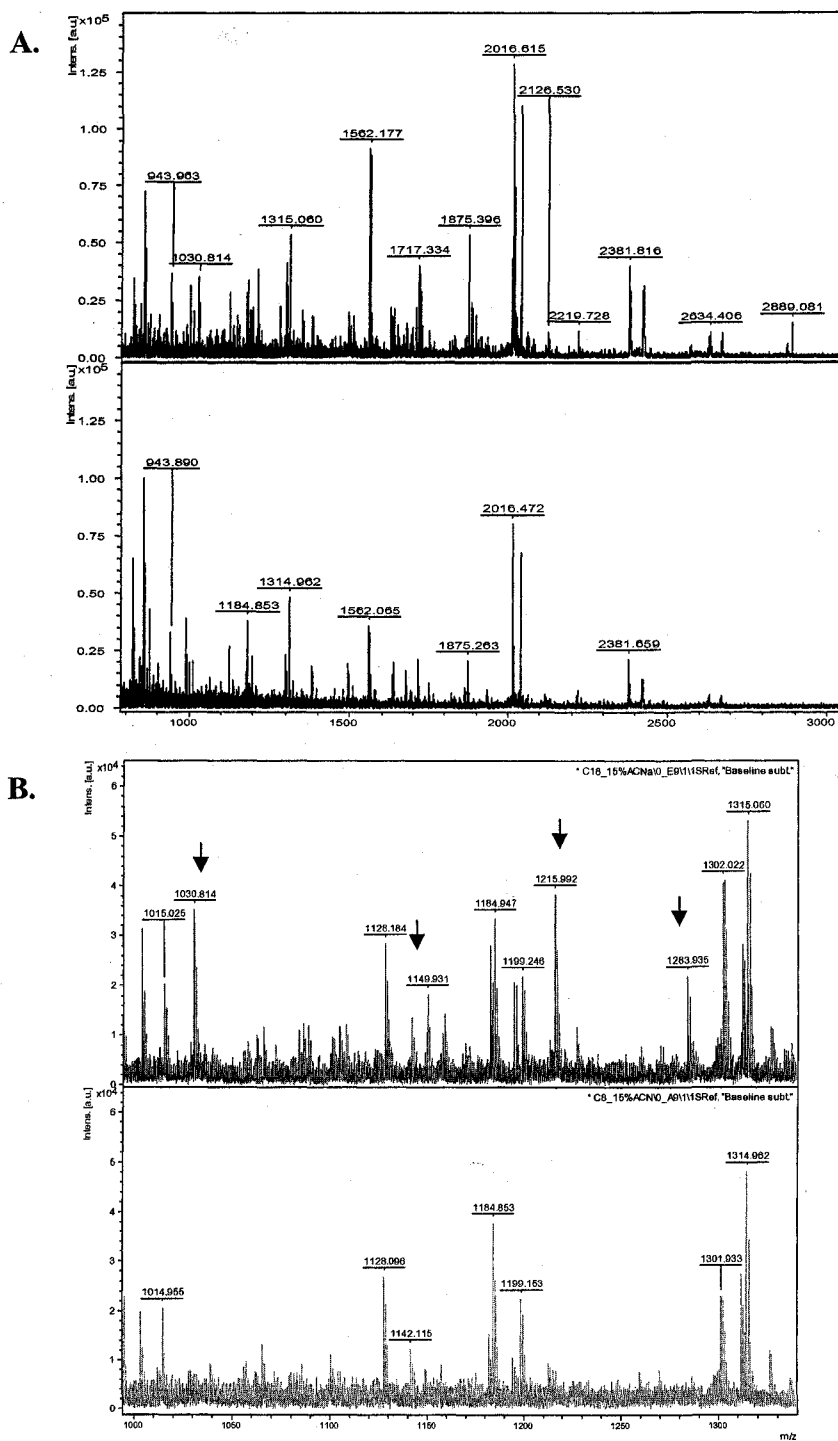


Figure 11. Comparison of C18 and C8 HIC bead 15% acetonitrile elutions. (A) The top panel shows a 15% ACN elution off of MB-C18 beads, while the bottom panel shows a 15% ACN elution off of MB-C8. (B) Peptides in the lower mass range (1000 – 1400 m/z) are seen captured by the C18 bead type (denoted by arrows in the top panel), but are lost by the C8 bead type (bottom panel). Samples were analyzed on the MALDI-TOF Ultraflex III in reflectron mode as stated in Figure 9 and Materials and Methods.

also discounts most matrix peaks. As shown in Figure 11, it is clear that there are certain peptides that are not captured by the C8 bead, but are captured by the C18 bead. This is expected as peptides are small and polar and thus need more carbons for successful binding, while larger proteins have more hydrophobic areas and thus will bind tenaciously to the available carbons (i.e. C4 or C8) (11). Therefore, we decided that we would continue on with the C18-HIC magnetic bead type. Additionally, during the course of this comparison, we found that the reverse thin-layer method, which we had utilized after the ZipTipC18 purification, would be impractical for an automated approach and also yielded less compact and uniform spots when compared to the dried droplet method (as is performed by the ClinProt robot). The dried droplet method also yielded more resilient spots that could withstand larger quantities of laser shots, which is advantageous for both profiling and LIFT identification.

During the optimization of the immobilized-trypsin bead method, we also determined the necessary protein concentration and pH of the fractionated sample for ideal trypsin efficiency and spectra quality. The SDS gel and spectra in Figure 12 illustrates that 8 μg is an optimal concentration for efficient trypsin digestion with immobilized-trypsin beads. Additionally, we found that digesting 20 μL of the digested sample and diluting this sample 1:3 with matrix minimized contaminating matrix peaks, while still producing robust tryptic spectra on the MALDI-TOF instrument. Using these specifications, 82 peaks were counted, when excluding matrix peaks. This approach works well with the C18 beads since, unlike ZipTipsC18, which concentrate 10 μL of digested sample into 5 μL of eluted sample; C18 HIC-magnetic beads concentrate 20 μL of digested sample into 10 μL of eluted sample. The elution buffer is a volatile 50%

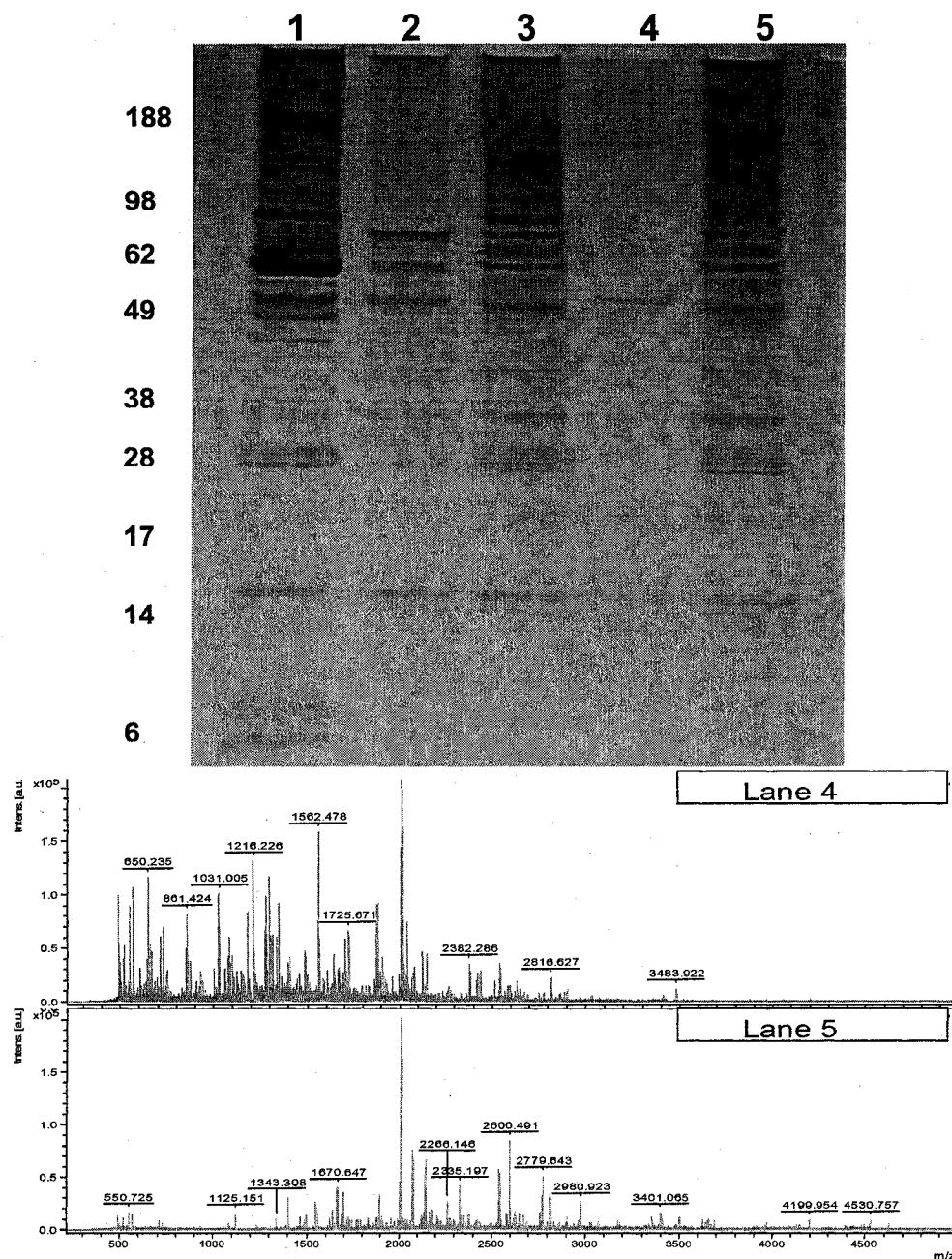


Figure 12. Concentration determination for ideal trypsinization using immobilized-trypsin beads. The gel shown is a NuPage 4-12% Bis-Tris with MES from Invitrogen. WCX fractionated sample was added varying amounts into reduction/alkylation and trypsinization reactions to generate the most efficient digestion. Lane 1: Undigested serum sample after WCX fractionation. Lane 2: 15.5 μ g of sample reduced/alkylated and 5.2 μ g was digested with immobilized-trypsin beads. Lane 3: 15.5 μ g of sample reduced/alkylated and 10.4 μ g was digested with immobilized-trypsin beads. Lane 4: 8 μ g of sample reduced/alkylated and 5.5 μ g (20 μ L of reduced/alkylated sample) was digested with immobilized-trypsin beads. Lane 5: 24 μ g of sample reduced/alkylated and 15.5 μ g was digested with immobilized-trypsin beads. Equivalent to 5 μ g of protein is loaded in each well. Gel was stained with a silver stain from Biorad. The lower panel shows the distribution of peaks from the best digest (Lane 4) and the worst digest (Lane 5).

acetonitrile solution, thus an elution with more volume allows for easier and more efficient handling that may be adapted to an automated workflow.

Trypsin works optimally at a pH between 7 and 9, however Figure 13 also shows that adjusting the pH to ~8 prior to reduction and alkylation (not just at the point of trypsin digestion) improves the digestion efficiency. This is important to know because each WCX fractionation preparation may vary slightly in terms of protein concentration, thus one would be adding more of the basic elution buffer (pH ~11) into the reduction reaction if the preparation had a lower protein concentration. These small variations turn out to be enough to alter the reduction reaction solution pH so that it affects digestion efficiency. Using the specification listed thus far (Figure 14), we found that this trypsin bead based method yielded on average a total of 85 peaks, discounting the matrix peaks. This is compared to 20 peaks that are seen on average with undigested WCX fractionated serum in the ideal range of reflectron mode for peptide identification. This clearly shows that trypsinizing the sample prior to MS analysis yields more peptides in the range of optimal MALDI-TOF/TOF analysis and thus may lead to direct identification of peptides via LIFT.

Comparison of free trypsin and immobilized trypsin digestions

We then compared the efficiency of the trypsin bead digestion of serum proteins with a standard soluble trypsin protocol as described in Materials and Methods. Pooled healthy serum was incubated with MB-WCX paramagnetic beads to reduce sample complexity, and the eluate proteins were used in subsequent digestions using either immobilized-trypsin beads or soluble trypsin. These tryptic eluates were applied to C18,

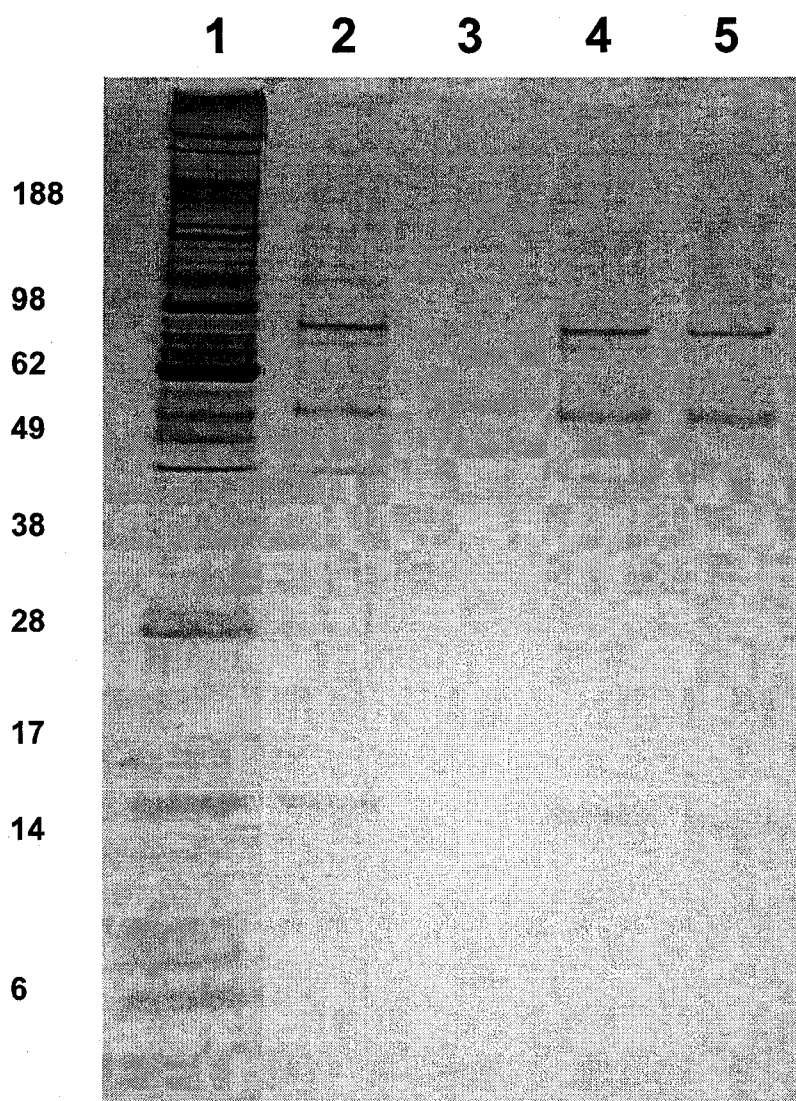


Figure 13. Effect of pH on ideal trypsinization using immobilized-trypsin beads. The gel shown is a NuPage 4-12% Bis-Tris with MES from Invitrogen. Lane 1: Undigested serum sample after WCX fractionation. Lane 2: 8 μg digested of serum sample after WCX fractionation and 20 μL ($\sim 5.5\mu\text{g}$) digested with immobilized-trypsin beads. Lane 3: Same as Lane 2 with the exception that after WCX fractionation the pH of the sample was adjusted to ~ 8 . Lane 4: Increasing the amount of trypsin beads (from 25 μL to 35 μL) to improve digestion efficiency. Lane 5: Digesting less reduced/alkylated sample ($\sim 2.75\ \mu\text{g}$) to improve digestion efficiency.

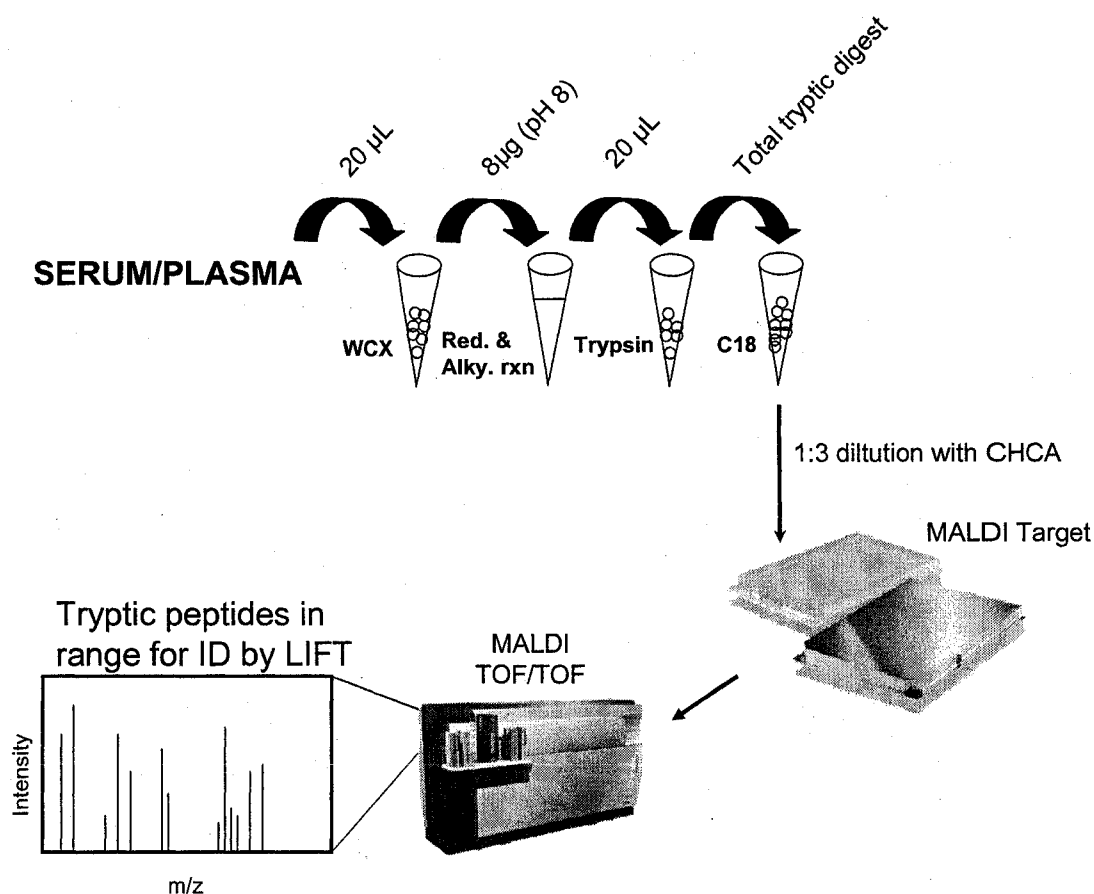


Figure 14. Workflow designed for immobilized-trypsin beads for peptide profiling using MALDI-TOF/TOF. Serum/Plasma samples are first fractionated by a chromatography-based magnetic bead (i.e. WCX or WAX). The pH of the sample is adjusted to ~ 8 and 8 μg of the fractionated sample is reduced and alkylated (final reaction volume is 29 μL). Twenty microliters of the reduced/alkylated sample is added to 25 μL of immobilized-trypsin beads and allowed to incubate at 37°C for 30 minutes. The total tryptic digest is then removed from the beads and added directly to a MB-C18 clean-up/concentration reaction. Two microliters of the C18 captured sample is then mixed with 4 μL CHCA matrix of which 1 μL is spotted on an AnchorChip plate and analyzed in reflectron mode of the MALDI-TOF/TOF Ultraflex III. Peaks of interest are subjected to MS/MS directly off of the profiled spot via the LIFT mode of the MALDI-TOF/TOF Ultraflex III. MS/MS spectra is analyzed and identified using the BioTools software from Bruker along with the MASCOT search engine.

and then spotted 1:3 with CHCA matrix for MALDI-TOF analysis. Representative spectra from the three tryptic digests (digested with trypsin beads for 30 minutes and digested with soluble trypsin for 30 min and overnight) along with the undigested MB-WCX eluate are shown in reflectron mode in Figure 15 and in linear mode in Figure 16. The trypsin beads were clearly more efficient for the conditions utilized and produced a greater number of lower mass peptides as compared to the standard soluble trypsin. However, the drawback with the trypsin-bound beads is that, as with most chromatography-based procedures, there is some extent of sample binding to the solid-support (it is estimated that $\sim 1/6^{\text{th}}$ of the sample binds to the solid-support). In terms of the soluble trypsin, it was noted that there were minimal differences between a 4 hour incubation time and an overnight incubation with the soluble trypsin. This is consistent with some trypsin digest protocols calling for the addition of more trypsin into the digestion reaction after the 4 hour mark or adding a higher starting concentration of trypsin to increase digestion efficiency (132, 133). As seen in Figure 17, there is already a certain level of contamination by trypsin peaks in the spectra and thus adding more trypsin to improve digestion efficiency would exacerbate this problem.

Reproducibility of immobilized trypsin protocol

The reproducibility of the MB-WCX and trypsin bead digest workflow described in Materials and Methods was applied to multiple aliquots of serum to determine the reproducibility of the technique. Six aliquots of the same serum samples were independently processed and spotted in triplicate for MALDI-TOF profiling. As shown in Figure 18, there was a high degree of reproducibility across the spectra, which is also

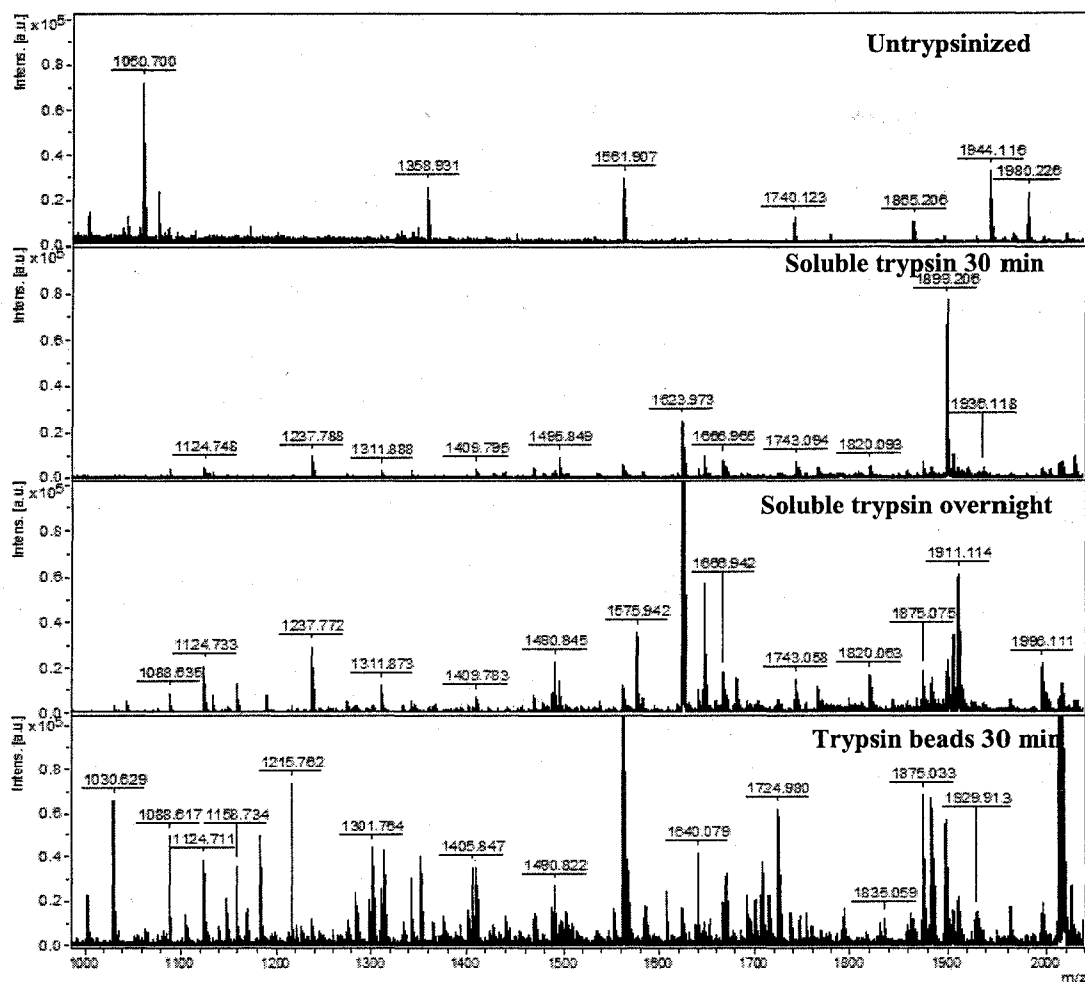


Figure 15. Reflectron mode MALDI-TOF comparison of WCX fractionated samples untrypsinized and trypsinized by either soluble trypsin or immobilized-trypsin. Samples were processed with 25 μ L of trypsin beads for 30 minutes or with soluble trypsin (1:40 sample-to-trypsin ratio) for 30 minutes or overnight as described in Materials and Methods. The tryptic peptides were captured by MB-C18 beads from Bruker and the eluted peptides were mixed 1:3 with CHCA matrix. One microliter of the sample/matrix mixture was spotted on an AnchorChip plate and analyzed using the reflectron mode of the MALDI-TOF UltraFlex III. The spectra were compared using the FlexAnalysis 2.0 software from Bruker.

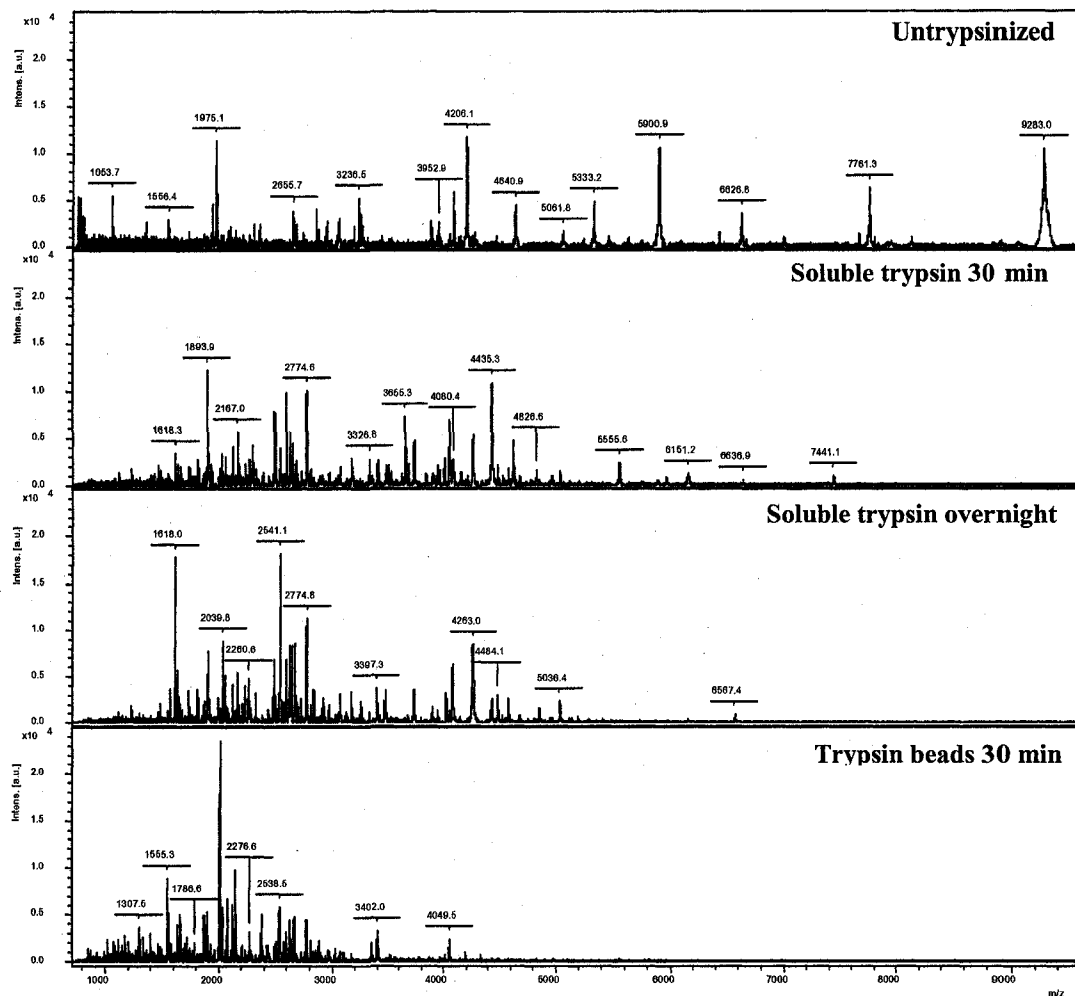


Figure 16. Linear mode MALDI-TOF comparison of WCX fractionated samples untrypsinized and trypsinized by either soluble trypsin or immobilized-trypsin. Samples were processed with 25 μ L of trypsin beads for 30 minutes or with soluble trypsin (1:40 sample-to-trypsin ratio) for 30 minutes or overnight as described in Materials and Methods. The tryptic peptides were captured by MB-C18 beads from Bruker and the eluted peptides were mixed 1:3 with CHCA matrix. One microliter of the sample/matrix mixture was spotted on an AnchorChip plate and analyzed using the linear mode of the MALDI-TOF UltraFlex III. The spectra were compared using the FlexAnalysis 2.0 software from Bruker.

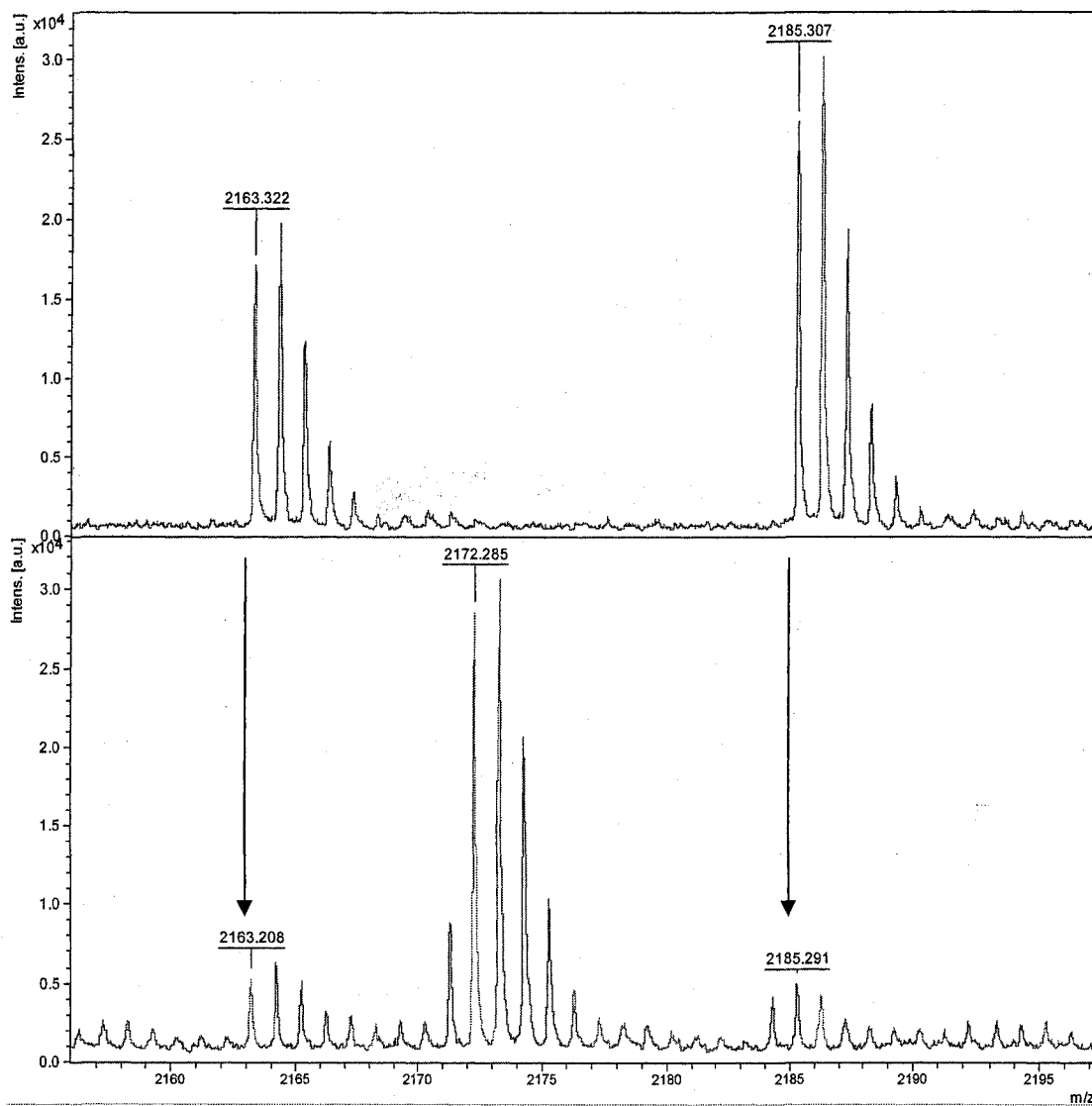


Figure 17. Example of trypsin contaminant peaks in soluble trypsin digests. The top panel shows representative peaks seen when the Roche trypsin is spotted on an AnchorChip plate with CHCA matrix. The bottom panel is a soluble, overnight trypsin digest of WCX fractionated serum. The arrows point to peaks found in the trypsin only spectra that are contaminating the soluble trypsin digest spectra. Samples were analyzed using the reflectron mode of the MALDI-TOF UltraFlex III and the resulting spectra were compared using the FlexAnalysis 2.0 software from Bruker.

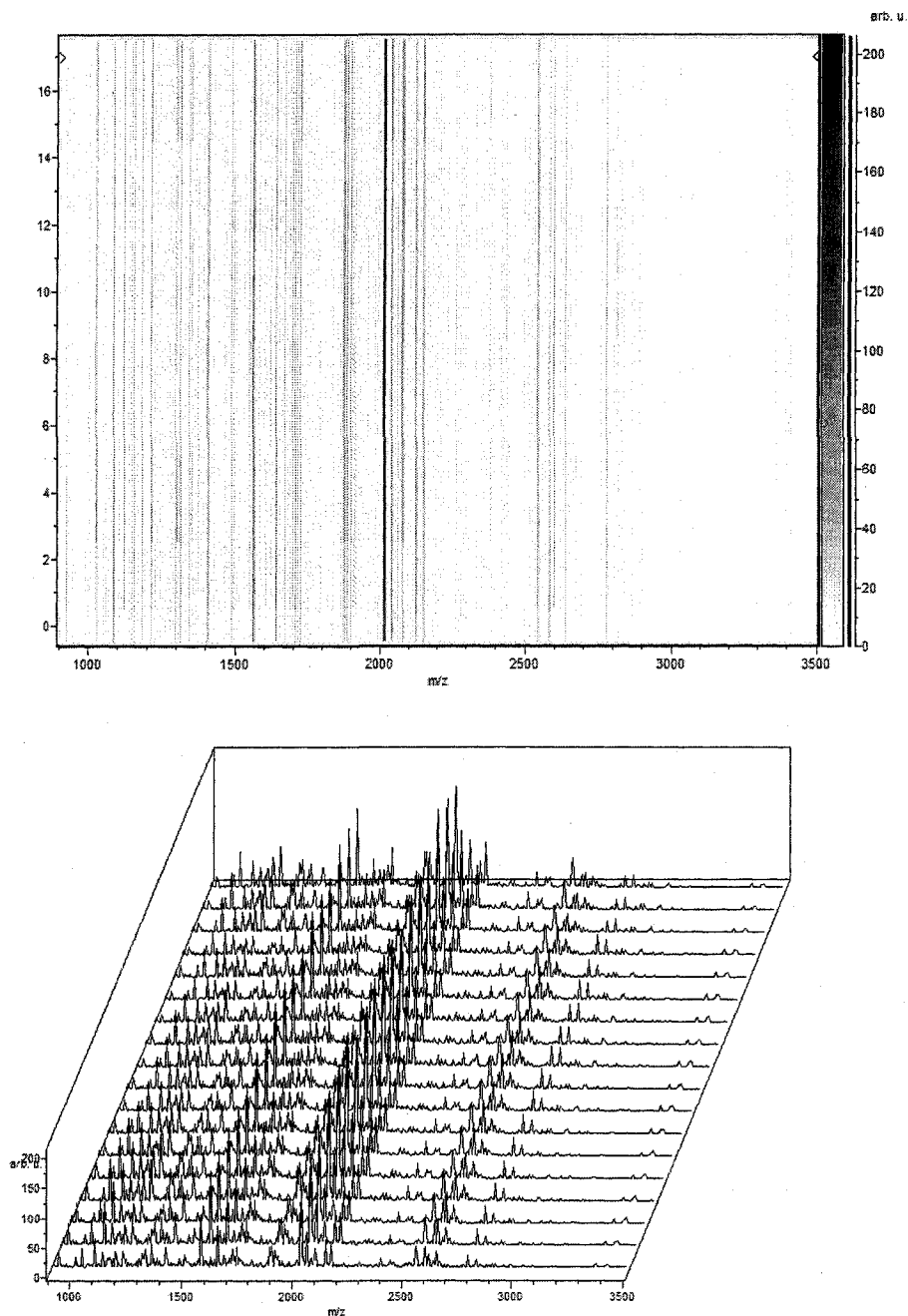


Figure 18. Reproducibility of immobilized trypsin bead method. Six aliquots of the same pooled serum sample were processed using the WCX fractionation and immobilized-trypsin digestion method as described in Figure 14 and Materials and Methods. Two microliters of each of the digested samples was mixed with 4 μ L CHCA matrix and 1 μ L of this mixture was spotted on an AnchorChip plate in triplicate. Spectra were generated in reflectron mode of the MALDI-TOF UltraFlex III and analyzed using ClinProTools 2.0. The top panel shows a heat map of all the samples in triplicate in the 900-3500 m/z range. The bottom panel shows the peak distribution of each individual sample in triplicate in the 900-3500 m/z range.

Table 8. Reproducibility of immobilized trypsin bead method as seen by the coefficient of variance (CV) of twelve representative peaks.

Mass	Manual run	
	Intensity	CV (%)
1124.81	5.33	8.41
1585.99	10.15	7.14
1667.97	8.32	4.96
1670.92	19.48	11.02
1694.93	6.75	13.65
1717.04	10.8	9.30
1885.02	39.18	8.88
1932.26	13.92	5.60
2017.31	154.16	11.34
2383.24	28.00	11.18
2425.80	46.22	11.49
2636.63	16.65	7.49

illustrated by the low coefficient of variance seen with twelve representative peak intensities (Table 8).

Bead-based Workflow with MALDI-TOF/TOF Identification of Tryptic Peptides

The WCX fractionation workflow was used in combination with trypsin beads for LIFT-MALDI-TOF/TOF sequencing identification of selected m/z peaks. Identified peptides are summarized in Table 9 and Figure 19 shows LIFT analysis for a representative peak. As expected, the protein identities represent common serum protein components from across all native mass ranges. Peaks with relatively low signal-to-noise ratio (S/N) values could be identified, in general, depending on whether any prevalent adjacent or co-migrating peaks were present to confound the LIFT fragmentation spectra. During this time we also examined two matrix formulations (a CHCA formulation in an EtOH/acetone solution and a CHCA formulation in an acetonitrile solution) and their effectiveness during the LIFT technique. Although both of these matrix formulations did not differ in spectra quality and LIFT results, it was found that the CHCA formulation in an acetonitrile solution was more resilient to the high laser energy needed for parent ion fragmentation and thus more shots per spot were able to be collected. However, this matrix would be difficult for large-scale spotting, either performed manually or robotically, due to the high content of acetonitrile, which is difficult to handle. Thus, the CHCA formulation used for the reproducibility study (EtOH/acetone based) was retained in the workflow for profiling and for direct LIFT application from the original spots used in profiling. Conversely, if sample quantities are limited and the LIFT procedure needs

Table 9. Peptides from immobilized-trypsin digest of WCX fractionated workflow identified by LIFT-MALDI-TOF/TOF.

Mass	Accession #	Peptide Identity	Score	Expect value	Peptide
1003.66	P01042	Kininogen-1	27	0.02	R.QVVAGLNFR.I
1030.63	P01042	Kininogen-1	44	0.0014	K.YFIDFVAR.E
1088.62	P00747	Plasminogen	32	0.029	R.WELCDIPR.C
1124.71	P04196	Histidine-rich glycoprotein	42	0.0011	R.DGYLFQLLR.I
1158.71	P01042	Kininogen-1	44	0.0012	K.KYFIDFVAR.E
1184.70	P02775	Platelet basic protein	49	0.00037	R.KICLDPDAPR.I
1215.76	P06727	Apolipoprotein A-IV	62	1.6e-05	K.ALVQQMEQLR.Q
1283.69	P02647	Apolipoprotein A-I	34	0.014	K.WQEEMELYR.Q
1301.77	P02647	Apolipoprotein A-I	33	0.015	R.THLAPYSDEL.R.Q
1314.81	P04004	Vitronectin	50	0.00023	R.RVDTVDPYPR.S
1342.77	P02768	Serum albumin	67	6.2e-06	K.AVMDDFAAFVEK.C
1352.79	P06727	Apolipoprotein A-IV	32	0.015	R.RVEPYGENFNK.A
1561.87	P00734	Prothrombin	68	4.8e-06	R.TATSEYQTFNPR.T
1623.83	P02768	Serum albumin	46	0.00084	K.DVFLGMFLYEYAR.R
1640.09	P02768	Serum albumin	66	6.9e-07	K.KVPQVSTPTLVEVSR.N
1724.99	P02775	Platelet basic protein	105	7.5e-10	K.GKEESLSDLYAELR.C
1875.05	P01042	Kininogen-1	76	6.8e-07	K.YNSQNQSNNQFVLYR.I
1884.06	P00734	Prothrombin	83	1.2e-07	R.TFGSGEADCGLRPLFEK.K
1898.12	P02775	Platelet basic protein	99	2.4e-09	K.GTHCNQVEVIATLKDGR.K
2012.17	P00734	Prothrombin	63	9.8e-06	R.TFGSGEADCGLRPLFEK.K.S
2016.24	P02765	Alpha-2-HS-glycoprotein	115	2.5e-11	R.TVVQPSVGAAAGPVVPPCPGR.I
2042.09	P02671	Fibrinogen alpha chain	92	1.7e-08	K.QFTSSTSYNRGDSTFESK.S
2081.20	P02765	Alpha-2-HS-glycoprotein	83	1e-07	R.HTFMGVVSLGSPSGEVSHPR.K
2126.16	P00747	Plasminogen	74	8.6e-07	R.ATTVTGTPCQDWAAQEPHR.H
2381.39	O00512	B-cell lymphoma 9 protein	34	0.005	K.KPEGPIQAMMAQSQSLGKGGP R.T + Oxidation (M)
2441.32	P01042	Kininogen-1	93	1.1e-08	K.SLWNGDTGECTDNA YIDIQLR.I
2585.33	P02768	Serum albumin	75	6.1e-07	K.VHTECCHGDLLECADDRADLAK .Y
2599.38	P02768	Serum albumin	72	1.2e-06	K.QNCELFEQLGEYKFQNALLVR.Y
2778.59	P02768	Serum albumin	32	0.0085	R.LVRPEVDVMCTAFHDNEETFLK K.Y
2815.52	P00751	Complement factor B	71	1.1e-06	R.LLQEGQALEYVCPSPGFYPVQVT R.T

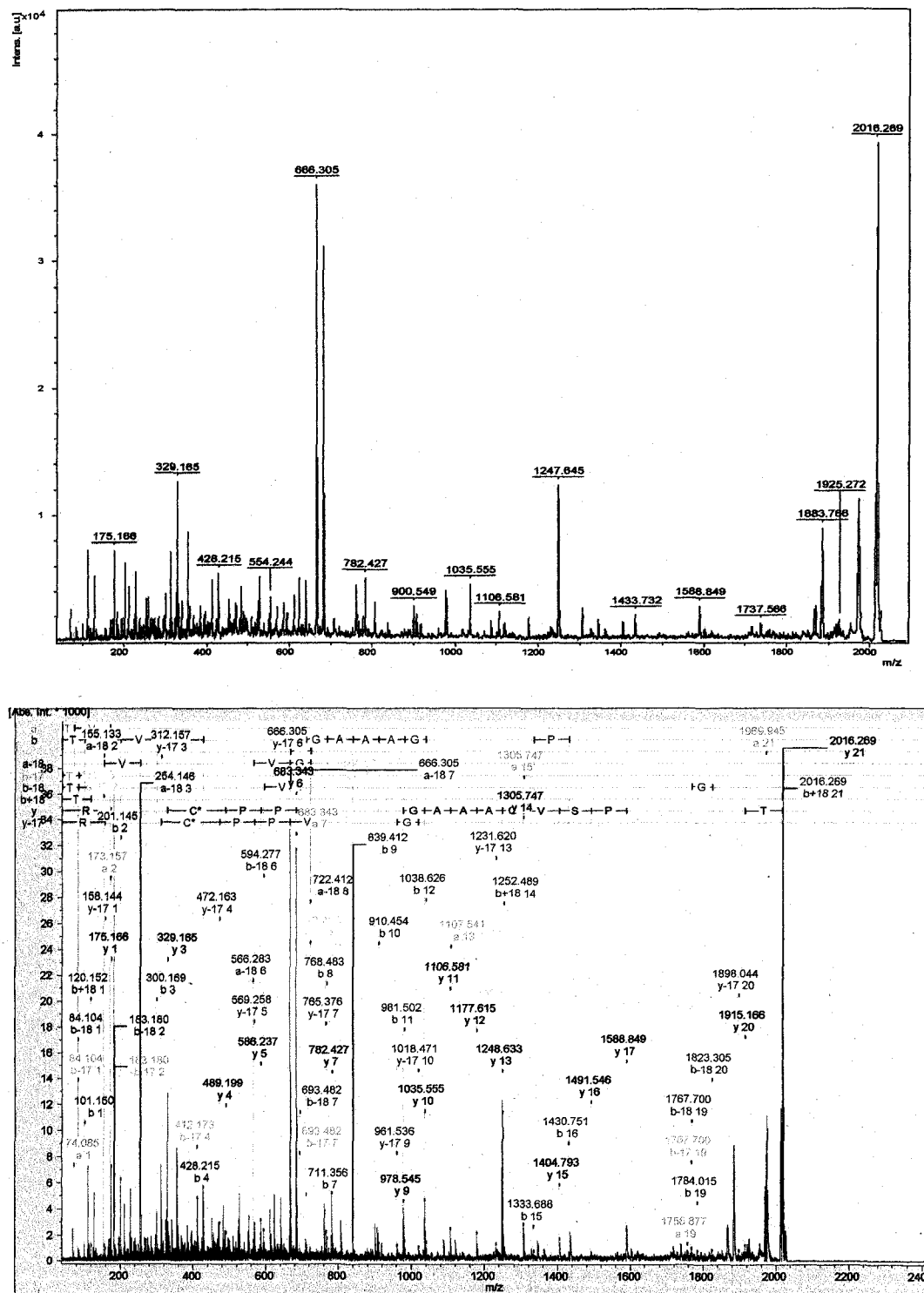


Figure 19. LIFT analysis of representative peak m/z 1616.24. The top panel shows the fragmented spectrum as is it is seen using FlexAnalysis. This MS/MS spectrum is then imported into the BioTools software program and MASCOT is used to search a selected database (SwissProt in this case) with the following criteria: carbamidomethyl and oxidation modifications; 100 ppm mass tolerance MS; 0.5 Da MS/MS tolerance.

to be repeated or performed on re-spotted sample, we recommend using the acetonitrile-based CHCA matrix.

The front-end bead fractionation may be varied for a more comprehensive sample analysis. Another front-end fractionation step easily adapted to this trypsinization technique utilizes MB-WAX beads. Examples of the types of tryptic peptides generated from MB-WAX fractionated samples are listed in Table 10. There are some shared commonalities (i.e. apolipoprotein AIV, prothrombin, vitronectin and alpha-2HS-glycoprotein), and there are also several differences between these WAX front-end fractionated tryptic peptides identified by LIFT-MALDI-TOF/TOF and those that were generated from WCX fractionated serum. For instance, the WCX fractionation workflow has such proteins as kininogen, histidine-rich glycoprotein, platelet-basic protein and B-cell lymphoma 9 protein that among others are not seen in the WAX fractionation workflow spectra. Vice versa, haptoglobin, ceruloplasmin, apolipoprotein CIII, and complement C4-A are amongst the proteins represented in the WAX workflow protocol that are not seen in the WCX workflow spectra. Additionally, the predominant protein in the WCX workflow is serum albumin, while the predominant protein in the WAX fractionated sample is inter-alpha-trypsin inhibitor. This is one reason why the pH 5.0 WAX binding buffer was utilized as specified in Materials and Methods. The average serum albumin pI is ~5.2 (main pI isoforms range from 4.7 to 5.6) (134), thus at the pH of the binding step the negatively charged (at neutral pH) serum albumin takes on an overall neutral charge and thus does not bind to the WAX bead. However, the other pH binding solutions (pH 7.4 and pH 9) would have made serum albumin negatively charged

Table 10. Examples of top peptides seen in the immobilized-trypsin digestion of WAX fractionated serum.

Mass	Accession #	Peptide Identity	Score	Expect value	Peptide
980.61	P00738	Haptoglobin	34	0.01	R.VGYVSGWGR.N
1182.69	Q9HB96	Fanconi anemia group E protein	37	0.0062	R.EEPVVQGPDGR.L
1215.79	P06727	Apolipoprotein A-IV	35	0.0064	K.ALVQQMEQLR.Q
1337.84	P19823	Inter-alpha-trypsin inhibitor heavy chain H2	44	0.00064	K.FYNQVSTPLLR.N
1401.82	P02751	Fibronectin	32	0.021	K.HYQINQWER.T
1470.62	P19827	Inter-alpha-trypsin inhibitor heavy chain H1	80	2.3e-07	K.QYYEGSEIVVAGR.I
1561.88	P00734	Prothrombin	79	3.5e-07	R.TATSEYQTFNPR.T
1647.00	P04004	Vitronectin	70	2.2e-06	R.DVWGIEGPIDAAFT R.I
1666.62	P04004	Vitronectin	69	9.1e-07	R.DWHGVPGQVDAA MAGR.I
1717.04	P02656	Apolipoprotein C-III	62	1.3e-05	K.DALSSVQESQVAQQ AR.G
1836.12	P01876	Ig alpha-1 chain C region	42	0.00071	R.QEPSQGTTFVAVTSI LR.V
2017.25	P02656	Apolipoprotein C-III	73	5e-07	K.TAKDALSSVQESQV AQQAR.G
2081.22	P02765	Alpha-2-HS-glycoprotein	102	1.1e-09	R.HTFMGVVSLGSPSG EVSHPR.K
2293.26	P00450	Ceruloplasmin	43	0.0011	R.FNKNNEGTYSPNY NPQSR.S
2551.46	P0C0L4	Complement C4-A	102	1.2e-09	R.TLEIPGNSDPNMIPD GDFNSYVR.V
2755.57	P01024	Complement C3	114	6.6e-11	R.EGVQKEDIPPADLS DQVPDTESETR.I
2994.16	P19827	Inter-alpha-trypsin inhibitor heavy chain H1	67	1.3e-06	R.GMADQDGLKPTIDK PSEDSPPLEMLGPR.R

and thus most amenable to binding to the WAX bead type. In addition to this observation, we also saw that compared to the other binding solution, pH 5.0 yielded the highest protein concentration, thus allowing the fractionated sample to be applied easily to the immobilized-trypsin workflow.

Additionally, we attempted a LIFT of an untrypsinized sample of both WCX fractionated samples and WAX fractionated samples. Table 11 shows the identified peptides from both the WCX and the WAX untrypsinized samples. As discussed above for the WCX fractionated samples, the spectra of untrypsinized samples are sparse as compared to the spectra generated immobilized trypsin samples. This is also true for WAX fractionated samples (on average there are 55 peaks for trypsinized samples versus 18 peaks for untrypsinized samples). Additionally, using WCX as an example, the peaks that qualify as good candidates for LIFT (i.e. good S/N ratio and not in the midst of a peak patch) are on average ~ 45 for the trypsinized samples as compared to ~12 for the untrypsinized samples. Furthermore, the lack of knowledge as to the identity of the enzyme that created the peptide of interest adds to the difficulty of identifying the endogenous peptide peaks in the untrypsinized samples. Of the 45 peaks that were subjected to LIFT-MS/MS from the serum samples processed with the WCX/immobilized-trypsin bead workflow, 30 peaks were successfully identified. However, of the 12 peaks that were subjected to LIFT-MS/MS from the untrypsinized WCX fractionated serum samples, only 5 peaks were successfully identified. This is true for the WAX workflow as well i.e. LIFT-MS/MS performed on 25 peaks from the trypsinized sample with 17 identified, while only 4 peaks were identified from the 10 peaks fragmented by LIFT-MS/MS from the untrypsinized sample. Thus, it is easier to

Table 11. Peptides identified from untrypsinized WCX and WAX fractionated serum.[†]

WCX fractionated serum					
Mass	Accession #	Peptide Identity	Expect value	Peptide	Enzyme denotation
1060.66	P01042	Kininogen-1	2.8e-05	K.RPPGFSPFR.S	Trypsin*
1192.66	P01042	Kininogen-1	0.014	R.RHDWGHEKQ.R	No enzyme
1263.80	Q14624	Inter-alpha-trypsin inhibitor H4	0.0059	R.MNFRPGVLSSR.L	Trypsin*
1561.82	P00734	Prothrombin	9.1e-06	R.TATSEYQTFNPR.T	Trypsin
1753.98	P02649	Apolipoprotein E	1.8e-05	A.KVEQAVETEPEPEL R.Q	No enzyme
WAX fractionated serum					
Mass	Accession #	Peptide Identity	Expect value	Peptide	Enzyme denotation
1206.72	P02671	Fibrinogen alpha	0.0015	G.EGDFLAEGGGVR.G	Semi-trypsin
1465.82	P02671	Fibrinogen alpha	1.1e-10	A.DSGEGDFLAEGGGV R.G	Semi-trypsin
1616.84	P02671	Fibrinogen alpha	2.4e-08	T.ADSGEGDFLAEGGG VR.G + Phospho (ST)	Semi-trypsin
2193.34	P01024	Complement C3	0.012	G.SPMYSIITPNILRLES EET.M	No enzyme

[†] SwissProt database searched with the following criteria: oxidation modifications; 120 ppm mass tolerance MS; 0.5 Da MS/MS tolerance. * Indicates that without enzyme selection the peptide identification would not be significant.

identify a peptide from the immobilized-trypsin workflow, (where we know that the enzyme that created that peptide is trypsin), than from the untrypsinized sample.

Trypsin-bead workflow on clinical samples: SOF study revisited

We next performed proof-of-concept trypsin bead workflows on clinical samples as described in Figure 14. The first set of samples utilized consisted of the scraped SOF serum samples from Aim 1. We processed these samples with the WCX fractionation and immobilized-trypsin digestion workflow as discussed in Materials and Methods in order to examine the reproducibility of this workflow with clinical samples and also to investigate whether there are any differential peptides that can be identified between the two sample sets (cases and controls). We made 24 pools (12 pools per group) with 8 samples per pool. Two peaks that were deemed most significant using the ClinProTools software, m/z 2017 and m/z 2383 were included in the generation of a genetic algorithm model, along with another significantly differential peak, m/z 1030 (it should be noted that the values we list in the text are average of isotopic values generated by ClinProTools, but the values used for the LIFT identification are of the 1st isotopic peak). Figure 20 shows the cluster plot of peaks m/z 2017 and m/z 2383. These peaks were identified as alpha-2HS-glycoprotein (m/z 2017), B-cell lymphoma 9 protein (Bcl-9) (m/z 2383), and kininogen (m/z 1031), which were all increased in patients that were going to develop breast cancer (cases). The genetic algorithm model had 100% recognition capability of the test set and the cross-validation yielded a sensitivity of 87.23% and a specificity of 77.36%.

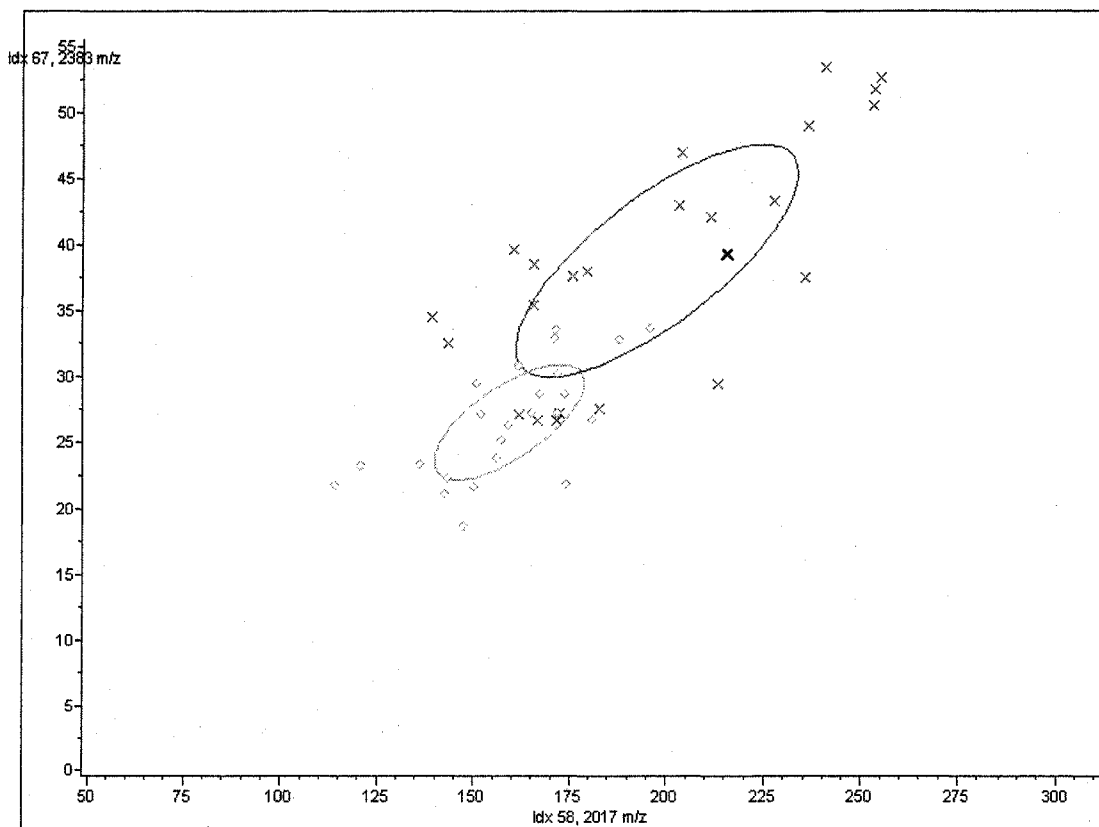


Figure 20. Cluster plot of relative intensity distributions of peaks m/z 2017 and 2383 in the initial run of the SOF cohort. The cluster plot was generated by the ClinProTools 2.0 software. The intensities of the 2 peaks, m/z 2017 and 2383, are plotted on 2 axes. The more clustered the points are in relation to their group and the more separated the clusters are from each other then the more significantly distinct the 2 groups are in relation to each other. In this cluster plot the intensities of peak m/z 2017 are found along the “x” axis, while the intensities of peak m/z 2383 along the “y” axis. The control sample peaks are designated by “o” and the case sample peaks are designated by “x”.

We re-processed the 24 samples and analyzed them again in duplicate on the MALDI-TOF/TOF instrument. A genetic algorithm model was generated using the same peaks (m/z 1031, 2017, and 2383) as for the previous run and was found to have a recognition capability of 100% of the test samples, with a cross-validation yielding a sensitivity of 70.83% and a specificity of 70.83%. A cluster plot using the peaks m/z 2017 and m/z 2383 is shown in Figure 21. We then used this newly generated model to see if it could properly externally validate the samples analyzed in the original run. The genetic algorithm model that was designed specifically for the repeated sample set was able to validate the 1st group of samples with a sensitivity of 70.83% (17/24 correctly classified) and a specificity of 83.3% (20/24 correctly classified). The combined peak intensity distributions for these SOF runs and the respective p-values are shown for kininogen (Figure 22), alpha-2HS-glycoprotein (Figure 23) and Bcl-9 protein (Figure 24). Interestingly, an independent iTRAQ analysis that was performed in our lab using lectin capture and different SOF sample pools showed that there was a 2.33-fold increase of alpha-2HS-glycoprotein in cases as compared to controls (1.0 to 0.43 iTRAQ ratio of case to control).

The results shown here are very promising in terms of the reproducibility of this method over time since there was a lengthy lapse between the analyses of these two SOF runs on the MALDI-TOF/TOF, during which time the instrument had been serviced by Bruker engineers. Yet the intensity and distribution patterns of the peaks used to construct our genetic algorithm models are still comparable enough to produce similar cross-validations between the two runs and also to allow for a good external validation of the first run by the genetic algorithm criteria set forth by the repeat run.

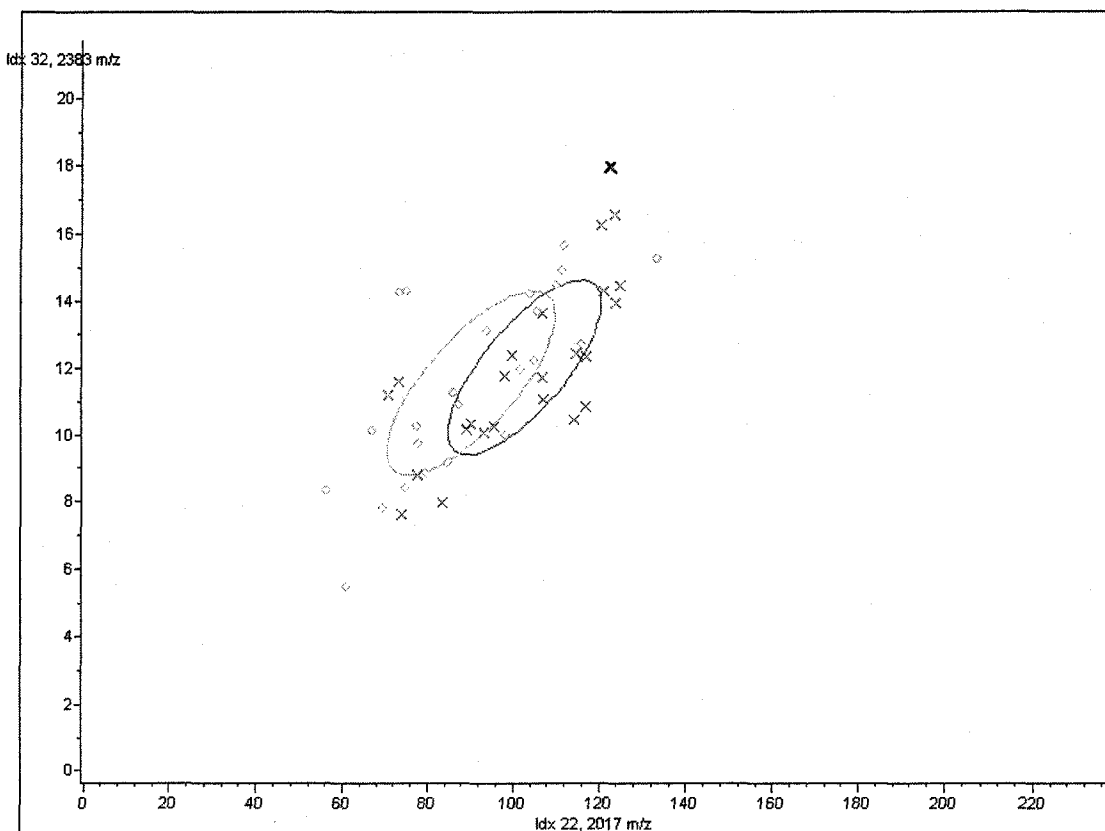


Figure 21. Cluster plot of relative intensity distributions of peaks 2017 and 2383 in the repeat run of the SOF cohort. The cluster plot was generated by the ClinProTools 2.0 software. The intensities of the 2 peaks, m/z 2017 and 2383, are plotted on 2 axes. The more clustered the points are in relation to their group and the more separated the clusters are from each other then the more significantly distinct the 2 groups are in relation to each other. In this cluster plot the intensities of peak m/z 2017 are found along the “x” axis, while the intensities of peak m/z 2383 along the “y” axis. The control sample peaks are designated by “o” and the case sample peaks are designated by “x”.

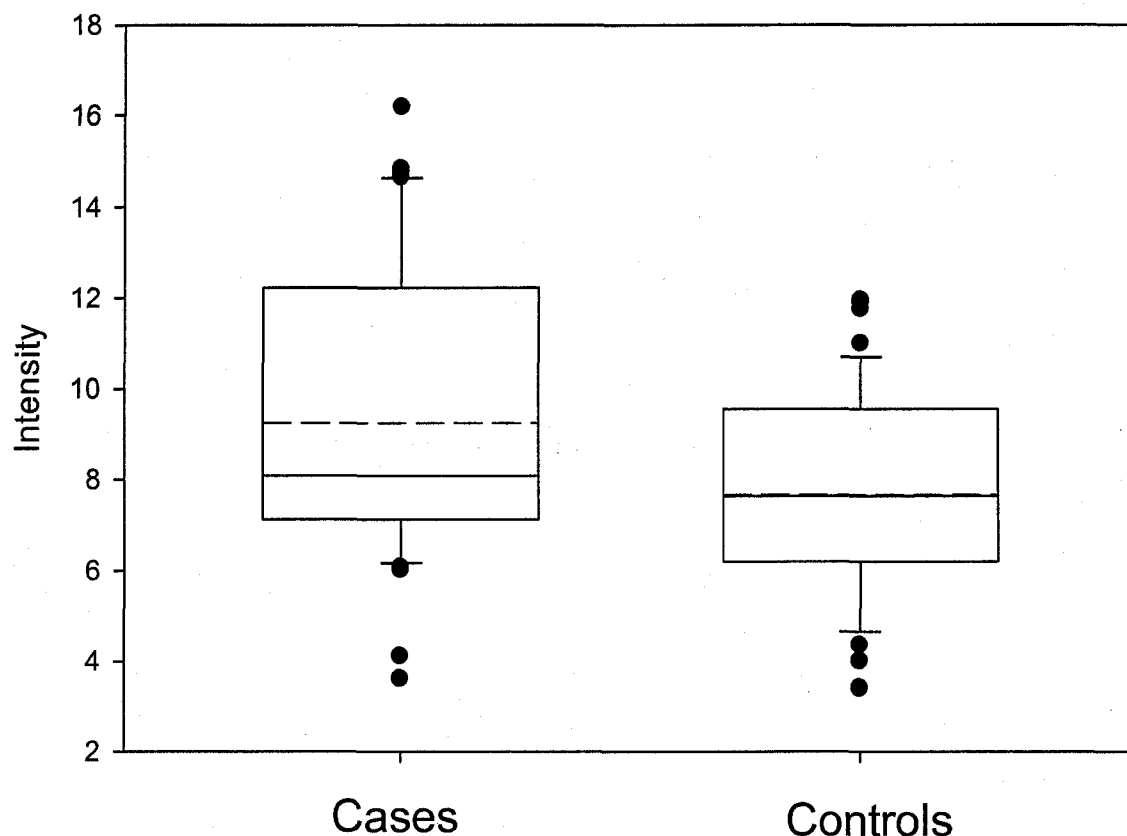


Figure 22. Box plot of relative intensity distributions of peak m/z 1031 in the SOF cohort. Relative intensities were plotted from both runs. The two runs were processed by the ClinProTools software independently with a 0.1% maximum peak shift tolerance and a beginning m/z cut-off of 800. The box denotes where the intensities of the majority of samples lie and the whiskers of the box plot demonstrate range of intensities of all samples that are not deemed outliers. All outliers are shown as individual points. The mean is depicted as a dotted line and the median is depicted as a solid line. For m/z 1031, the mean was 9.25 for cases and 7.67 for controls, while the median was 8.08 for cases and 7.63 for controls. Using a student t-test, the P-value for this peak was determined to be 0.0064. Significance is < 0.05 .

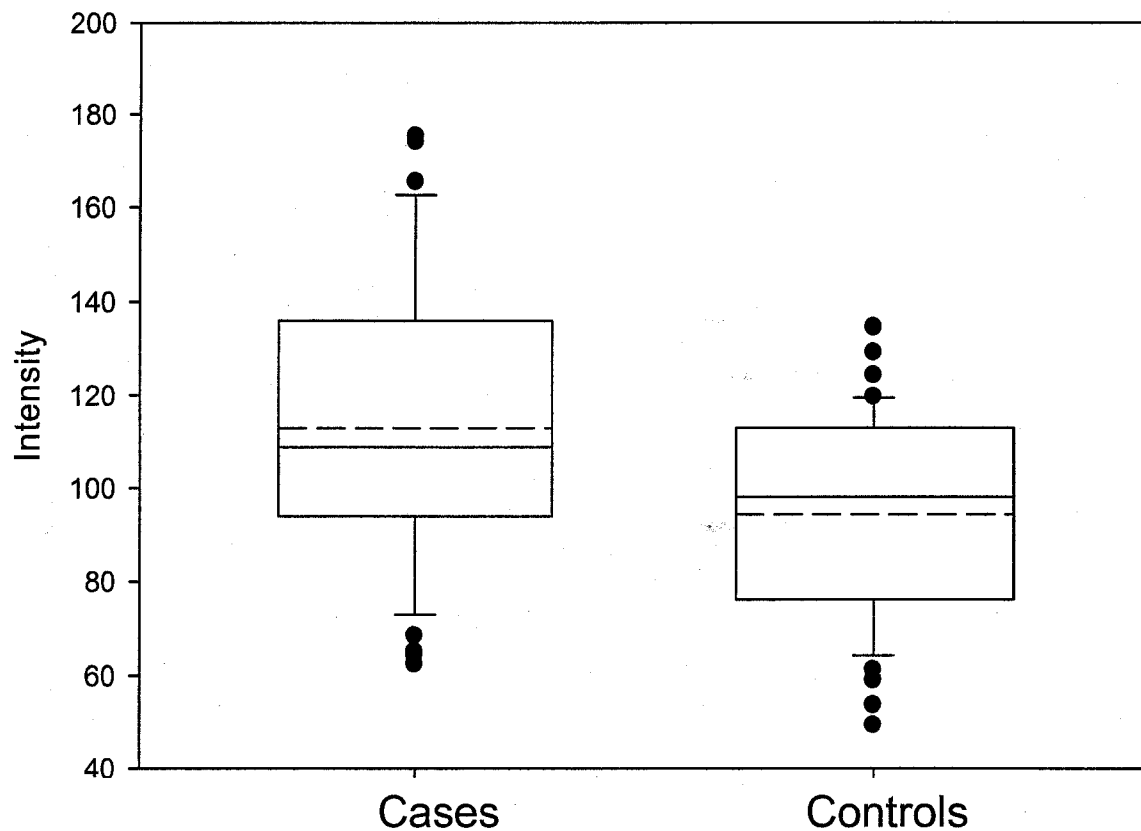


Figure 23. Box plot of relative intensity distributions of peak 2017 in the SOF cohort. Relative intensities were plotted from both runs. The two runs were processed by the ClinProTools software independently with a 0.1% maximum peak shift tolerance and a beginning m/z cut-off of 800. The box denotes where the intensities of the majority of samples lie and the whiskers of the box plot demonstrate range of intensities of all samples that are not deemed outliers. All outliers are shown as individual points. The mean is depicted as a dotted line and the median is depicted as a solid line. For m/z 2017, the mean was 112.88 for cases and 94.36 for controls, while the median was 108.81 for cases and 98.14 for controls. Using a student t-test, the P-value for this peak was determined to be 0.00095. Significance is < 0.05 .

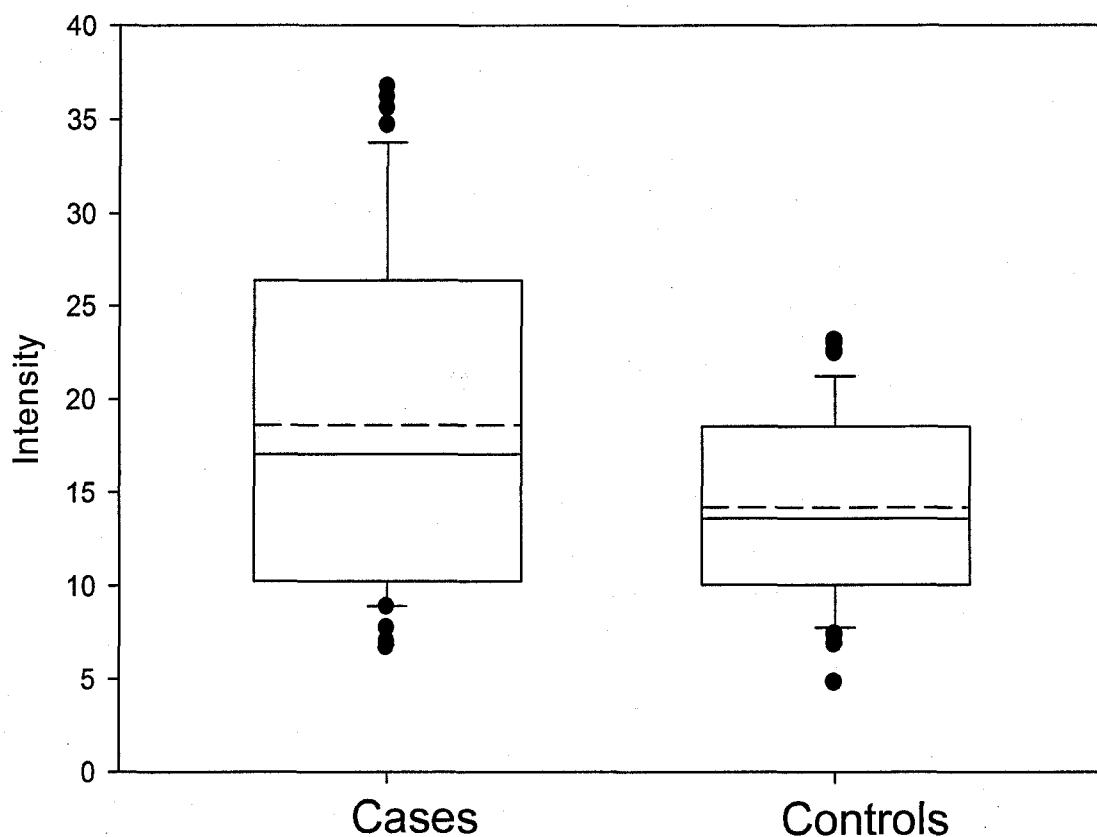


Figure 24. Box plot of relative intensity distributions of peak 2383 in the SOF cohort. Relative intensities were plotted from both runs. The two runs were processed by the ClinProTools software independently with a 0.1% maximum peak shift tolerance and a beginning m/z cut-off of 800. The box denotes where the intensities of the majority of samples lie and the whiskers of the box plot demonstrate range of intensities of all samples that are not deemed outliers. All outliers are shown as individual points. The mean is depicted as a dotted line and the median is depicted as a solid line. For m/z 2383, the mean was 18.59 for cases and 14.19 for controls, while the median was 17.03 for cases and 13.57 for controls. Using a student t-test, the P-value for this peak was determined to be 0.0046. Significance is < 0.05 .

Trypsin-bead workflow on clinical samples: Serum from patients diagnosed with Benign Prostatic Hyperplasia and Prostate Cancer

In addition to the SOF sample cohort we also processed a set of serum samples through the workflow described in Figure 14 from individuals diagnosed with BPH and those diagnosed with prostate cancer (PCa) (all the patients had PSA levels between 2 and 10 ng/mL). Twenty pooled samples (10 pooled samples per group) were run in duplicate as described in the Materials and Methods. We found that two peptides identified as Apolipoprotein AIV (ApoAIV) (m/z 1216 and 1353) were increased in patients that were diagnosed with BPH. In addition, a kininogen peptide (m/z 1031) was also noticeably increased in patients that had PCa. ClinProTools was used to generate a genetic algorithm model using these three peaks and this model was able to differentiate 100% between the two test groups. Upon cross-validation this model was able to distinguish between sample groups with a sensitivity of 84.11% (correctly classified cases i.e. PCa) and specificity of 75% (correctly classified controls i.e. BPH).

To determine the reproducibility of this method using clinical samples we randomly re-processed 12 of the samples (6 samples per group) with the front-end MB-WCX fractionation and trypsinization scheme. These samples were blinded and analyzed in duplicate by the MALDI-TOF using the exact specification and setting used for the initial run. We utilized the model generated from the original 20 samples to classify the blinded samples. In this manner we were able to classify 10/12 correctly as BPH and 10/12 correctly as PCa, giving us a specificity and sensitivity of 83.3%. A cluster plot was created using the ApoAIV m/z 1216 peptide and the kininogen peptide intensity distributions from both runs (Figure 25). Figure 26 is a box plot of the combined

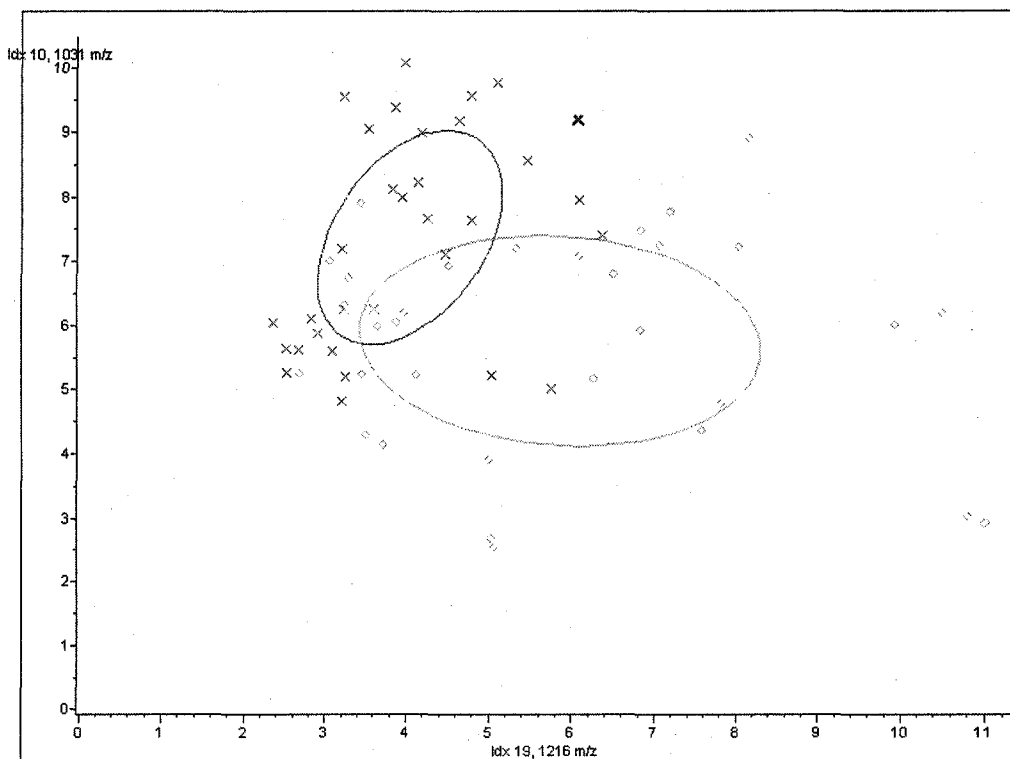


Figure 25. Cluster plot of relative intensity distributions of peaks 1031 and 1216 in the PCa vs. BPH cohort. The cluster plot was generated by the ClinProTools 2.0 software. The intensities of the 2 peaks, m/z 1031 and 1216, are plotted on 2 axes. The more clustered the points are in relation to their group and the more separated the clusters are from each other then the more significantly distinct the 2 groups are in relation to each other. In this cluster plot the intensities of peak m/z 1216 are found along the “x” axis, while the intensities of peak m/z 1031 along the “y” axis. The control sample peaks are designated by “o” and the case sample peaks are designated by “x”.

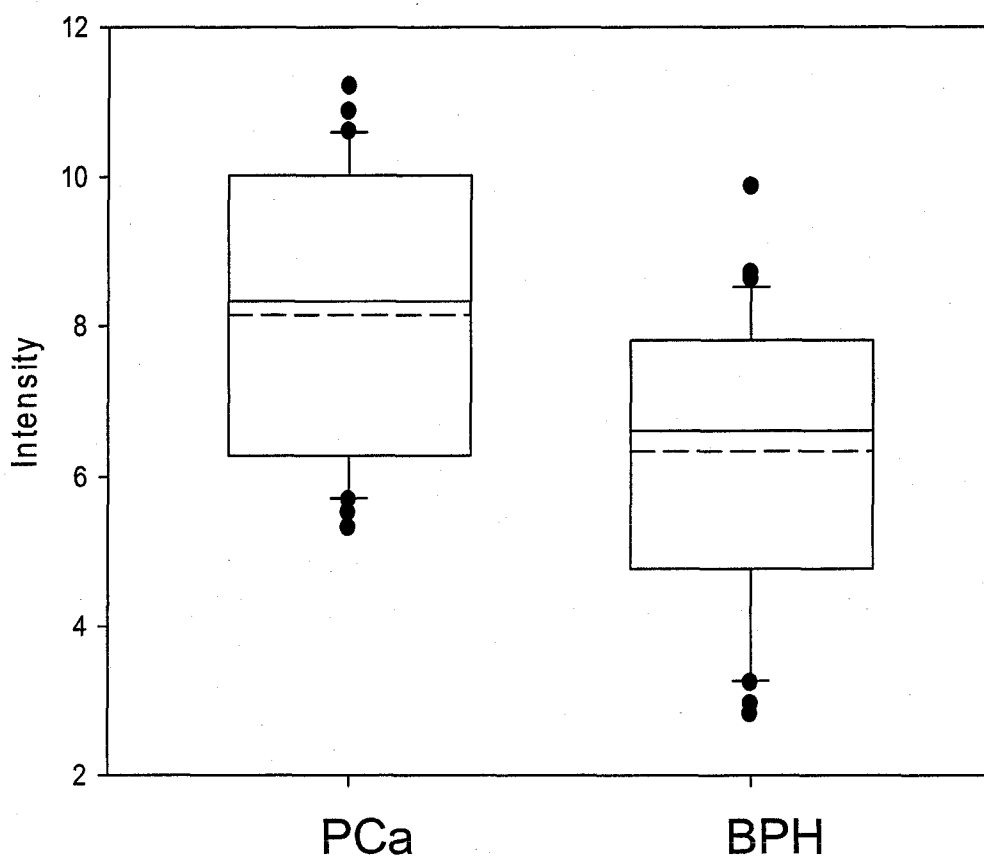


Figure 26. Box plot of relative intensity distributions of peak 1031 in the PCa vs. BPH cohort after WCX fractionation. Relative intensities were plotted from both runs. The two runs were processed by the ClinProTools software independently with a 0.1% maximum peak shift tolerance and a beginning m/z cut-off of 800. The box denotes where the intensities of the majority of samples lie and the whiskers of the box plot demonstrate range of intensities of all samples that are not deemed outliers. All outliers are shown as individual points. The mean is depicted as a dotted line and the median is depicted as a solid line. For m/z 1031, the mean was 8.14 for PCa and 6.34 for BPH, while the median was 8.32 for PCa and 6.61 for BPH. Using a student t-test, the P-value for this peak was determined to be 0.00024. Significance is < 0.05 .

intensity distributions of kininogen and Figure 27 is the box plot of the combined intensity distributions of the two ApoAIV peaks between BPH and PCa using the WCX fractionation/immobilized-trypsin bead workflow

Additionally, we knew that an ApoAIV peak is also found in the WAX fractionation workflow, we thus processed the 20 samples by that workflow and found that the peptide for ApoAIV (m/z 1216) was also increased in the BPH sample set using this fractionation scheme. Figure 28 shows the intensity distribution of the ApoAIV peak between BPH and PCa in the WAX scheme.

We also compared the raw values of peak intensities between the first group of WCX processed and immobilized-trypsin bead digested samples and their repeated counterparts. These workflows are reproducible with clinical samples over-time when specific instrumental settings are utilized, since the intensities did not vary much between the same samples which were processed and analyzed weeks apart. For example, two BPH samples had an intensity (duplicates averaged) of 11.48 and 7.36, respectively at m/z 1216. The same samples re-processed and re-analyzed weeks later, had an intensity of 12.29 and 7.47, respectively at m/z 1216. On the PCa side, two PCa samples had peak intensities of 4.52 and 3.34 at m/z 1216. These same PCa samples were re-processed and re-analyzed weeks later and had intensities of 4.57 and 3.75 at m/z 1216.

4.4 Discussion

The development of new instrument configurations and continued improvement in existing proteomic mass spectrometry technologies has allowed for unprecedented

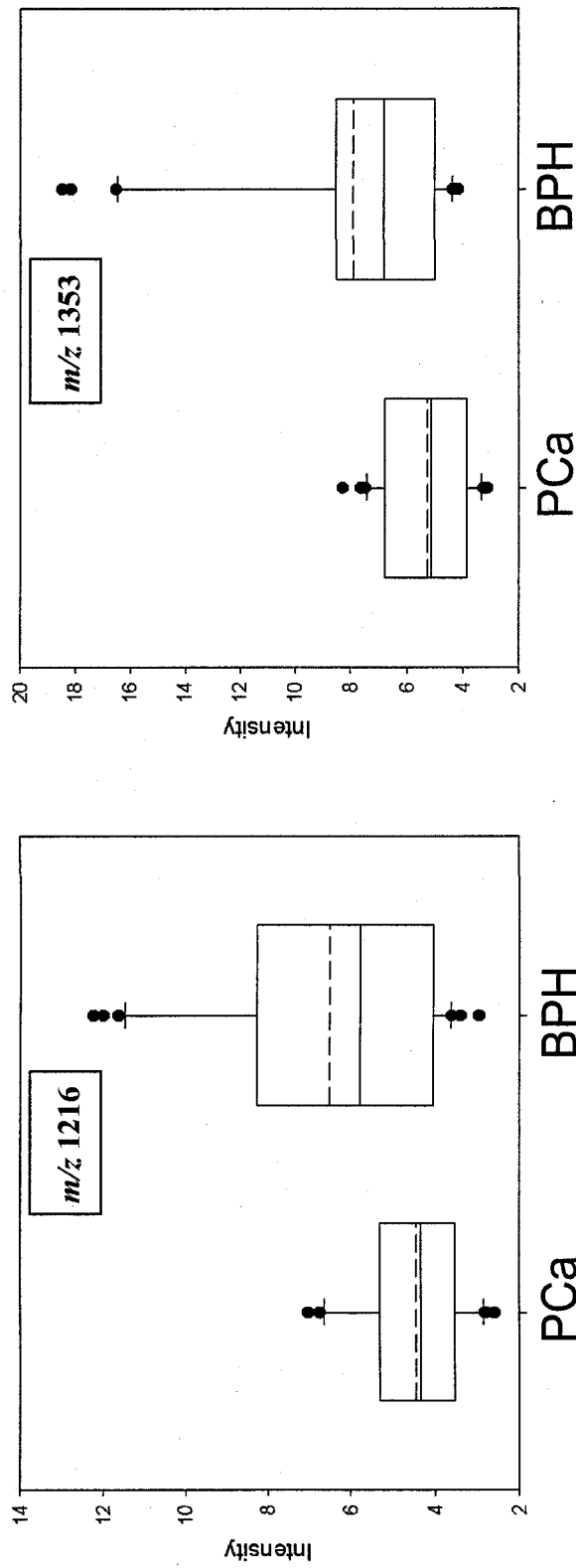


Figure 27. Box plot of relative intensity distributions of the ApoAIV peaks in the PCa vs. BPH cohort after WCX fractionation. Relative intensities were plotted from both runs. The two runs were processed by the ClinProTools software independently with a 0.1% maximum peak shift tolerance and a beginning m/z cut-off of 800. The box denotes where the intensities of the majority of samples lie and the whiskers of the box plot demonstrate range of intensities of all samples that are not deemed outliers. All outliers are shown as individual points. The mean is depicted as a dotted line and the median is depicted as a solid line. For m/z 1216, the mean was 4.47 for PCa and 6.52 for BPH, while the median was 4.35 for PCa and 5.78 for BPH. Using a student t-test, the P-value for this peak was determined to be 0.00025. For m/z 1353, the mean was 5.29 for PCa and 7.9 for BPH, while the median was 5.15 for PCa and 6.83 for BPH. Using a student t-test, the P-value for this peak was determined to be 0.00096. Significance is < 0.05 .

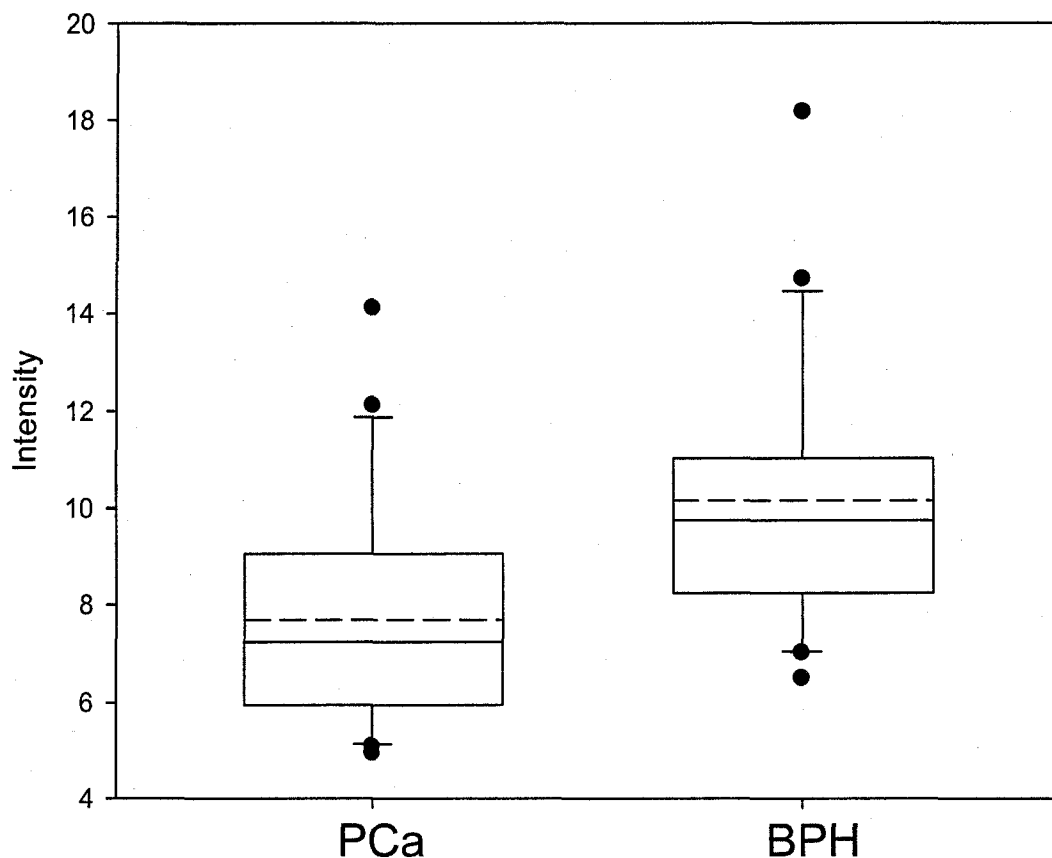


Figure 28. Box plot of relative intensity distributions of peak 1216 in the PCa vs. BPH cohort after WAX fractionation. Relative intensities were plotted after the spectra was processed by the ClinProTools software with a 0.1% maximum peak shift tolerance and a beginning m/z cut-off of 800. The box denotes where the intensities of the majority of samples lie and the whiskers of the box plot demonstrate range of intensities of all samples that are not deemed outliers. All outliers are shown as individual points. The mean is depicted as a dotted line and the median is depicted as a solid line. For m/z 1216, the mean was 7.69 for PCa and 10.15 for BPH, while the median was 7.24 for PCa and 9.75 for BPH. Using a student t-test, the P-value for this peak was determined to be 0.0042. Significance is < 0.05 .

opportunities in biomarker protein discovery and analysis of complex biological proteomes. In particular, analysis of serum and plasma will always be a challenge both because of the inherent dynamic characteristics of the fluid and the persistent clinical variables that affect the quality of the starting material. Any proteomic study using clinically obtained samples will be influenced by issues involving sample collection, processing and storage, the level of epidemiological input and study design biases. The different sample processing workflows described in this Aim demonstrate the feasibility of sequential chromatographic fractionation and trypsinization protocols to elucidate proteomic differences in complex samples like serum and the peptides generated are ideal for direct MALDI-TOF/TOF sequence determinations.

The technique described in this method paper is complementary to the mining of the low molecular weight (LMW) proteome in that it allows for the examination of peptides that may have been previously outside the m/z range of the mass spectrometer. In addition, the high-efficiency of the immobilized trypsin may allow for the release of less abundant proteins from carrier molecules or protein complexes. These newly created, low abundant peptides may be enriched by using various up-front fractionation methods. For example, many biological processes are influenced and identified through altered glycosylation events on proteins and may be targeted by using specific lectins, as will be demonstrated in Aim III, to capture specific carbohydrate moiety carrying proteins. Many other post-translational modifications and truncations are also possible. Thus, inclusion of the trypsinization step can facilitate detection of these modifications or truncations due to the altered spectra of the affected peptides.

The majority of the proteins identified in our cataloguing analysis have been reported previously in the serum after the use of techniques that enrich for LMW proteins (64). Many of the proteins we identified are found in the median concentration ranges of serum and plasma proteins. For example, the protein Platelet-basic protein(PBP)/beta-thromboglobulin/NAP-2 belongs to a family of CXC cytokines and is reportedly found in serum as multiple isoforms in the $\mu\text{g/mL}$ concentration range (135). Additionally, the differential protein Bcl-9 from the SOF study is a ubiquitously expressed, but low-abundant, nuclear protein. It is unlikely that the WCX fractionation alone could account for the enrichment of some of these low-abundant proteins. However, as previously mentioned, it is possible that these proteins are enriched on carrier proteins or bound to other larger protein complexes become released during the efficient trypsinization process. Specifically, the aforementioned PBP/beta-thromboglobulin/NAP-2 protein has previously been reported as albumin-associated (136).

The trypsinization step which we have integrated after fractionation of serum is an effective means to profile serum and acquire identifications of peaks of interest given that the LIFT- MALDI-TOF/TOF is effective at the $< 4,000 m/z$ range (19). Traditional in-solution protein digestions can be tedious, require long incubation times, and the amount of trypsin included is limiting to minimize interfering auto-digestion peaks. The method we describe takes advantage of the fact that immobilizing enzymes can yield reactions that are faster, more efficient and have high-throughput (31, 32). This is due in part to the increased stability of the immobilized enzyme and also to the ability of using higher enzyme-to-substrate ratios. There have been other approaches to immobilize trypsin onto solid supports to increase its catalytic ability, thus minimizing the time needed for

digestion and streamlining the trypsinization process. Innovative approaches such as trypsin adsorbed directly onto a metal MALDI plate (33, 34), linked to copolymer MALDI sample array chips (35) or immobilized onto different monolithic HPLC columns (32, 36) have been described. Currently, trypsin is also commercially available bound to agarose beads or immobilized as individual spin columns. However, in terms of profiling studies, the method we describe is advantageous in that it allows for the total automation of the trypsin digest protocol, in a timely, efficient manner. Additionally, it may be effectively combined with various robotic fractionation techniques in tandem, creating workflows that allow for the effective chromatographic separation of large sample sets with minimal operator error for the purpose of exploring complex proteomes in greater detail. This robotic workflow will also be examined further in Aim III.

Currently, we are using both the WCX/trypsin beads/C18 scheme and the WAX/trypsin beads/C18 scheme to profile different sample sets. A challenge when peptide profiling using the MALDI-TOF/TOF is that the spectrum is crowded in the <4000 Da mass range. Thus, there are peaks that are much more difficult to identify using LIFT due to other peptides over-shadowing their signal or fragmenting along with the peak of interest. Thus, one way to determine the identity of a peptide that is presenting in this manner is to employ an LC-MALDI-TOF/TOF to reduce the complexity of individual spots on the target plate, while retaining all the information in the increased number of elutions of a specific sample.

Fortunately, in both our clinical cohort examples we were able to identify all major peaks of interest. In the SOF study, processed with the WCX fractionation/immobilized-trypsin bead workflow, the most compelling peptide in terms

of reproducibility and intensity was the m/z 2017 peptide from the alpha-2HS-glycoprotein. This protein was elevated in individuals that will develop breast cancer in the future. The true biological role of alpha-2HS-glycoprotein is unknown, but it has been implicated as a negative acute-phase protein. It is highly expressed in developing fetal tissue, but is dramatically decreased in adults. Additionally, it has been found that normal circulating levels in adults (300-600 $\mu\text{g/ml}$) fall significantly during injury and infection (137). Interestingly, one group of researchers, using Lewis lung carcinoma cells as a cancer model, reported finding that the lack of the mouse form of alpha-2HS-glycoprotein (fetuin-A) significantly protects those mice from developing tumors *in vivo*. Furthermore, according to the authors, fetuin-A is capable of promoting the growth of more aggressive tumor cells, but not benign and normal cells *in vitro* (138). Hence it would be plausible that the elevation of this protein, which is thought to act in the neutralization of inflammatory components, may yield a permissive microenvironment for breast cancer to develop without the intervention of the immune system.

Another peptide that was elevated in women that develop breast cancer was m/z 2383, a peptide from the Bcl-9 protein. The Bcl-9 protein, whose overexpression is associated with B-cell malignancies, is a coactivator (and also a nuclear shuttle protein) that binds unphosphorylated beta-catenin thereby aiding in the activation of Wnt target genes (139, 140). This Wnt signaling pathway is pathologically activated in many different types of human cancers and research has shown that resulting WNT proteins are overexpressed in different tumors (141, 142). Interestingly, components of the Wnt/ β -catenin pathway have also been found activated in up to 60% of breast carcinomas (143, 144).

As for the BPH/PCa cohort, the most striking differential protein was ApoAIV. Two peptides from this protein, m/z 1216 and 1353, were increased in patients diagnosed with BPH as compared to patients diagnosed with PCa for samples processed with the WCX fractionation/immobilized-trypsin bead workflow. Additionally, the ApoAIV m/z 1216 peptide was also increased in the BPH patients as compared to PCa patients as seen by the WAX fractionation/immobilized-trypsin bead method. ApoAIV is a plasma protein that circulates freely in solution or associates with chylomicrons and high-density lipoprotein (HDL) (145, 146). This apolipoprotein does not have a known function though it is thought to have a role in lipid absorption, transport and metabolism, and may act as a satiety signal. It has also been suggested that higher levels of ApoAIV leads to an increase of chylomicron formation, which is then responsible for a more efficient absorption and amplified lymphatic output of the carotenoid, lycopene (147). Lycopene is a powerful antioxidant that has been shown to scavenge oxygen free radicals and also to interact with reactive oxygen species thereby protecting cells from oxidative damage. Oxidative damage caused by free radicals to cellular proteins, lipids and DNA has been implicated as a possible mechanism for the propagation of cancer, including PCa (148). Additionally, lycopene has been implicated in the prevention of PCa by affecting various signaling pathways (i.e. insulin-like growth factor) (149). However, the extent of lycopene's ability in thwarting PCa is still very controversial with conflicting reports coming out regularly (150).

One study recently found that plasma lycopene concentrations are decreased in subjects with localized and metastatic prostate cancer as compared to normal and BPH diagnosed subjects (151, 152). This is also validated by the finding that malodialdehyde

levels (which are used as an indicator of lipid oxidation) were increased in PCa patients as compared to BPH and normal individuals (152). Interestingly, ApoAIV has also been shown to be a potent inhibitor of lipid oxidation (153). Thus, a decrease of ApoAIV may lead to an elevated risk of PCa owing to the increase in lipid oxidation caused by the diminishment of antioxidant activity through lycopene and/or ApoAIV itself.

In both serum cohorts that we examined, the protein kininogen-1 (represented by m/z 1031) was elevated in women that were going to develop breast cancer and also men that were diagnosed with PCa. It is interesting that kininogen-1 is increased in patients that have or will develop cancer. Kininogen-1, also known as alpha-2-thiol proteinase inhibitor, is a blood protein involved in the kallikrein-kinin system, which is involved in blood clotting. There are two types of kininogen, high molecular weight (HMW) and low molecular weight (LMW), which are both composed of a heavy chain and a light chain. HMW-kininogen (HK) and LMW-kininogen (LK) have an identical heavy chain sequence, but a differing light chain sequence. HK binds to endothelial cells where it can be cleaved by plasma kallikrein to release bradykinin (BK). The remaining portion of this protein is termed cleaved high-molecular-weight kininogen (HKa) (154). Recent studies have found that HKa is anti-angiogenic, while HK, BK and LK are pro-angiogenic (155). The HKa anti-angiogenic properties, which include inhibition of endothelial cell migration and proliferation, have all been allocated to the D5 domain (located in the light chain). Whether HKa or HK are further proteolyzed *in vivo* to release D5 is unknown, but current *in vitro* data support this possibility (154). The peptide we found to be increased in patients that will develop or currently have cancer

could be a component of HK, LK or HKa (but is not part of D5 or BK), since it is located in the heavy chain region.

In adults the vasculature is quiescent and tightly regulated by a balance of pro- and anti-angiogenic factors (156). Thus, it could be possible that in individuals at risk for cancer, or who have cancer, this system could be deregulated. For example, there may be an over-abundance of pro-angiogenic factors (HK, BK and LK), while the anti-angiogenic factor, HKa, may be misfolded (hiding the D5 domain) or may be degraded or subjected to missed cleavage by proteases, rendering it incapacitated. Alternatively, since angiogenesis is a complicated process involving many factors, the kininogen protein may be elevated, but there may be contact or signaling problems downstream in the anti-angiogenic pathway, thus making the pro-angiogenic pathway dominant. However, a problem with analyzing the true importance of this protein in clinical sample studies is that HK (but not LK), as mentioned above, is part of the clotting cascade and thus must be looked at with caution, since improper sample handling and storage may affect its presence.

These cohort studies demonstrate that it is unlikely that there will be one sole biomarker that will effectively (with perfect sensitivity and specificity) be used in the early detection of cancer, be it breast cancer or prostate cancer. As discussed in the introduction of this thesis, cancer is a very complicated state, with many elements working together and against each other to create a favorable tumor microenvironment. Thus, it is more reasonable that a panel of biomarkers, along with specialized algorithms, will assess an individual's risk of developing a certain type of cancer (or detect that cancer early in its development).

CHAPTER V

AIM III: DEVELOPMENT OF INTEGRATED FRACTIONATION PROTOCOLS FOR IN-DEPTH AND AUTOMATED MALDI-TOF/TOF ANALYSIS

5.1 Introduction

In general serum expression profiling is reproducible and portable across multiple laboratories, especially when rigorous study design and sample handling are combined with carefully controlled instrument calibration, automated sample preparation, and supervised bioinformatic data analysis (47, 50, 62, 126). Furthermore, the recent development of TOF/TOF technology has brought with it the capability of protein identification (m/z 4000 or less) through the generation of fragment ions and homology searching. As discussed in Chapter IV, this has alleviated one common problem in expression profiling: the intricacy of determining protein identities of potential biomarker peaks. However, another difficulty that still remains in expression profiling, particularly for serum and plasma studies, is the issue of protein dynamic range and complexity, highlighting the need for new strategies to increase the utility of these techniques for clinical biomarker assay development (43, 47, 71, 125).

The majority of the proteins that are seen as peptides by the WCX and WAX workflows described in Chapter IV are all high-abundant proteins, whose abundance may not necessarily be reflective of disease state in relation to cancer. These proteins are considered acute phase proteins, or host response proteins, that are mainly synthesized in the liver and reflective of the host immune response (157). Additionally, many criticize

the use of these host response proteins as biomarkers, since some are altered by the clinical processing of samples (especially serum samples) (57, 65).

This issue may be addressed by using a targeted fractionation method, which involves using immobilized lectins to separate glycosylated proteins from non-glycosylated proteins and even separate differentially glycosylated proteins. This is comparatively a more biologically relevant fractionation method since glycans participate in many biological processes such as cell adhesion, molecular trafficking and clearance, cell recognition and the immune response (124). Notably, it has been shown that glycosylation profiles in the cell change significantly during oncogenesis and that altered glycoforms may lead to cancer progression. Since, the blood is enriched for secreted and shed cell surface glycoproteins, it therefore contains many potential biomarkers that may reflect disease specific glycan differences (157). One example of a secreted glycoprotein is the aforementioned PSA, of which multiple glycoforms have been described (71, 158, 159).

The “bottom-up” expression profiling approaches, described in Chapter IV and further elaborated on in this Chapter, typically utilize magnetic bead surfaces and are thus designed to be compatible with automated robotic sample processing. Automation, using the ClinProt robot and the MALDI-TOF/TOF instrument from Bruker, would be advantageous because it is high-throughput, reproducible, limits operator error and consumes small amounts of the patient’s sample (62, 71, 72). This leads not only to more significant results due to the increase in sample numbers, but may also provide an ideal technique that translates effectively into clinical, diagnostic laboratories.

In this Chapter we revisit the immobilized-trypsin bead technique, developed in Chapter IV, and adapt it into two different approaches with the goal of a more comprehensive look at the serum profile. We began by first examining tandem bead approaches for the purpose of sample preservation and also to determine if more information may be garnished from an already fractionated sample through fractionation with another chemical-affinity bead type. We next explore lectin capture strategies with our established immobilized-trypsin bead workflow. All of these approaches are designed with the goal of automation since, as discussed previously, the benefit to robotic automation is that it allows for the opportunity of analysis of large numbers of samples comprising many clinical cohorts with limited operator variation. Thus, we end this Aim with a look at the automation of the WCX workflow that was performed manually in Chapter IV. Challenges of this automated workflow are discussed, with a final robotic schematic design presented.

5.2 Materials and Methods

Serum Samples

A pooled human serum sample collected from over 360 donors (50) was used for method development procedures.

Tandem-bead workflows

Initial fractionation of serum was done with either MB-WCX (weak cationic exchange) or MB-WAX (weak anionic exchange) paramagnetic beads essentially as described by the manufacturer's protocols (Bruker Daltonics, Bremen, Germany).

Briefly, 20 μL of serum was mixed with 40 μL of binding solution (the WAX protocol utilized the pH 5 binding solution) supplied with the beads and 20 μL MB-WCX or MB-WAX beads (note that the WAX beads were equilibrated with activation solution prior to this step) for 15 minutes (mixing every 5 minutes). A magnetic bead separator was used to concentrate the beads and for the wash/rinse processes. Unbound serum proteins were removed for processing with the opposite bead-type (if on WCX then added to WAX and vice versa) and the bound serum proteins (after washing) were eluted with 10 μL of MB-WCX or MB-WAX elution solution supplied by the manufacturer. Eight microliters of HPLC water and 1 μL of MB-WCX stabilization solution were added to the WCX eluate, and 11 μL of MB-WCX elution was added to the WAX eluate, to give a final sample pH of 7.5-8.5. Finally, the unbound serum proteins were processed with the opposite bead type using the same workflow as described above. The pH was adjusted according to their secondary fractionation bead type.

Magnetic bead-bound lectin workflows

Magnetic bead (MB)-ConA (concanavalin A) and MB-WGA (wheat germ agglutinin), from Bruker Daltonics, were used to initially fractionate the serum according to the manufacturer's protocol (20 μL serum to 20 μL lectin beads). These lectin beads were used either separately or together in a mixture, where the solutions (binding and wash) utilized were from the MB-ConA kit.

Three separate schemes (discussed in Results) were tested to yield the best MS pattern: 1) Elution using acidic elution solution provided by manufacturer, 2) elution with competitive sugars, and 3) a tandem-bead approach after elution off of lectin beads. Ten

microliters of elution solution provided with the Bruker lectin kit was utilized to elute the bound glycoproteins. To make the pH fall into the ideal range for reduction, alkylation and trypsinization (pH 7-9), 10 μ L of WCX elution solution was added to the lectin elution. Alternatively, between 1.5 and 2 μ L of 1M NaOH was added to the 10 μ L of lectin elution. Two competitive sugar elutions were also tested using MB-ConA as an example lectin (10 μ L of 200 mM or 400 mM mannose was used to elute off of the ConA magnetic beads). For the tandem bead approach, each of the lectin bead types were eluted with 10 μ L of Bruker elution solution, brought up to 20 μ L with 25 mM Ammonium bicarbonate (pH 7.8), and then added directly to a MB-WCX reaction (20 μ L sample, 40 μ L binding solution and 20 μ L beads). Additionally, 2 MB-WGA elutions were pooled into one sample, concentrated under reduced pressure, and reconstituted in their original volume with 25 mM Ammonium bicarbonate (pH 7.8). The entire lectin fractionated sample was then manually processed through the WCX fractionation protocol.

Agarose bead-bound lectin workflow

Serum was first depleted using ProteoPrep ImmunoAffinity Albumin and IgG Depletion Kit (Sigma, Saint Louis, MO). Three-hundred microliters of ConA/WGA agarose-bound lectins (E.Y. Labs, San Mateo, CA) were washed with binding buffer (25 mM Tris, 150mM NaCl, 1mM MnCl₂ and 1 mM CaCl₂). The depleted serum was then incubated overnight at 4°C with the washed lectin beads in a total volume of 200 μ L (the volume was adjusted using the binding buffer). The sample was eluted with 100 μ L of the competitive sugar. This glycoprotein elution was subsequently acetone precipitated

and the resulting pellet was reconstituted in 80 μL of 25 mM Ammonium bicarbonate (pH 7.8). Twenty-two microliters of sample was added into the reduction and alkylation reaction described below.

Automated MB-WCX workflow utilizing the ClinProt robot

The automated workflow for the MB-WCX beads was performed very similarly to the manual method described in Chapter IV. Briefly, 20 μL of serum was mixed with 40 μL of binding solution supplied with the beads and 20 μL MB-WCX (in the first automated run 10 μL of beads were added). A magnetic bead separator was used to concentrate the beads and for the wash/rinse processes. Unbound serum proteins were removed and the beads were washed 3 times with the MB-WCX wash solution. Bound serum proteins were eluted with 10 μL of MB-WCX elution solution supplied by the manufacturer. The final addition of the stabilization solution supplied with the bead kit was replaced with an addition of 8 μL of water.

Manual and automated reduction, alkylation and digestion using immobilized-trypsin beads

For reduction and alkylation, $\sim 8 \mu\text{g}$ of the fractionated samples were reduced with 8 mM DTT in 25 mM ammonium bicarbonate (pH 7.8) at 56 $^{\circ}\text{C}$ for one hour (24 μL total volume). For the automated run, 8 μg was estimated as 10 μL of the final WCX eluate. Also, apart from the agarose-bound lectin eluate, the rest of the lectin eluates had very low protein concentrations and therefore 10 μL of each of these eluates was added to their respective reduction reaction (as this was the maximum volume recommended).

The reduced samples were then alkylated with 17 mM iodoacetamide in 20 mM ammonium bicarbonate total solution (29 μ L total volume).

The resulting tryptic peptides were re-captured and concentrated with HIC-C18 paramagnetic beads (Bruker Daltonics, Bremen, Germany) as follows. Twenty microliters of the tryptic digest was incubated with 10 μ L HIC-C18 beads and 40 μ L HIC-C18 binding buffer. Bound peptides were washed with the manufacturer's wash solution and eluted in 10 μ L of 50% ACN as per the manufacturer's specifications. It should be noted that for the lectin capture studies we also attempted a C8 clean-up, however the results were similar to those seen for the WCX trypsinization workflow in Chapter IV, in that the C18 bead type produced more robust spectra. Thus, MB-C18 was retained as the preferential bead-type for clean-up and concentration of peptides for analysis on the MALDI-TOF instrument.

MALDI-TOF analysis

The tryptic peptide samples (for the WCX and WAX tandem bead workflow and the automated WCX workflow) after HIC-C18 clean-up were mixed in a 1:3 dilution (for manual method of spotting) or 1:5 dilution (for automatic method of spotting) with CHCA matrix solution (4 mL ethanol, 2 mL acetone, 0.008 g CHCA, and 0.1% TFA) and 1 μ L of the mixture was manually spotted (or robotically spotted by the ClinProt robot where indicated) onto an AnchorChip plate using a dried droplet spotting technique. The lectins were spotted using a reverse-thin layer spotting technique described in the Materials and Methods of Chapter IV.

The UltraFlex III MALDI-TOF/TOF instrument (Bruker Daltonics, Bremen, Germany) was used to analyze peptides in reflectron mode and the resulting spectra were processed using FlexAnalysis 2.0 or ClinProTools 2.0 (Bruker Daltonics, Bremen, Germany). Peaks of interest were further analyzed on a separate platform using the LIFT function of a MALDI-TOF/TOF Ultraflex III instrument. The BioTools software and the MASCOT search engine (www.matrixscience.com) were used to compare the TOF/TOF spectra against primary sequence databases (SwissProt) to determine peptide sequence identities (unless otherwise noted the search criteria is as follows: carbamidomethyl and oxidation modifications; 100 ppm mass tolerance MS; 0.5 Da MS/MS tolerance).

5.3 Results

Tandem-bead workflows

We first investigated whether the workflows discussed in Chapter IV would be compatible if performed in tandem and if so would these types of new workflows yield any new protein observations. Figure 29 illustrates the workflows from Chapter IV and the tandem workflows that are described in this section.

We first analyzed a tandem workflow that involved an initial WAX fractionation, with the unbound fraction going directly into a WCX fractionation reaction. After the WCX fractionation, the eluate was reduced, alkylated, trypsinized using immobilized-trypsin beads, and the peptides were captured onto a C18 bead type. Unfortunately, the resulting spectra, compared to a WCX alone workflow, seemed to be enriched in serum albumin and only 1 peptide, alpha-1B-glycoprotein (peptide m/z 2295.98, R.TPGAAANLELIFVGPQHAGNYR.C, from protein P04217, with expect value of

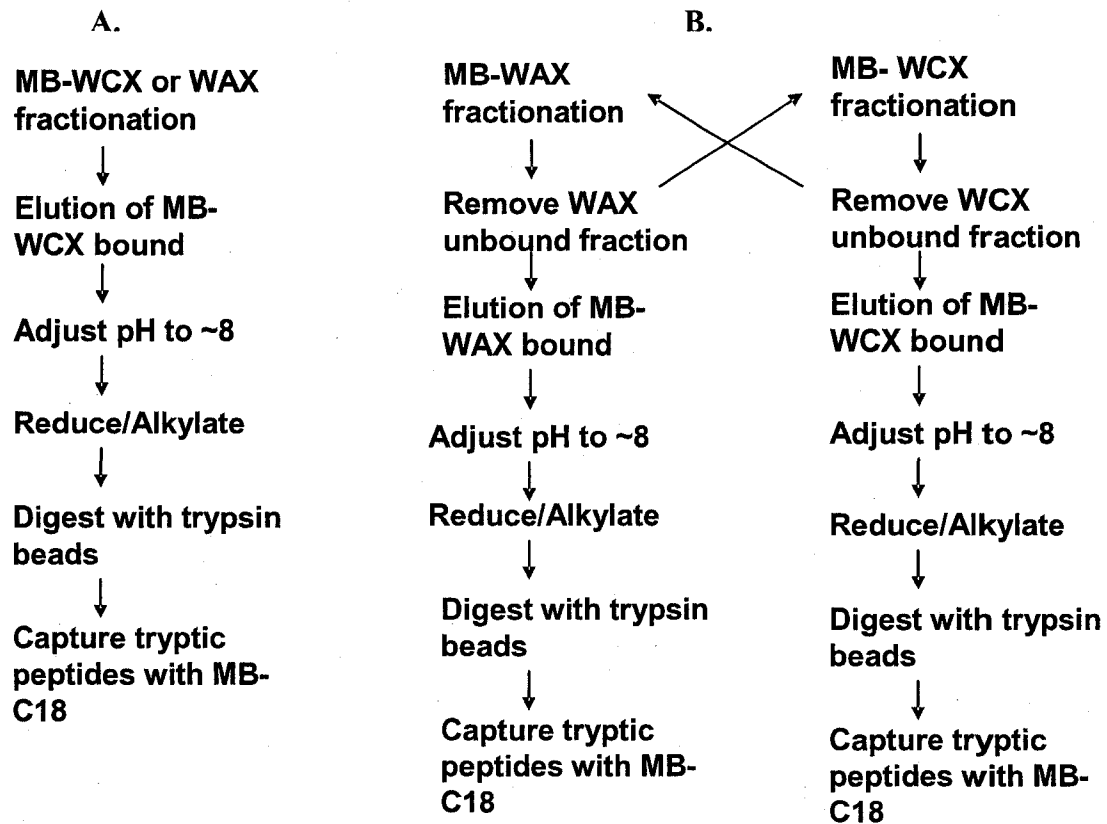


Figure 29. Single bead and tandem bead workflows. A) Workflow depicting the use of one bead type (i.e. WCX or WAX) prior to trypsin digestion (as was described in Chapter IV). B) Workflows depicting utilization of tandem bead types where the unbound serum fraction from one bead type is introduced into another bead type workflow.

4.2e-05) was able to be identified as a peptide that was not seen in the WCX or WAX alone schemes (Figure 30).

We next tested another workflow that involved an initial WCX fractionation, with the unbound sample being directly WAX fractionated. After the WAX fractionation, the eluate was reduced, alkylated, trypsinized using immobilized-trypsin beads, and the peptides were captured onto a C18 bead type. This tandem bead workflow also did not yield more information as compared to the WAX single bead type workflow. In fact the sample appeared to be more enriched with Ig alpha-1 C chain region, with 3 of the most robust peptides originating from this protein. However, one peptide was represented in this WCX/WAX workflow that was not prominent in the WAX (or WCX alone) fractionation scheme and this was alpha-1-acid glycoprotein (Figure 31).

In light of these results, we are not interested in further pursuing these tandem bead approaches for MALDI-TOF profiling. The resulting spectra did not yield enough novel peptides, as compared to the single front-end fractionation described in Chapter IV, to be deemed useful. In fact, both of these tandem workflow methods seemed to be enriched with proteins that are typically thought of as the most abundant in serum i.e. serum albumin and Ig-alpha. However, we have demonstrated that the buffers are compatible between the two fractionation methods; thus, this method may be utilized for samples that are limited in quantity (and thus two separate bead type runs are not an option). Using this approach one may pick the primary bead type of interest, generate spectra from that initial fractionation step, and then still be able to utilize the unbound sample for further analysis with another bead type.

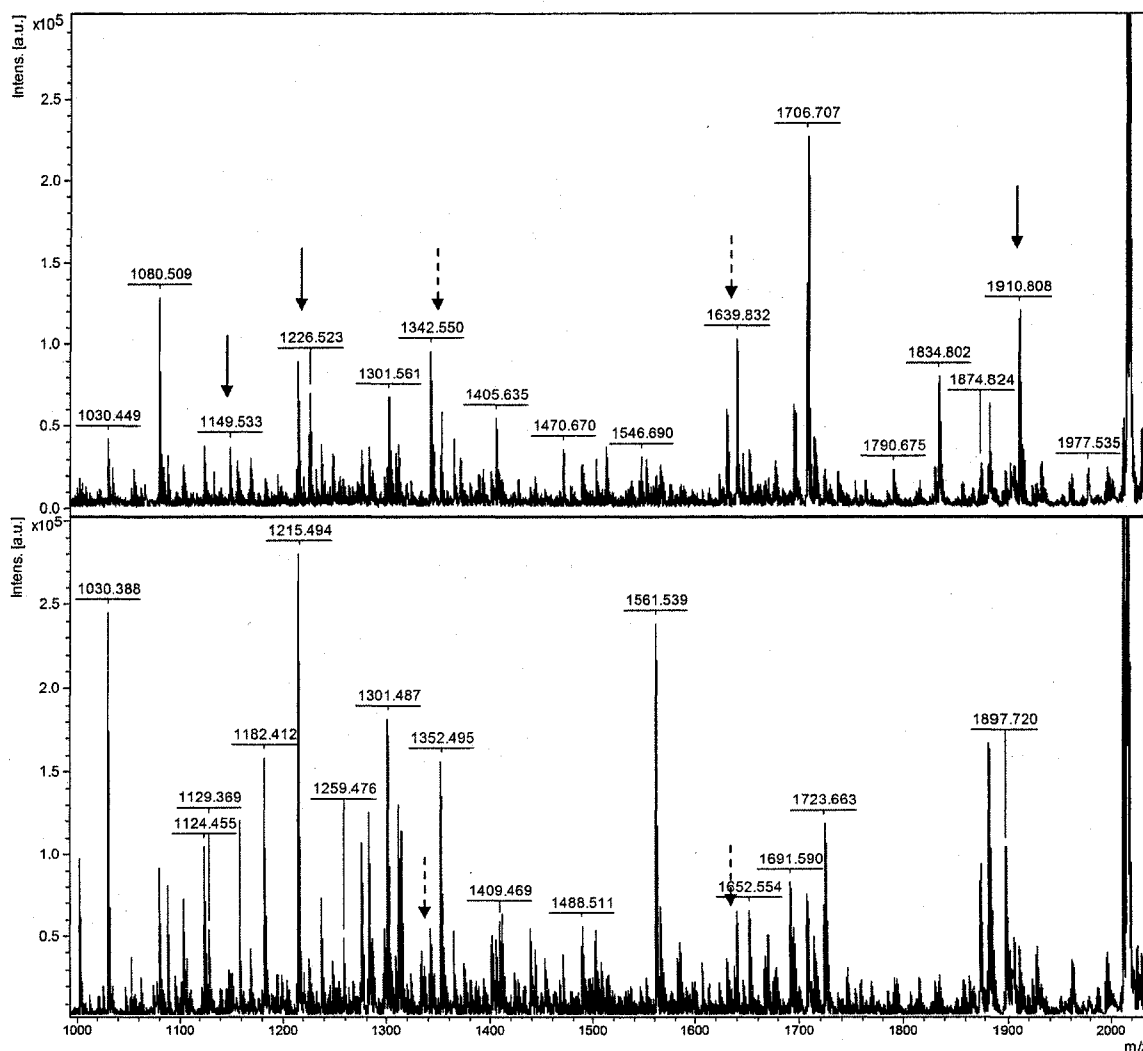


Figure 30. Comparison of spectra from a tandem fractionation scheme and a single fractionation scheme of serum. The top panel illustrates a spectrum of the unbound from a MB-WAX fractionation scheme that was later subjected to MB-WCX/trypsin bead workflow. The bottom panel is a single fractionation of serum with MB-WCX followed by subsequent digestion with immobilized-trypsin beads. All samples were cleaned-up and concentrated with MB-C18 and spotted onto an AnchorChip plate. The samples were then analyzed by the MALDI-TOF/TOF UltraFlex III in reflectron mode and processed by the FlexAnalysis 2.0 software. The arrows denote serum albumin peaks, with the solid arrows indicating peaks that are not shared between the two spectra and the dashed arrows indicating peaks that are shared between the two spectra.

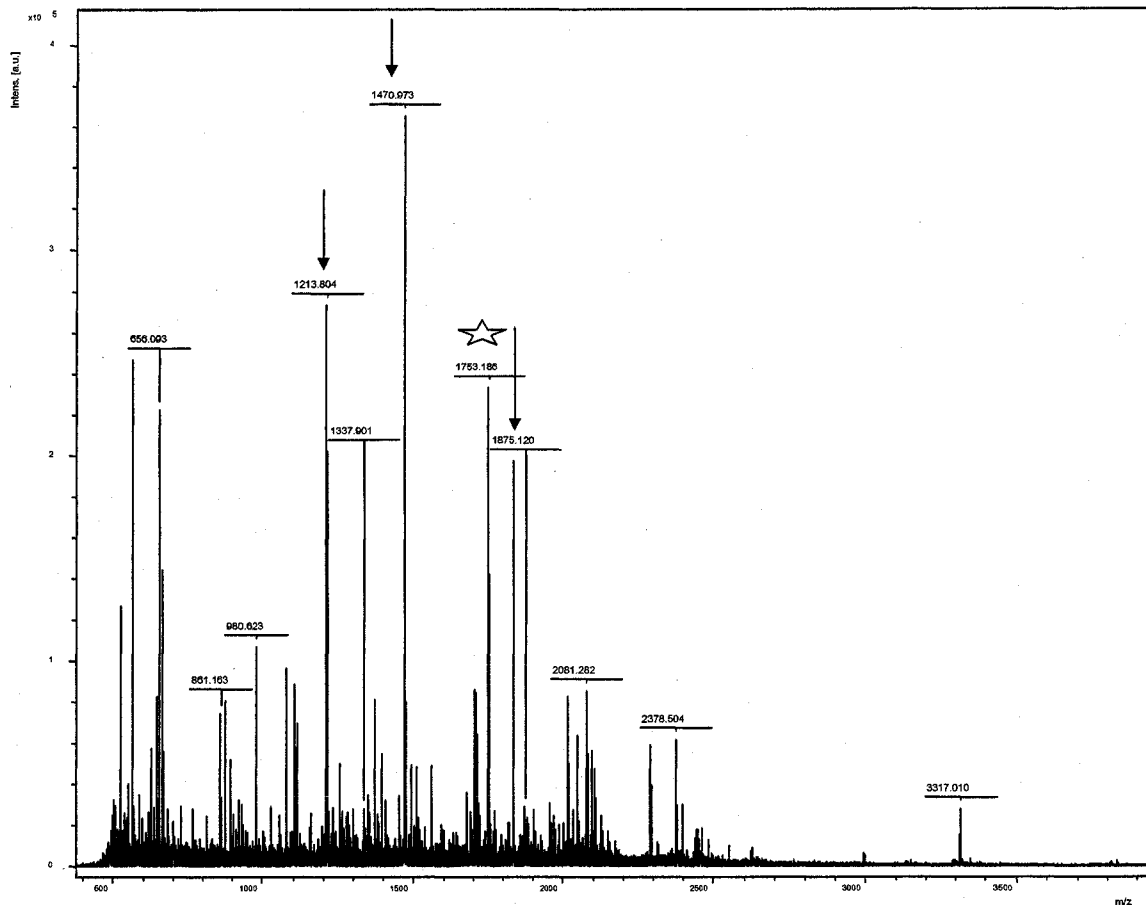


Figure 31. Spectrum of a tandem fractionation scheme. Shown is a spectrum generated from the unbound of a MB-WCX fractionation scheme that was later subjected to MB-WAX/trypsin bead workflow. The tryptic peptides were captured and concentrated using MB-C18 and spotted onto an AnchorChip plate. The samples were then analyzed by the MALDI-TOF/TOF UltraFlex III in reflectron mode and processed by the FlexAnalysis 2.0 software. The arrows indicate the 3 peaks, 1213.80, 1470.97, and 1836.29, which were found to be from the protein Ig alpha-1 C region. Peaks 1213.8 and 1470.97 are typically not seen in a WAX alone fractionation and digestion scheme. However, 1836.29 is seen in both the tandem fraction and single fractionation scheme. The star designates a peptide that is not prominent in a WAX alone fractionation and digestion scheme: alpha-1-acid glycoprotein 1 (expect value: 1.6×10^{-5} and peptide R.YVGGQEHFALLLR.D).

Targeted capture using immobilized-lectins

Because tandem bead workflows using chemical affinity beads did not give us a more in-depth profile of the serum samples, we decided to take a targeted capture approach and interrogate the serum proteome profile with a lectin bead type. We first began with a magnetic bead approach that could be easily adapted to the ClinProt robot. To this end, Bruker lectin beads, MB-ConA and MB-WGA, were utilized individually or together. Unfortunately, it was discovered that the eluted samples had very low protein concentration. Given that the quantity and the type of proteins in solution play an important role in the ability to alter the pH of that solution; we encountered a problem of effectively being able to bring the eluted sample into the correct pH range for reduction/alkylation/trypsinization. Adding a dilute, but high pH base (i.e. WCX elution buffer) made the concentration of the eluted sample very low (i.e. $\sim 0.3 \mu\text{g}/\mu\text{L}$, compared to $\sim 1.3 \mu\text{g}/\mu\text{L}$ for WCX or WAX fractionation schemes after pH adjustment) and yielded spectra with very minimal peaks that were seen in the WCX or WAX scheme (Figure 32). As a reference, the MB-WCX workflow utilized in Chapter IV yielded spectra with an average of 85 peaks in the m/z range ideal for TOF/TOF analysis. Conversely, the peak numbers seen after the lectin workflows were 22, 14, and 8 for ConA/WGA, ConA and WGA respectively. On the other hand utilizing a concentrated, high pH base yielded variable results with a danger of precipitating the proteins from solution. Thus, this was not an acceptable result since this step would not be able to be standardized.

We next attempted to perform a competitive sugar elution off of the lectin Bruker beads (in lieu of the acidic elution solution). Using MB-ConA as an example lectin we

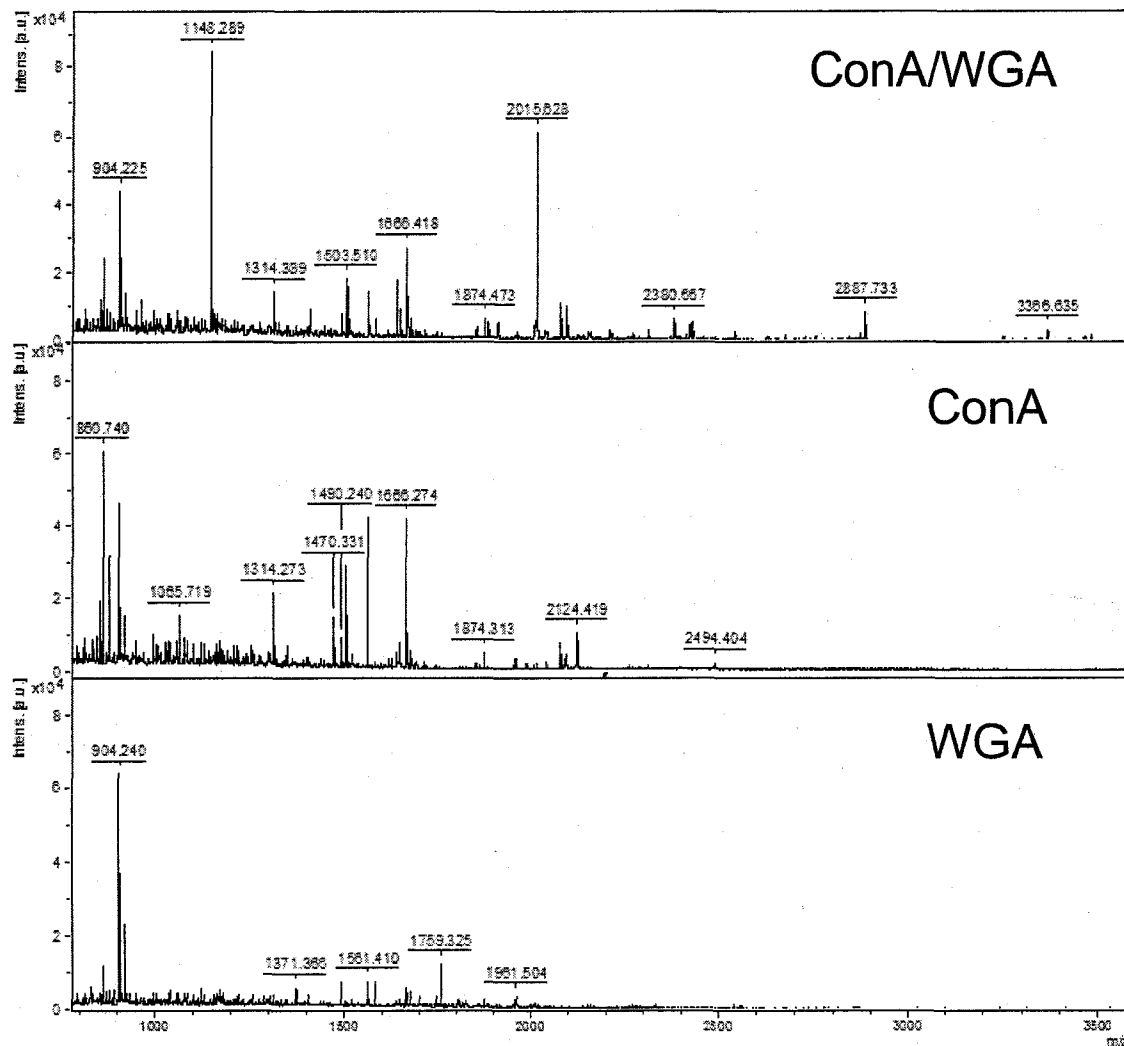


Figure 32. Spectra comparison of serum fractionated with MB-ConA/WGA, MB-ConA, and MB-WGA and eluted with Bruker elution buffer. Samples were processed using either a mixture of ConA and WGA magnetic beads from Bruker or with a single bead type of ConA or WGA. The samples were eluted with the buffer provided by the manufacturer and the pH was adjusted for optimal trypsinization. The samples were then reduced, alkylated and digested with immobilized trypsin beads. The tryptic peptides were captured using MB-C18 and the resulting elutions were spotted using a reverse thin-layer spotting technique. The tryptic eluates were analyzed in reflectron mode using the MALDI-TOF UltraFlex III and spectra were processed using FlexAnalysis 2.0. The top panel shows a spectrum from a ConA/WGA fractionation scheme, the middle panel shows a spectrum from a ConA alone scheme and the bottom panel shows a spectrum from a WGA alone scheme.

found that 400 mM mannose slightly improved the amount of protein eluted and the robustness of the spectra. However, the resulting eluate again had a very similar minimal protein concentration. Thus after reduction/alkylation/trypsinization there was very little complexity (i.e peak number of 10) to the resulting spectra (Figure 33). An additional consideration with this approach is that the competitive sugars used to elute the glycoproteins are more than likely still in the trypsinized sample after MB-C18 clean-up. These carbohydrates bind the C18 bead type and thus when eluted with the peptides may interfere with the generation of the spectra. We utilized a C8 bead type in hopes that a smaller amount of carbons may decrease the binding abilities of the elution sugars. However, using the C8 bead type did not improve the spectra and in fact resulted in reduced abundance of peaks (as it did for the WCX workflow in Chapter IV).

Our final try at using the lectin-immobilized magnetic beads was to utilize a tandem workflow to nullify the elution solution pH concern (since WCX binding buffer has a pH ~ 4). This workflow used an initial lectin bead to capture glycoproteins from the serum and the subsequent elution was added to a MB-WCX fractionation reaction. This approach was moderately successful, but still produced rather sparse spectra. A comparison of the Bruker elution spectra (14 peaks recognized in spectra) and the tandem workflow spectra (20 peaks recognized in spectra) is shown in Figure 34 (using a MB-ConA approach as an example).

In order to demonstrate that these less-than-ideal spectra results are mainly due to low concentration of proteins eluted off of the lectin bead we dried down the 2 lectin elutions (using WGA eluted sample as an example since it yielded the weakest spectra

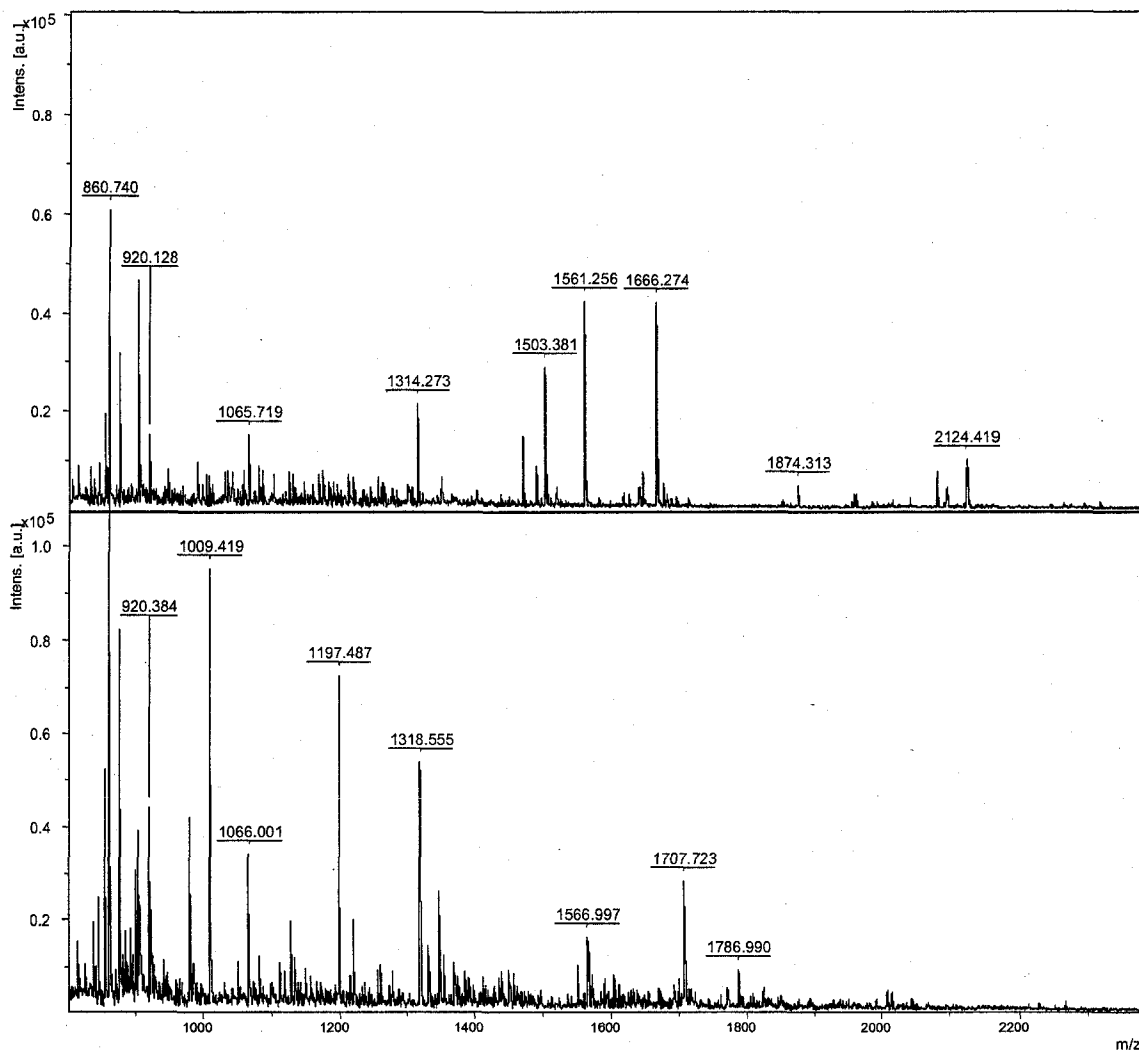


Figure 33. Spectra comparison of serum fractionated with MB-ConA and eluted either with Bruker elution buffer or with a competitive sugar. Samples were processed using ConA magnetic beads from Bruker. The samples were eluted with either the buffer provided by the manufacturer, in which case the pH was adjusted for optimal trypsinization, or with a mannose competitive sugar elution. The samples were then reduced, alkylated and digested with immobilized trypsin beads. The tryptic peptides were captured using MB-C18 and the resulting elutions were spotted using a reverse thin-layer spotting technique. The tryptic eluates were analyzed in reflectron mode using the MALDI-TOF UltraFlex III and spectra were processed using FlexAnalysis 2.0. The top panel shows a spectrum from a sample eluted with the manufacturer's elution buffer and the bottom panel shows a spectrum from a sample eluted with 400 mM of mannose.

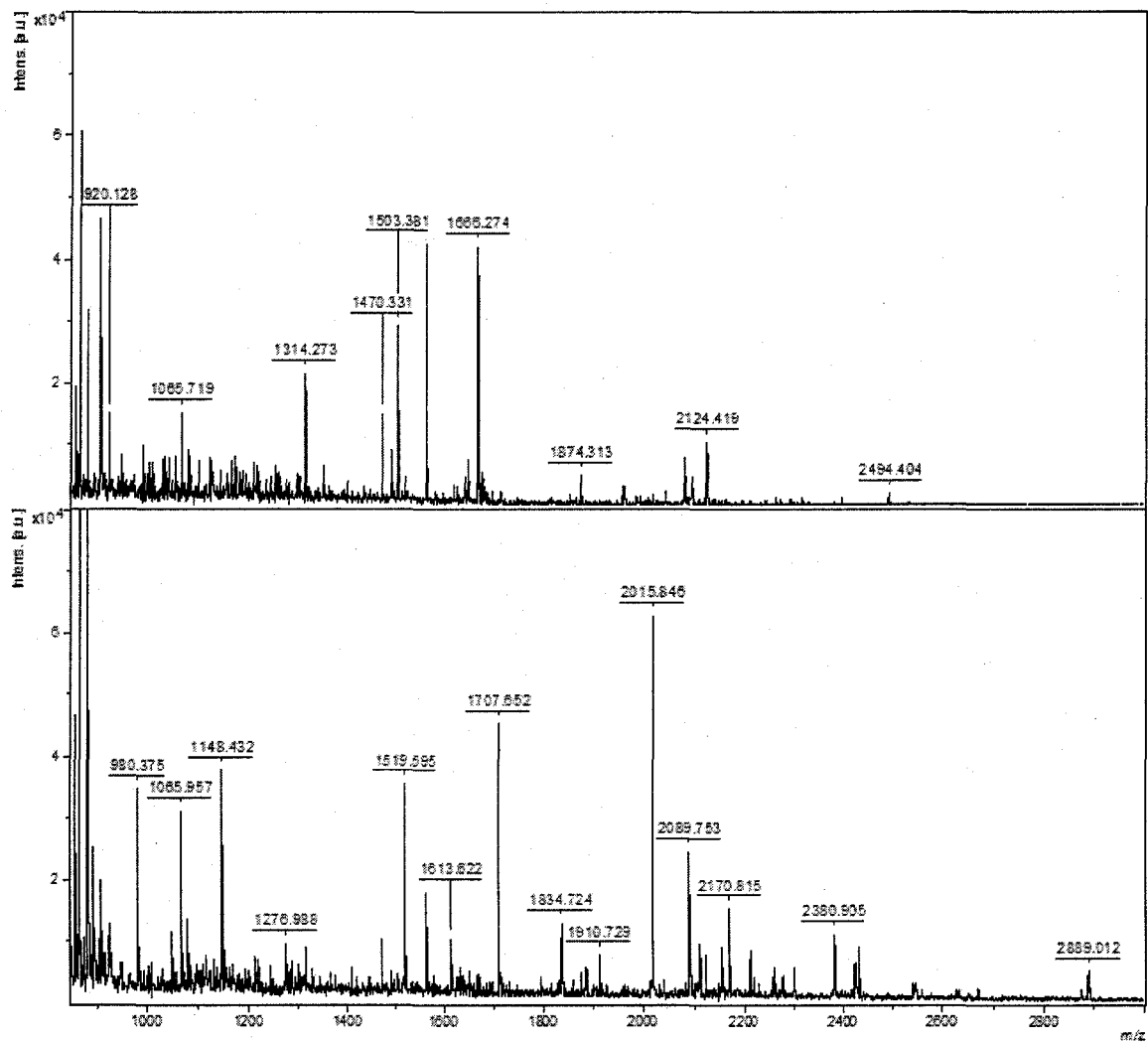


Figure 34. Spectra comparison of serum fractionated with MB-ConA or MB-ConA/WCX. Samples were first processed using ConA magnetic beads from Bruker. The samples were eluted with the buffer provided by the manufacturer and the pH was either adjusted for optimal trypsinization or the entire eluted sample was subjected to MB-WCX fractionation. The samples were then reduced, alkylated and digested with immobilized trypsin beads and the tryptic peptides were captured using MB-C18. The resulting elutions were spotted using a reverse thin-layer spotting technique and analyzed in reflectron mode using the MALDI-TOF UltraFlex III. The generated spectra were processed using FlexAnalysis 2.0. The top panel showcases a spectrum from a ConA alone scheme, while the bottom panel showcases a spectrum from a tandem ConA/WCX fractionation approach.

results) and reconstituted the eluted glycoproteins in 25 mM Ammonium bicarbonate (to yield the acceptable WCX binding pH). The entire glycoprotein solution was then added to a WCX fractionation reaction. Figure 35 shows the resulting WGA/WCX spectra, which has 45 peaks identified by the FlexAnalysis software, compared to the WGA alone scheme, which only had 8 peaks identified by FlexAnalysis. However, we discarded this approach for use with clinical samples because the drying down of samples during an automated procedure would not be very feasible. Additionally, this type of step performed with many samples would add unwanted variation into the protocol, seeing as reconstituting each sample reproducibly may pose a logistics problem.

Since there is not enough material being eluted off of the Bruker magnetic lectin beads (for visualization by MALDI-TOF after trypsinization) and increasing bead volume to capture more glycoproteins would not be cost-effective and also beyond the volume capabilities of the liquid handling robot, we opted to use agarose bound lectins to capture glycoproteins in higher quantities. Our lab has previously worked out the protocol for agarose-bound lectin capture, which includes an initial serum albumin depletion step (71). Additionally, the most abundant elution comes from a combination capture using a mixture of ConA and WGA bead types. Figure 36 shows the spectra generated from the immobilized-trypsin digestion of the ConA/WGA lectin capture. This compares to the WCX workflow method optimized in Chapter IV (i.e. 88 peaks for the agarose-bound ConA/WGA scheme and 85 peaks for the MB-WCX scheme). Table 12 lists the identities of the top 10 peaks from these spectra. Although, the initial capture of these glycoproteins is not compatible with an automated system, the subsequent steps (reduction/alkylation, digestion with immobilized trypsin, MB-C18 clean-up and sample

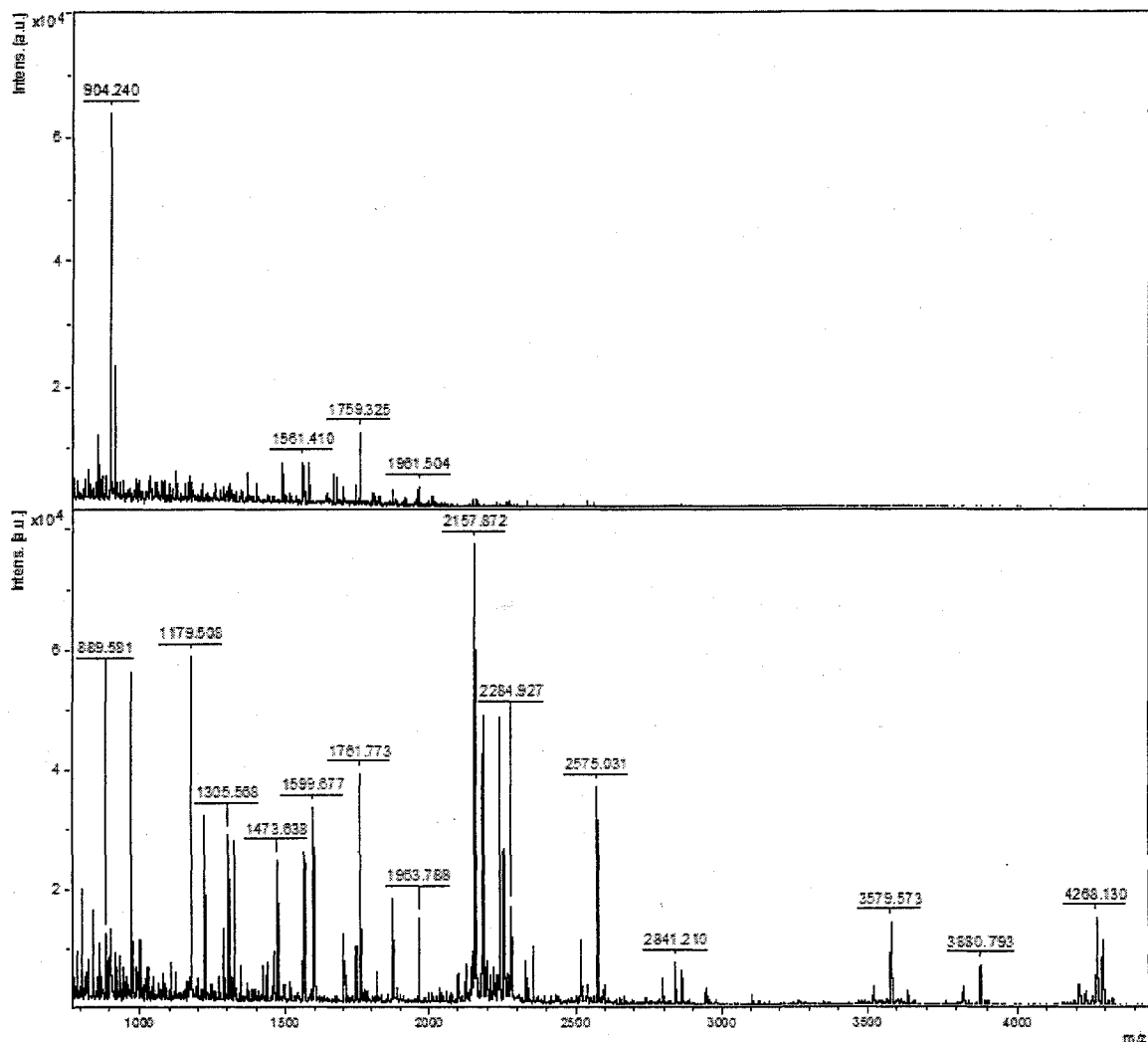


Figure 35. Spectra comparison of serum fractionated with MB-WGA or MB-WGA/WCX. Samples were first processed using WGA magnetic beads from Bruker. The samples were eluted with the buffer provided by the manufacturer. One eluted sample had its pH adjusted for optimal trypsinization, while two of the eluted samples were pooled, dried down, resuspended with 25 mM ammonium bicarbonate and then the entire pooled elution was subjected to MB-WCX fractionation. All the samples were then reduced, alkylated and digested with immobilized trypsin beads and the tryptic peptides were captured using MB-C18. The resulting elutions were spotted using a reverse thin-layer spotting technique and analyzed in reflectron mode using the MALDI-TOF UltraFlex III. The generated spectra were processed using FlexAnalysis 2.0. The top panel showcases a spectrum from a WGA alone scheme, while the bottom panel showcases a spectrum from a tandem WGA/WCX fractionation approach (using double the WGA fractionation sample for the subsequent WCX fractionation scheme).

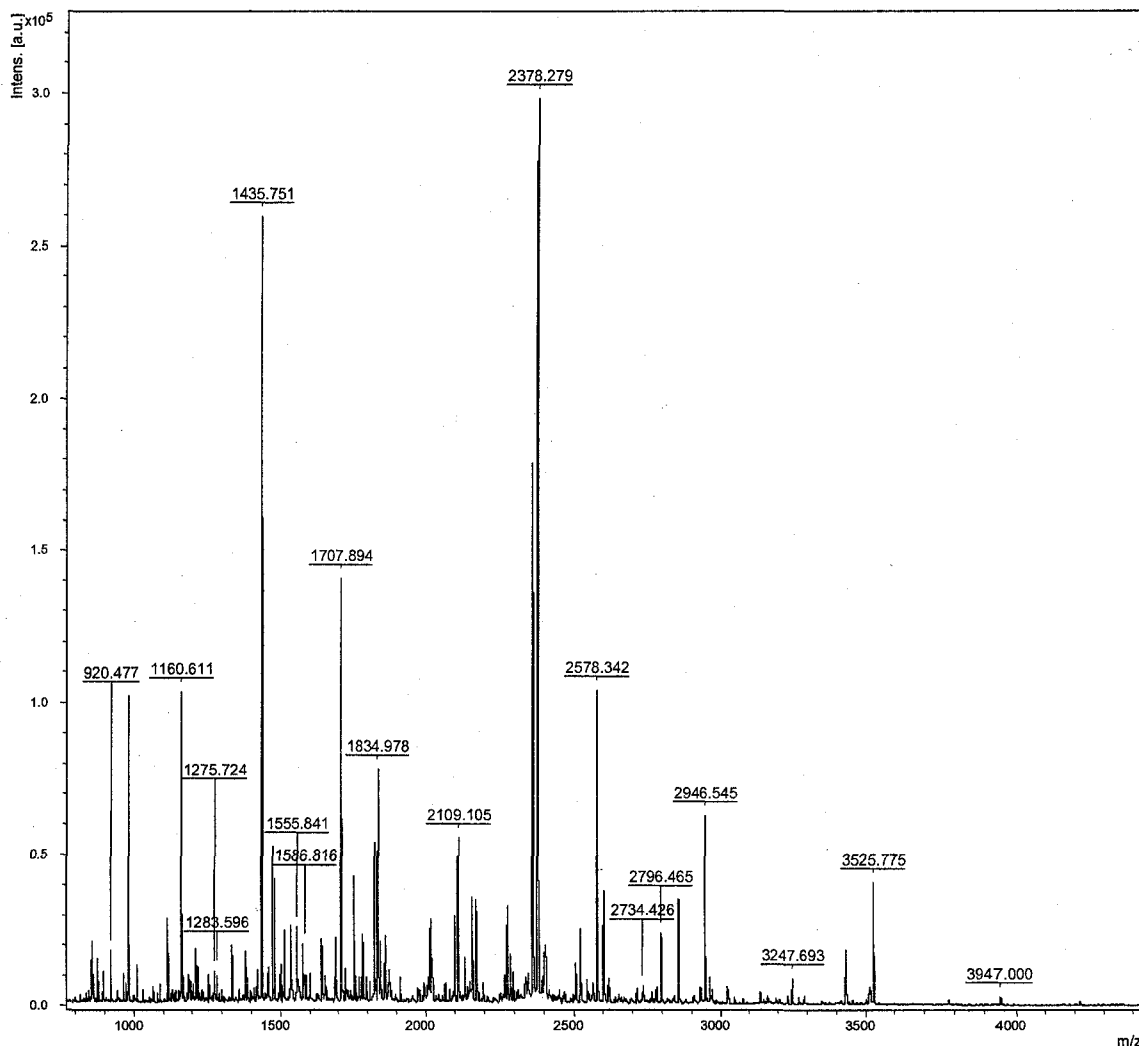


Figure 36. Spectra of serum fractionated with ConA/WGA agarose immobilized lectins. Serum was first albumin and IgG depleted using Sigma immunocapture depletion columns. The depleted serum was incubated overnight with a mixture of agarose-bound, ConA and WGA beads. The sample was eluted with a competitive sugar, acetone precipitated and reconstituted in 25 mM Ammonium bicarbonate. Eight micrograms of the glycoprotein elute was then reduced, alkylated, and digested with immobilized trypsin beads. The tryptic peptides were captured on MB-C18 beads and after elution were diluted with CHCA matrix in a 1:3 ratio and spotted on an AnchorChip plate using a dried-droplet technique. Spectra were generated using the reflectron mode of the MALDI-TOF UltraFlex III. This spectrum had 88 peaks as recognized by the FlexAnalysis 2.0 software.

Table 12. Top 10 peptides seen in the immobilized-trypsin digestion of ConA/WGA fractionated serum after serum albumin and IgG depletion.

Mass	Accession #	Peptide Identity	Score	Expect value	Peptide
980.50	P00738	Haptoglobin	58	6.10E-05	R.VGYVSGWGR.N
1117.66	P01876	Ig alpha-1 chain C region	47	0.00083	R.GFSPKDVLVR.W
1160.61	P02763	Alpha-1-acid glycoprotein 1	62	3e-05	K.WFYIASAFR.N
1642.95	P04217	Alpha-1B-glycoprotein	48	0.00038	R.ATWSGAVLAGRDAVLR.C
1753.01	P02763	Alpha-1-acid glycoprotein 1	46	0.00057	R.YVGGQEHFAHLLLR.D
1825.00	P00739	Haptoglobin-related protein	58	4.4e-05	R.ILGGHLDAKGSFPWQAK. M
1834.98	P00738	Haptoglobin	59	4e-05	R.VMPICLPSKDYAEVGR.V
2016.12	P02765	Alpha-2-HS-glycoprotein	87	4.2e-08	R.TVVQPSVGAAGPVVPPC PGR.I
2097.19	P02763	Alpha-1-acid glycoprotein 1	31	0.016	R.YVGGQEHFAHLLLRDTK. T
2109.11	P00738	Haptoglobin	132	1.6e-12	R.TEGDGVYTLNNEKQWIN K.A

spotting on a target plate) are all designed to be able to be processed by the ClinProt robot.

Automation of the immobilized-trypsin bead protocol

Because the majority of the workflows described in this thesis were designed with automation in mind, we next decided to write parameters for the ClinProt robot (robotic sample handling system equipped with magnetic separation capabilities) to determine if the MB-immobilized-trypsin bead workflow could be reproducibly automated. The robot was first utilized to process aliquots of the same serum sample using the MB-WCX workflow. Since at this point we were only looking at the reproducibility of the immobilized-trypsin beads, we pooled the WCX fractionated serum and introduced 4 aliquots into the trypsin-bead workflow as processed by the ClinProt robot. As with the manual method, the robot first reduced, alkylated and then added the reduced/alkylated samples to the trypsin beads (after first washing the trypsin beads to neutralize their pH). These 4 digested samples were then independently subjected to the robotic MB-C18 workflow and spotted in duplicate. Figure 37, shows the reproducibility of this method visually and Table 13 lists the CVs of 12 representative peaks (these are the same 12 peaks that were used to assess the reproducibility of the manual method in Table 8 of Chapter IV).

Given that the immobilized-trypsin workflow can reproducibly digest the WCX fractionated sample using an automated workflow, we next attempted to completely automate the entire MB-WCX/immobilized-trypsin beads/MB-C18 workflow. Fifteen samples were processed as described in Materials and Methods and spotted in duplicate on an AnchorChip plate by the ClinProt robot. Figure 38 shows the resulting spectra in

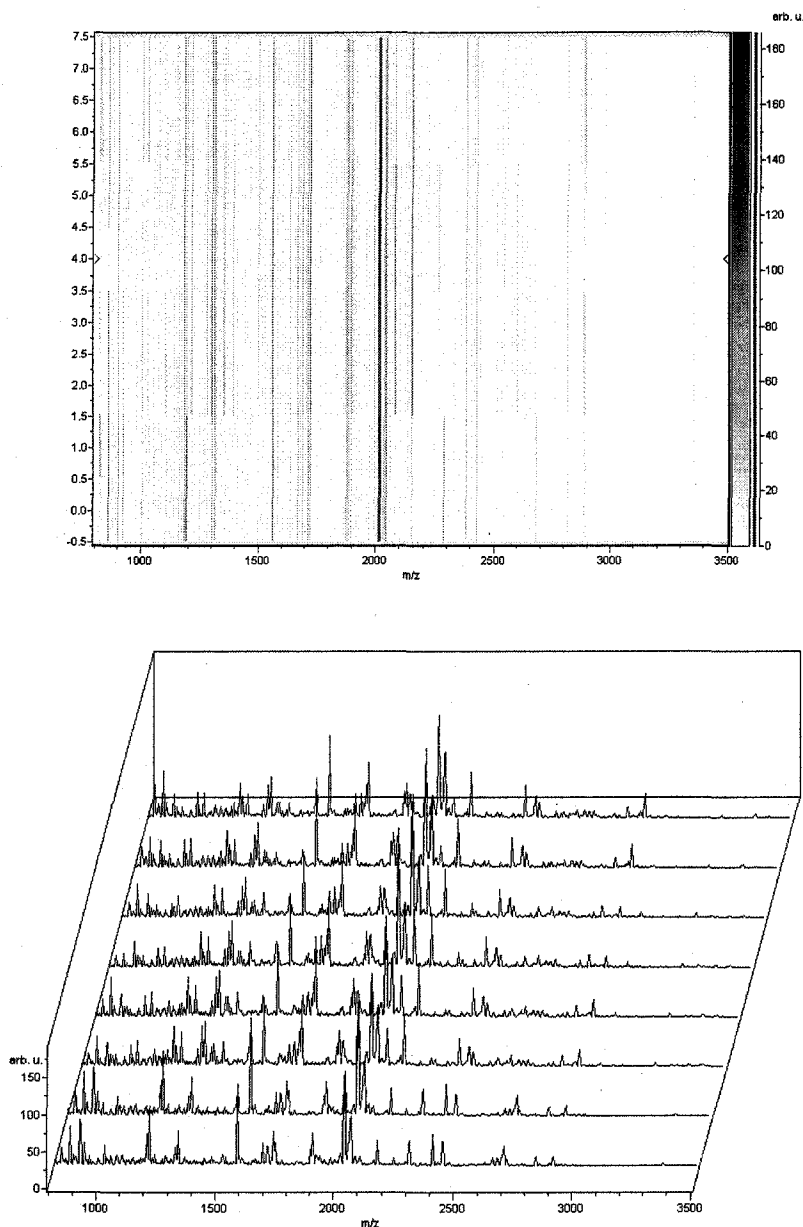


Figure 37. Reproducibility of automated immobilized-trypsin/MB-C18 workflow. Four aliquots of the same serum pool were processed with the MB-WCX workflow by the ClinProt robot. The eluates were pooled together and re-aliquoted into 4 separate samples and then reduced, alkylated and trypsin digested by the ClinProt robot. The robot next processed the tryptic peptides through a MB-C18 workflow and spotted the samples in duplicate on an AnchorChip plate. The samples were analyzed on the MALDI-TOF UltraFlex III instrument and the resulting spectra were analyzed by the ClinProTools 2.0 software. The top panel showcases a heatmap of the spectra and the bottom panel shows the individual spectra.

Table 13. Reproducibility of automated immobilized trypsin bead method as seen by the coefficient of variance (CV) of twelve representative peaks.

Mass	Automatic run	
	Intensity	CV (%)
1124.81	2.28	6.84
1585.99	7.46	7.30
1667.97	6.09	12.30
1670.92	20.99	7.04
1694.93	13.92	8.74
1717.04	23.90	9.07
1885.02	29.19	10.28
1932.26	10.15	6.38
2017.31	211.19	10.30
2383.24	49.79	3.71
2425.80	96.94	6.69
2636.63	18.10	6.03

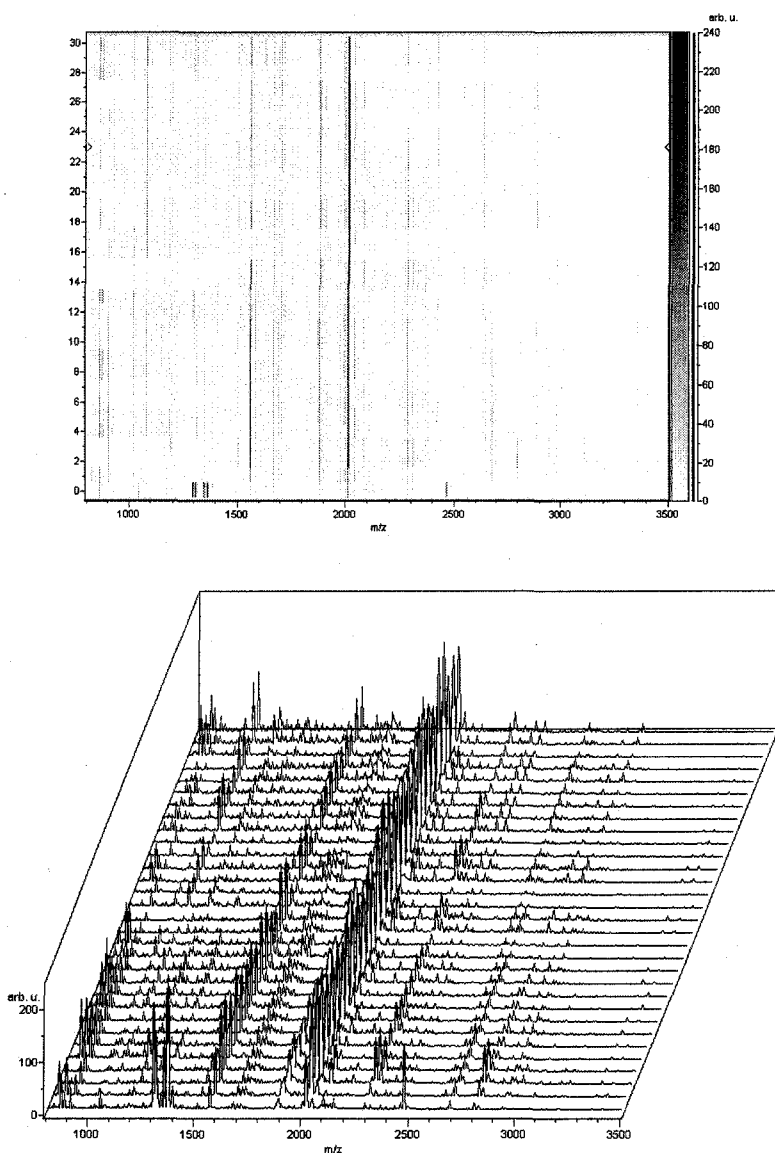


Figure 38. Reproducibility of the automated MB-WCX/immobilized-trypsin/MB-C18 workflow. Fifteen aliquots of the same serum pool were processed with the MB-WCX workflow by the ClinProt robot. The eluates were then individually reduced, alkylated and trypsin digested by the ClinProt robot. The robot next processed the tryptic peptides through a MB-C18 workflow and spotted the samples in duplicate on an AnchorChip plate. The samples were analyzed on the MALDI-TOF UltraFlex III instrument and the resulting spectra were analyzed by the ClinProTools 2.0 software. The top panel showcases a heatmap of the spectra and the bottom panel shows the individual spectra.

Table 14. Reproducibility of immobilized trypsin bead method determined by the coefficient of variance (CV) of twelve representative peaks (with and without outliers).

	15 samples	10 samples
Mass	CV (%)	CV (%)
1124.87	20.21	15.54
1586.21	27.91	13.97
1667.24	18.57	16.36
1671.14	26.52	24.57
1695.17	16.28	12.28
1717.21	21.79	16.95
1885.26	37.15	32.67
1933.40	21.28	16.84
2017.47	36.92	32.15
2383.50	35.68	31.30
2426.20	36.89	32.08
2637.72	32.57	29.19

heatmap and individual spectra form. Additionally, Table 14 lists the CVs of the same 12 representative peaks that were highlighted in Table 8 and Table 12. The CVs were improved on average by 5% if 5 sample “outliers” were removed, leaving 20 samples (10 samples with duplicates) for CV determination (Table 14). Unlike the first automated run (where the WCX fractionated sample was initially pooled giving the digestion reaction a similar starting point) or the manual method described in Chapter IV (where protein concentrations were first determined before adding a set amount of protein into the reduction/alkylation/trypsinization reaction), this automated workflow was not as reproducible. This was mainly due to the ClinProt robot occasionally removing beads from the reactions and mishandling buffers, so that the eluted WCX fractionated samples had variable protein concentrations.

5.4 Discussion

The goal of most proteomic research is to extensively investigate the proteome of the system under investigation in a reproducible and high-throughput manner. Serum expression profiling is a valuable technique as one may compare up to 192 samples on one MALDI plate using the same settings, thus greatly reducing sample to sample variability. However, one drawback to this is that the each sample spot on the MALDI plate is dense with peptides/proteins and thus typically only highly abundant proteins are seen. There is much controversy as to whether these abundant proteins, which are considered acute phase proteins, are truly specific to the disease being profiled, especially when considering how dilute tumor-generated biomarkers are in the blood. Additionally, these acute phase proteins may be influenced by sample collection, processing, and

storage and are therefore analyzed with caution. One way of overcoming this obstacle is to look beyond the abundant proteins through aggressive fractionation protocols.

Another method is to use a targeted capture approach, such as a differential capture of glycan groups on proteins. In this manner the focus is not on the actual abundance of the protein, but on its carbohydrate decorations. As mentioned earlier, several disease, including cancer, have been associated with alterations in glycan moieties.

In this Aim we have seen that tandem bead workflows may be performed as long as buffer compatibility is kept in mind and as long as the final protein yield is at least 5 μg in a 10 μL volume (though 8 μg is ideal). Although the tandem bead workflows described here did not give us a more encompassing look into the serum proteome, we did conclude that this is a good sample sparing technique (critical if your sample set is minimal and precious), in that one may generate both a WCX and a WAX profile from one sample aliquot.

Additionally, a depletion step (of some of the abundant proteins) may be introduced into the tandem fractionation workflow in order to look at lower abundant proteins. This approach, however, is accompanied by many caveats. Firstly, most depletion kits commercially available target serum albumin and/or IgG (such as the Sigma kits used in the lectin workflow). Since serum albumin and IgG make up about 50% of the blood proteome they are wise targets. However, once serum albumin is taken out of the picture some of the lower abundant proteins, such as beta-thomboglobulin (NAP2), which is a cytokine seen in our WCX workflow, would more than likely not be seen. As discussed in Chapter IV, serum albumin is a carrier protein, which when isolated has been shown to be rich of cellular peptides, cytokines and other lower

abundant proteins that typically would not be seen in the blood proteome (64, 136).

Additionally, there are several other abundant proteins, which together with serum albumin make-up about 99% of the blood proteome. Thus, by removing serum albumin and IgG, we are simply making room for the other abundant proteins. A remedy to this would be to use a depletion step that removes the top (i.e. 20) most abundant proteins. However, most of these depletion kits are in single column form and therefore are not adaptable to a high-throughput workflow. Additionally, there may be an issue of sample cross-contamination if trying to process many samples in tandem through one column.

The other goal of this Aim was to adapt our tryptic peptide profiling workflow to incorporate a targeted capture of glycoproteins, thereby turning the focus on glycan moiety differences, instead of strictly protein abundance differences. To this end we first investigated the feasibility of using lectins immobilized to magnetic beads, which would be amenable to a complete automated workflow. However, of these only a ConA/WGA lectin mix proved to be passably satisfactory. We demonstrated that the reason we were seeing weak spectra was mainly due to the low protein concentration of glycoproteins eluted off of the lectin beads. Unfortunately one remedy to this, pooling two eluates and concentrating them into one eluate, would be very low-throughput and highly susceptible to variation as discussed in the Results section. The other solution of increasing the lectin magnetic beads is impractical, as the robot can only handle a certain volume and also at a certain point this would no longer be a cost-effective experiment. A more high-throughput solution involves using agarose-bound lectins. Thus, the front-end fractionation step would be manual, but the reduction, alkylation, trypsinization, MB-C18

clean-up and sample spotting on a MALDI plate could be performed on an automated platform.

Finally, as the goal of all these workflows was to make this tryptic peptide workflow automated, we designed protocols that could be operated by the ClinProt robot. Using a front-end WCX fractionation for the method design, we examined the reproducibility of an automatic workflow starting from the reduction step and terminating with the robot spotting on a MALDI target plate after MB-C18 capture of tryptic peptides. Additionally, we also investigated performing a complete automatic workflow starting with the MB-WCX fractionation and ending with the MB-C18 protocol and spotting of peptides on a MALDI target plate.

We found that the automated trypsin digestion/MB-C18 protocol was highly reproducible, given that the CVs for 12 representative peptides were under 13% (these were the same peptides analyzed for the reproducibility of the manual workflow). However, the robot had a very difficult time reproducibly processing the MB-WCX workflow, resulting in WCX fractionated elutions with varying concentrations. This led to the generation of spectra that were far less reproducible than the manual workflow described in Chapter IV and the automated trypsin bead/MB-C18 workflow described in this Aim (both of these protocols added a constant amount of WCX fractionated serum into the reduction reaction). The same peaks that were found to have CVs of 13% or less in both the manual workflow (Table 8) and the automatic trypsin bead/MB-C18 workflow (Table 13) had CVs ranging from 16% to 37% in the completely automated MB-WCX/trypsin bead digestion/MB-C18 workflow (without 5 sample “outliers” the CVs ranged from 12% to 33%). We therefore recommend (taking into account the high

reproducibility of this method performed both manually and robotically from the reduction step) that serum samples may be initially fractionated manually and then depending on the cohort size may be either further processed manually (if a small cohort) or robotically (if a large cohort).

The importance of creating a workflow that is entirely automated is without question as it will allow for the processing of large numbers of samples that will satisfy statistical requirements of experimental data and will also limit operator interference, thus making the protocols less susceptible to variation between laboratories. However, we conclude that the ClinProt robot is not optimal the platform for automation of the workflows designed in this thesis. Our main concern with this platform is its habit of removing beads from samples during wash steps. This is a haphazard occurrence and therefore produces variation between the samples. Currently, our front-end fractionation methods performed manually produce much more consistent and reproducible results. We will discuss further alternatives to, and adaptations of, the ClinProt platform in the main Conclusions and Future Directions section.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Aim I (Chapter III): Development of Precious Sample Sparing Techniques for Mass Spectrometry Analysis.

A. The results indicated that there was no significant difference in quality of data between scraping a frozen serum sample and thawing the whole sample. Therefore, since our SOF cohort was a compelling, yet limited sample set, we decided to utilize the scrape technique for the processing of all SOF samples and thus preserve the samples for future use without the addition of unnecessary freeze-thaw cycles.

B. SELDI-TOF MS analysis using IMAC fractionation of 42 cases and 42 controls yielded 11 peaks that were predictor variables used in the generation of the classification and regression trees. A tree containing 4 terminal nodes, with 3 peaks used as splitting factors (m/z 7850.989, 9303.888, and 9190.488), was deemed optimal and possessed a recognition capability of 85.7% of cases and 78.6% of controls of the test set. Cross-validation of the generated tree yielded correct classification of 31 of the 42 cases (74% cases correctly classified) and 30 of the 42 controls (71% of controls correctly classified). Unfortunately, this model did not prove successful in validating samples at an independent institution.

C. The best MALDI-TOF MS genetic algorithm model generated using the WCX fractionation scheme with the initial 84 sample set yielded an overall 71.1% recognition

capability between groups. However, this 7 peak genetic algorithm model proved to be over-fitted for the 84 sample set and produced very low external validation sensitivity and specificity (overall recognition capability was below 50%). A 5 peak genetic algorithm model yielded the highest sensitivity and specificity comparatively during external validation of 112 samples, with an overall recognition capability of 61.2% between cases and controls. However, this model performed poorly when used again to identify case and control status of 96 blinded samples (overall recognition capability of 52.1%).

6.2 Future directions of Aim I

The main success of this Aim is the knowledge that serum samples may be scraped without being thawed and will still produce comparable spectra to samples that were completely thawed. In this way the integrity of the serum proteome is conserved, while still retaining the convenience of retaining the samples in their original storage vials. Unfortunately, the classification success of the SELDI-TOF MS and MALDI-TOF MS analyses of the SOF sample set processed with the scrape technique left much to be desired. A very complex question was addressed in this sample set: Can serum protein profiles predict whether women will develop breast cancer in the future based on their serum profiles. This is a very difficult question to pose to two MS platforms that are limited in their ability to penetrate the immense concentration range of the serum proteome, especially for the SELDI-TOF MS due to its flat chip surface fractionation. However, MALDI-TOF MS would be amenable to several front-end fractionations performed in tandem or targeted capture (i.e. lectin capture of glycoproteins) for a more

encompassing examination of the serum proteome of these women. Additionally, as demonstrated in Chapter IV this cohort is being investigated on a small scale with pooled samples subjected to such bottom-up approaches as the immobilized-trypsin bead workflow and quantitative iTRAQ analysis. This will continue with the inclusion of more samples and fractionation techniques. In the future it would also be interesting to see what the albuminome of these patients holds in terms of predictive peptides/proteins for breast cancer risk.

Another consideration in the interrogation of this sample set is that many of these women may not even have tumor development at the time of blood draw. Thus, there may not be tumor-specific biomarkers present in the blood. As mentioned several times in this dissertation, breast cancer is a very heterogeneous disease with many factors (i.e. ER and HER-2 receptor status) playing a role in influencing prognosis and response. Additionally, many genetic and environmental variables have been implicated as factors influencing breast cancer risk. Two such variables, race and hormone-replacement therapy, have been eliminated from the study. However, many other risk factors remain that may be used to stratify the samples. Certain factors such as ER status are beyond the scope of the current sample set, since the original goal of the SOF study was osteoporosis and not breast cancer risk evaluation. However, stratifying patients based on their serum levels of known breast cancer prognosis markers, such as HER-2 and MUC-1, is still an interesting possibility. Although, these biomarkers are typically assayed to determine breast cancer aggressiveness, progression and drug response it would still be interesting to determine if these biomarkers are altered pre-cancer diagnosis and thus underlying genetic factors in breast cancer risk. It may also prove useful to look into BRCA-1 and

BRCA-2 genetic status as a possible stratification factor (though *BRCA* mutations typically affect a small percentage of the breast cancer population) since mutations of these proteins leads to a high risk of developing breast cancer. Furthermore, depending on the information collected along with the original patient consent, we may be able to classify patients based on factors used to evaluate breast cancer risk in the Gail model. These would include age at menarche and age at first live birth, which may be known since these incidents also correlate to osteoporosis risk. Another factor that increases breast cancer risk is obesity and thus BMI (body mass index) may be an additional factor used to stratify the samples. However, this stratification may not be as relevant in 1980s, when this sample set was collected, as it would be today given the steady rise of obesity in the United States. According to the Centers for Disease Control and Prevention, in the 1970s and early 1980s only 15% of the adult population was obese. However, by 2004 33% of the adult population in the United States was considered obese and this situation is worsening with each passing year.

By using these different variables to put patients into more specific groups we may begin to focus on biomarkers related to breast cancer risk, rather than to biological variation between patients that are as complex as the disease itself. These stratifications, along with novel MALDI-TOF technology and aggressive fractionation techniques discussed above, may allow for the discovery of predicative biomarkers for breast cancer. These biomarkers may be incorporated with current breast cancer disease progression biomarkers and risk models in order to design a more sensitive and specific diagnostic tool for physicians to predict an individual woman's breast cancer risk.

6.3 Aim II (Chapter IV). Increasing the Effectiveness of the MALDI-TOF/TOF for Analysis of Large Molecular Weight Proteins.

A. The immobilized-trypsin beads are very efficient at digesting proteins in short incubation times (30 minutes compared to 4 hrs to overnight with traditional soluble trypsin protocols) and digestion efficiency is increased when proteins are reduced and alkylated. Adjusting the pH of the fractionated serum eluates to ~ pH 8 prior to reduction and alkylation also improves the digestion efficiency of the bead. We furthermore determined that for robust and reproducible spectra 8 μg is the ideal starting protein content for reduction and alkylation, with ~ 5 μg of reduced/alkylated sample being digested with 25 μL trypsin beads. Additionally, since the trypsin is attached to paramagnetic beads it can be removed from the reaction by placing the reaction tube against a magnet, thereby not contaminating the spectra with autocatalytic trypsin peptides, a common draw-back to soluble trypsin digests.

B. To create more uniform peptide/matrix spots, and make the peptide spectra more robust, a clean-up step using reverse-phase chromatography beads is necessary. Either ZipTipsC18 or MB-C18 may be used; however the latter is more adaptable to an automated platform. Additionally, it was found that MB-C18 produced more intense and full spectra as compared to MB-C8.

C. A final immobilized-trypsin bead workflow was established that may be applied regardless of the front-end fractionation type: (1) Fractionate serum and adjust pH of eluate to ~8, add 8 μg of the eluate into a reduction reaction, (2) follow with an alkylation

step (29 μL total volume), (3) add 20 μL of the reduced/alkylated sample to 25 μL of washed immobilized-trypsin beads, (4) incubate together at 37 °C for 30 minutes and remove the digested sample, (5) add the 20 μL digest into a MB-C18 workflow, (6) elute the captured peptides off of the C18 beads and spot on an AnchorChip plate (1:3) with CHCA matrix using the dried droplet technique.

D. This final immobilized-trypsin bead workflow was found to be reproducible with minimal intra-spectra variation of samples processed individually, but originating from the same pooled serum sample.

E. Both MB-WCX and MB-WAX front-end fractionation workflows were incorporated successfully with the immobilized-trypsin bead scheme and using LIFT-MS/MS peptide identities were able to be determined for most of the top peaks in the respective spectra. We also demonstrated the minimal amount of information garnered in terms of visualized and identified peptides from undigested MB-WCX and MB-WAX workflows.

F. The immobilized-trypsin workflow followed by ClinProTools analysis and peptide identification by LIFT-MALDI-TOF/TOF proved to be an effective and reproducible scheme for profiling clinical samples and directly identifying differential peptides.

F.1. Pools of the SOF samples (described in Chapter III) were processed with this workflow and 3 peaks were found to be differential: m/z 1031 (kininogen-1), m/z 2017 (alpha-2HS-glycoprotein), and m/z 2383 (Bcl9 protein). The forced peak genetic algorithm model that was generated using these three peaks had a 100% recognition

capability of the test set and a cross-validation of 87.23% correctly classified cases and 77.36% of correctly classified controls. These samples were re-processed months later and the same 3 peaks were used to generate a forced peak genetic algorithm model. This model once again had 100% recognition of the test set and cross-validation yielding 70.83% for both sensitivity and specificity. Additionally, this model was used to externally validate the 1st data set and the result was 17/24 correctly classified cases (70.83% sensitivity) and 20/24 correctly classified controls (83.3% specificity).

F.2. Pools of a BPH vs. PCa sample set were processed using the WCX/trypsin bead workflow and three peaks were determined to be differential between the two groups: *m/z* 1031 (kininogen-1), *m/z* 1216 (apolipoprotein AIV) and *m/z* 1353 (apolipoprotein AIV). The forced peak genetic algorithm that was created using these 3 peaks had 100% recognition of the test set and a cross-validation that correctly classified 84.11% of PCa and 75% of BPH. The samples were re-processed in a blinded manner and then this model was used to classify the blinded samples into their respective groups. The genetic algorithm model was able to correctly classify 10/12 PCa cases correctly and 10/12 BPH controls correctly. This outcome demonstrates the reproducibility of the immobilized-trypsin bead method with clinical samples.

6.4 Aim III (Chapter V). Development of Integrated Fractionation Protocols for In-Depth and Automated MALDI-TOF/TOF Analysis.

A. Tandem bead workflows (i.e. MB-WCX unbound fraction processed by MB-WAX and vice versa and subjected to immobilized-trypsin bead digestion) are a good way to maximize information generated from one sample aliquot. This is therefore a

recommended process if the samples being interrogated are of limited quantity.

However, this tandem workflow does not yield more revealing spectra results for peptide profiling than single bead front-end fractionation methods. In actuality, the tandem bead workflow tended to have a profile skewed slightly towards the two most abundant proteins, serum albumin and Ig alpha, as compared to single bead workflows.

B. Bruker lectins were found to be successfully integrated into the immobilized-trypsin bead workflow if the eluates were concentrated (i.e. using a dry down method of pooled samples), followed by a MB-WCX fractionation workflow, or two lectin magnetic bead types were used simultaneously to capture glycoproteins (i.e ConA/WGA). All of these methods increased the eluted concentration of glycoproteins for further processing by the immobilized-trypsin beads. However, the most ideal glycoprotein output was generated with ConA/WGA lectins immobilized on agarose beads following albumin and IgG depletion. This lectin workflow generated tryptic peptides of the same quality and intensity as the WCX and WAX workflow results in Chapter IV.

C. Because all of the workflows described in this thesis dissertation are either amenable to partial or complete automation, we assessed the reproducibility and feasibility of using the ClinProt robot for processing these workflows (using the MB-WCX scheme as a representative workflow. We found that the partial automation (starting at the reduction step and terminating at the MB-C18 and sample spotting step) is very reproducible as demonstrated by low coefficient of variance values. However, the complete automation (starting at the MB-WCX step and going through to the MB-C18 and sample spotting

step) was not as reproducible, since the ClinProt robot generated initial WCX eluates with variable protein concentrations, which negatively impacted the rest of the protocol. This was mainly due to the robots tendency towards removing trace amounts of magnetic beads during wash steps and other manipulations leading to the elution. Therefore, we conclude that the ClinProt robot may be used for partial automation, but that it may not be the ideal platform for the automation of these workflows.

6.5 Future Directions of Aim II and Aim III

The immobilized-trypsin bead technique described in these Aims is effective at generating peptides that ionize in the ideal m/z peak range for most effective utilization of LIFT-MALDI-TOF/TOF for direct identification of differential peaks. Since the trypsin is immobilized onto paramagnetic beads, this scheme may be adapted to a robotic front-end processing system and allow for a relatively high-throughput, rapid and reproducible method for peptide profiling. This immobilized-trypsin technique may essentially be modified to suite any strategy that would normally utilize in-solution trypsin digestions. For example, we are currently looking into streamlining this method with comparative isotopic mass tagged quantitative approaches (160-162). Additionally, if one adheres to the basic scheme designed in this dissertation then this trypsin bead method may be made compatible with many fractionation workflows still to be envisioned. One such method may involve designing a depletion/capture strategy that is more amenable to high-throughput processing than current commercially available depletion kits. For example, antibodies to human serum albumin may be immobilized on magnetic beads and thus utilized to remove the serum albumin and by the same token

capture it for further analysis. Consequently, from one sample aliquot two separate fractions will be generated: 1) a depleted serum sample that may be further processed with other fractionation schemes to look at peptides and proteins that may have been overshadowed by the bulk of serum albumin, and 2) a serum albumin fraction, which may be used to profile the differences between sample groups based on peptides/proteins that are carried by this molecule.

One obstacle that may be encountered when using this peptide profiling technique is that certain peaks of interest may not be able to be identified by LIFT. This may occur for a number of reasons, one being that a peak of interest may be embedded in a dense peak cluster, thus when the parent ion is fragmented the other contaminating peaks may be fragmented along side it. Another common reason for a peak of interest's inability to be identified by LIFT-MS/MS is that surrounding dominant peaks in the spectra may be suppressing the signal of the protein of interest. A possible solution to both of these problems is the use of LC-MALDI-TOF/TOF to fractionate the sample into many spots onto a MALDI AnchorChip plate based on the peptides chemical affinity properties. In this way the fractionated sample results in the deconvolution of the original tryptic peptide spectra that was crowded into a single spot on the MALDI plate. Thus, we are currently working on adapting the immobilized-trypsin beads for compatibility with the LC-MALDI-TOF/TOF platform.

In our two cohorts investigated in Chapter IV, we discovered several potential biomarkers. It will be necessary to process larger samples sets with our workflow to assess the viability of these biomarkers and the genetic algorithm models using these biomarkers for classification. If these biomarkers hold true for the larger sample set, then

we will be ready to independently validate the results. Therefore, another consideration that will be further investigated is the validation of discovered biomarkers. One approach would be to use commercially available Enzyme-Linked ImmunoSorbent Assays (ELISAs) to measure protein levels of the identified proteins in many serum samples. However, when a peptide is identified as differential between two sample sets, as in our workflow, it is not known if the entire protein expression level is elevated or if the increase is due to the elevation of an isoform, glycoform, or post-translational modified version of this protein. Initial validation of biomarkers using an ELISA would only look at the global change of the identified protein, thus all variations on that protein (if the epitope is present for detection by the antibody) would be incorporated in the expression levels without distinction. However, doing individual western blots for each sample would be tedious and labor-intensive. A remedy to this would be to use an immunocapture strategy with specific antibodies for the protein of interest immobilized on magnetic beads (43, 46). The captured proteins would be eluted, digested and analyzed by MALDI-TOF MS to determine if there are specific peptides generated that are elevated, while the other peptides from this molecule are comparable between groups. The identity of the differential peptide may then be identified using LIFT-MS/MS. However, if the post-translational modification is too complex or more specific information needs to be gleaned from the peptide, such as the site and composition of a glycosylation, multiple-reaction monitoring (MRM) may be utilized (163). A triple-quadrupole linear ion trap instrument is available to our lab for performing such analysis if deemed necessary. Once the reason behind the protein level increase is elucidated, a

more specific ELISA may be designed and processed with many samples for true validation.

Finally, we have demonstrated that by using the Bruker ClinProt instrument with the immobilized-trypsin bead workflow at least partial automation may be attained. The ClinProt robot's main problem is the premature, random removal of magnetic beads with bound protein. Thus, for the full automation capability of the developed workflows described in this thesis dissertation we would need to either evaluate other robotic platforms or look for ways to improve the ClinProt platform. There are other liquid handling robotic systems available, such as the Tecan Freedom Evo, which may be evaluated for their performance with our workflows. However, after observing the ClinProt robot we believe that a single modification in plastic ware may improve its performance. We will therefore evaluate the use of alternative 96-well work-plates than are currently used with this platform. The current work-plates end in a sharp point at the bottom of each well and overall the wells are very narrow. Both of these qualities allow the needles of the ClinProt robot to agitate the magnetic beads during the handling of the buffers, thus permitting beads to be aspirated along with the liquid. Additionally, a rounded 96 well plate may make for easier mixing of the samples, which currently tend to result in the uneven distribution of the magnetic beads throughout the solution due to the clumping of the beads (this especially true for the trypsin beads). Changing the velocity of mixing performed by the robot does not improve this issue. Unfortunately, changing the work-plates is easier said than done, given that the module which holds the plate in place is specifically tailored to this type of plate shape. However, it would be more cost-

efficient to replace this component rather than the entire robot if it was indeed discovered that the work-plate shape influences the performance of the robot.

6.6 Concluding Remarks

SELDI and MALDI are reproducible and portable methods, however as demonstrated by McLerran et al (119), great care has to be put into standardizing sample collection, processing and storage of samples as each of these factors may add sample bias and lead to inaccurate conclusions. One main drawback to SELDI-TOF MS is that it does not possess the capability of identifying peaks found to be differentially expressed. Also, the chip-surface area utilized in SELDI-TOF MS analysis may not lead to efficient fractionation, which is needed for the simplification of the large dynamic range of serum proteome. Since MALDI-TOF MS is not limited by on-chip fractionation protocols it may allow for more extensive fractionation. These front-end fractionation techniques typically employ paramagnetic beads, which not only have more surface area than a flat surface chip, but also allow for in-tandem use of these bead types in an automated manner.

Serum expression profiling typically occurs in the linear range, which is more sensitive in terms of peptide detection. However, it is the mass precision of the reflectron mode that is needed for MS/MS identification of peptide sequences. We have shown that common fractionation schemes such as WCX and WAX do not yield robust spectra necessary for endogenous peptide sequencing. Additionally, the peptides present are difficult to identify since the enzyme that generated each peptide is not known. Overall endogenous peptide profiling with MALDI-TOF is a step above SELDI-TOF MS

profiling since some peptide identity information may be garnered from the spectra analyzed by the MALDI platform compared to no possibility for direct peptide sequencing by the SELDI platform. However, it is necessary to design a scheme that may be used as a companion to endogenous peptide profiling; one that will allow for the visualization of many peptides in the mass range of the reflectron mode and thus permit relatively easy identification by LIFT-MS/MS.

Based on this necessity we tailored a workflow that utilizes immobilized-trypsin beads. We found that this technique is highly reproducible and may be adapted to many different front-end fractionation schemes performed individually or in tandem to reduce the complexity of the serum. Since this workflow utilizes paramagnetic beads it may be either partly or completely automated, though current robotic technology is only capable of reproducible and consistent partial automation. Using these workflows we were able to identify several promising biomarkers in two serum cohorts that are pending future analysis and validation. However, the most important discovery is that these workflows produce consistent results on clinical samples, yielding similar results and conclusions on independent runs. There are many exciting applications for these workflows, as described in the future directions section, and the fractionation schemes described herein are only the start of many different possibilities. We expect that work outlined in this dissertation may serve as a guide for all future analysis.

REFERENCES

1. Patterson SD, Aebersold RH. Proteomics: the first decade and beyond. *Nat Genet* 2003;33 Suppl:311-23.
2. Tyers M, Mann M. From genomics to proteomics. *Nature* 2003;422:193-7.
3. Klose J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* 1975;26:231-43.
4. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 1975;250:4007-21.
5. Aebersold R. A mass spectrometric journey into protein and proteome research. *J Am Soc Mass Spectrom* 2003;14:685-95.
6. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* 2000;97:9390-5.
7. Unlu M, Morgan ME, Minden JS. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* 1997;18:2071-7.
8. Tonge R, Shaw J, Middleton B, et al. Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics* 2001;1:377-96.
9. Edman P. A method for the determination of amino acid sequence in peptides. *Arch Biochem* 1949;22:475.
10. Edman P, Begg G. A protein sequenator. *Eur J Biochem* 1967;1:80-91.
11. Simpson RJ. *Proteins and Proteomics*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2003.
12. Abderhalden E, Brockmann H. The contribution determining the composition of proteins especially polypeptides (German). *Biochem Z* 1930;225:386-408.
13. Barber M, Bordoli RS, Sedgwick RD, Tyler AN, Bycroft BW. Fast atom bombardment mass spectrometry of bleomycin A2 and B2 and their metal complexes. *Biochem Biophys Res Commun* 1981;101:632-8.

14. Morris HR, Panico M, Barber M, et al. Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis. *Biochem Biophys Res Commun* 1981;101:623-31.
15. Hunt DF, Buko AM, Ballard JM, Shabanowitz J, Giordani AB. Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomed Mass Spectrom* 1981;8:397-408.
16. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 1989;246:64-71.
17. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 1988;60:2299-301.
18. Baumann S, Ceglarek U, Fiedler GM, et al. Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin Chem* 2005;51:973-80.
19. Suckau D, Resemann A, Schuerenberg M, et al. A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Anal Bioanal Chem* 2003;376:952-65.
20. Wilm M, Shevchenko A, Houthaeve T, et al. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* 1996;379:466-9.
21. Macht M, Asperger A, Deininger SO. Comparison of laser-induced dissociation and high-energy collision-induced dissociation using matrix-assisted laser desorption/ionization tandem time-of-flight (MALDI-TOF/TOF) for peptide and protein identification. *Rapid Commun Mass Spectrom* 2004;18:2093-105.
22. Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 1995;67:1426-36.
23. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994;66:4390-9.
24. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001;291:1304-51.
25. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.

26. Yates JR, 3rd, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* 1995;67:3202-10.
27. Griffin PR, MacCoss MJ, Eng JK, et al. Direct database searching with MALDI-PSD spectra of peptides. *Rapid Commun Mass Spectrom* 1995;9:1546-51.
28. Yates JR, Eng JK, Clauser KR, Burlingame AL. Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *J Am Soc Mass Spectrom* 1996;7:1089-98.
29. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551-67.
30. Pappin DJ, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 1993;3:327-32.
31. Massolini G, Calleri E. Immobilized trypsin systems coupled on-line to separation methods: recent developments and analytical applications. *J Sep Sci* 2005;28:7-21.
32. Zhang K, Wu S, Tang X, Kaiser NK, Bruce JE. A bifunctional monolithic column for combined protein preconcentration and digestion for high throughput proteomics research. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007;849:223-30.
33. Colangelo J, Orlando R. On-target exoglycosidase digestions/MALDI-MS for determining the primary structures of carbohydrate chains. *Anal Chem* 1999;71:1479-82.
34. Harris WA, Reilly JP. On-probe digestion of bacterial proteins for MALDI-MS. *Anal Chem* 2002;74:4410-6.
35. Ibanez AJ, Muck A, Halim V, Svatos A. Trypsin-linked copolymer MALDI chips for fast protein identification. *J Proteome Res* 2007;6:1183-9.
36. Jmeian Y, El Rassi Z. Tandem affinity monolithic microcolumns with immobilized protein A, protein G', and antibodies for depletion of high abundance proteins from serum samples: integrated microcolumn-based fluidic system for simultaneous depletion and tryptic digestion. *J Proteome Res* 2007;6:947-54.
37. Guiochon G. Monolithic columns in high-performance liquid chromatography. *J Chromatogr A* 2007;1168:101-68.

38. Tholey A. MALDI Mass Spectrometry for Quantitative Proteomics – Approaches, Scopes and Limitations. *In: Dagstuhl Seminar Proceedings: Computational Proteomics, Saarland, Germany, 2006.*
39. Gygi SP, Rist B, Gerber SA, et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999;17:994-9.
40. Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* 2005;5:4-15.
41. Ross PL, Huang YN, Marchese JN, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004;3:1154-69.
42. Comuzzi B, Sadar MD. Proteomic analyses to identify novel therapeutic targets for the treatment of advanced prostate cancer. *Cellscience* 2006;3:61-81.
43. Semmes OJ, Malik G, Ward M. Application of mass spectrometry to the discovery of biomarkers for detection of prostate cancer. *J Cell Biochem* 2006;98:496-503.
44. Wright GL, Jr. SELDI proteinchip MS: a platform for biomarker discovery and cancer diagnosis. *Expert Rev Mol Diagn* 2002;2:549-63.
45. Conrads TP, Hood BL, Issaq HJ, Veenstra TD. Proteomic patterns as a diagnostic tool for early-stage cancer: a review of its progress to a clinically relevant tool. *Mol Diagn* 2004;8:77-85.
46. Malik G, Ward MD, Gupta SK, et al. Serum levels of an isoform of apolipoprotein A-II as a potential marker for prostate cancer. *Clin Cancer Res* 2005;11:1073-85.
47. Semmes OJ. The "omics" haystack: defining sources of sample bias in expression profiling. *Clin Chem* 2005;51:1571-2.
48. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 2004;20:777-85.
49. Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. *J Natl Cancer Inst* 2005;97:315-9.
50. Semmes OJ, Feng Z, Adam BL, et al. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clin Chem* 2005;51:102-12.

51. Merkle CJ, Loescher LJ. Cancer nursing: Principles and practices, 6th edition, p. 3-25. Boston: Jones and Bartlett; 2005.
52. Franks LM, Knowles MA. Introduction to the Cellular and Molecular Biology of Cancer, 4th edition, p. 1-24. Leeds: Oxford University Press; 2005.
53. Bertucci F, Houlgatte R, Nguyen C, et al. Gene expression profiling of cancer by use of DNA arrays: how far from the clinic? *Lancet Oncol* 2001;2:674-82.
54. Bertucci F, Birnbaum D, Goncalves A. Proteomics of breast cancer: principles and potential clinical applications. *Mol Cell Proteomics* 2006;5:1772-86.
55. Hoffman SA, Joo WA, Echan LA, Speicher DW. Higher dimensional (Hi-D) separation strategies dramatically improve the potential for cancer biomarker detection in serum and plasma. *J Chromatogr B Analyt Technol Biomed Life Sci* 2006.
56. Petricoin EF, Belluco C, Araujo RP, Liotta LA. The blood peptidome: a higher dimension of information content for cancer biomarker discovery. *Nat Rev Cancer* 2006;6:961-7.
57. Aebersold R, Anderson L, Caprioli R, et al. Perspective: a program to improve protein biomarker discovery for cancer. *J Proteome Res* 2005;4:1104-9.
58. Villanueva J, Shaffer DR, Philip J, et al. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest* 2006;116:271-84.
59. Banks RE, Stanley AJ, Cairns DA, et al. Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. *Clin Chem* 2005;51:1637-49.
60. Drake RR, Cazares LH, Malik G, et al. Quality control, preparation and protein stability issues for blood serum and plasma used in biomarker discovery and proteomic profiling assays. *Bioprocessing J* 2004;3:45-50.
61. Liotta LA, Petricoin EF. Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. *J Clin Invest* 2006;116:26-30.
62. Villanueva J, Philip J, Entenberg D, et al. Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal Chem* 2004;76:1560-70.
63. Zolg W. The proteomic search for diagnostic biomarkers: lost in translation? *Mol Cell Proteomics* 2006;5:1720-6.

64. Tirumalai RS, Chan KC, Prieto DA, et al. Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* 2003;2:1096-103.
65. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002;1:845-67.
66. Granger J, Siddiqui J, Copeland S, Remick D. Albumin depletion of human plasma also removes low abundance proteins including the cytokines. *Proteomics* 2005;5:4713-8.
67. Tang N, Tornatore P, Weinberger SR. Current developments in SELDI affinity technology. *Mass Spectrom Rev* 2004;23:34-44.
68. Simpkins F, Czechowicz JA, Liotta L, Kohn EC. SELDI-TOF mass spectrometry for cancer biomarker discovery and serum proteomic diagnostics. *Pharmacogenomics* 2005;6:647-53.
69. Aivado M, Spentzos D, Alterovitz G, et al. Optimization and evaluation of surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) with reversed-phase protein arrays for protein profiling. *Clin Chem Lab Med* 2005;43:133-40.
70. Hortin GL. The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. *Clin Chem* 2006;52:1223-37.
71. Drake RR, Schwegler EE, Malik G, et al. Lectin capture strategies combined with mass spectrometry for the discovery of serum glycoprotein biomarkers. *Mol Cell Proteomics* 2006;5:1957-67.
72. Pusch W, Kostrzewa M. Application of MALDI-TOF mass spectrometry in screening and diagnostic research. *Curr Pharm Des* 2005;11:2577-91.
73. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2007. *CA Cancer J Clin* 2007;57:43-66.
74. Thibodeau GA, Patton KT. *Anatomy and Physiology*, 6th edition, p. 1139-40. St. Louis: Mosby; 2007.
75. Russo J, Hu YF, Yang X, Russo IH. Developmental, cellular, and molecular basis of human breast cancer. *J Natl Cancer Inst Monogr* 2000:17-37.
76. Hilakivi-Clarke L, de Assis S. Fetal origins of breast cancer. *Trends Endocrinol Metab* 2006;17:340-8.
77. Ely S, Vioral AN. Breast cancer overview. *Plast Surg Nurs* 2007;27:128-33; quiz 34-5.

78. Kufe DW, Pollock; RE, Weichselbaum RR, et al. *Cancer Medicine*, 6th edition. Hamilton: BC Decker Inc; 2003.
79. Li CI, Malone KE, Daling JR. Differences in breast cancer stage, treatment, and survival by race and ethnicity. *Arch Intern Med* 2003;163:49-56.
80. Kelsey JL. A review of the epidemiology of human breast cancer. *Epidemiol Rev* 1979;1:74-109.
81. Sakorafas GH, Krespis E, Pavlakis G. Risk estimation for breast cancer development; a clinical perspective. *Surg Oncol* 2002;10:183-92.
82. Vogel VG. Breast cancer prevention: a review of current evidence. *CA Cancer J Clin* 2000;50:156-70.
83. Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990;250:1684-9.
84. Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* 1994;265:2088-90.
85. Arnold MA, Goggins M. BRCA2 and predisposition to pancreatic and other cancers. *Expert Rev Mol Med* 2001;2001:1-10.
86. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* 2001;93:358-66.
87. Rockhill B, Byrne C, Rosner B, Louie MM, Colditz G. Breast cancer risk prediction with a log-incidence model: evaluation of accuracy. *J Clin Epidemiol* 2003;56:856-61.
88. Berry DA, Iversen ES, Jr., Gudbjartsson DF, et al. BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *J Clin Oncol* 2002;20:2701-12.
89. Ozanne EM, Klemp JR, Esserman LJ. Breast cancer risk assessment and prevention: a framework for shared decision-making consultations. *Breast J* 2006;12:103-13.
90. Rebbeck TR, Levin AM, Eisen A, et al. Breast cancer risk after bilateral prophylactic oophorectomy in BRCA1 mutation carriers. *J Natl Cancer Inst* 1999;91:1475-9.

91. Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol* 2007;25:5287-312.
92. Duffy MJ, Shering S, Sherry F, McDermott E, O'Higgins N. CA 15-3: a prognostic marker in breast cancer. *Int J Biol Markers* 2000;15:330-3.
93. King CR, Kraus MH, Aaronson SA. Amplification of a novel v-erbB-related gene in a human mammary carcinoma. *Science* 1985;229:974-6.
94. Slamon DJ, Clark GM, Wong SG, et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987;235:177-82.
95. Leitzel K, Teramoto Y, Konrad K, et al. Elevated serum c-erbB-2 antigen levels and decreased response to hormone therapy of breast cancer. *J Clin Oncol* 1995;13:1129-35.
96. Colomer R, Montero S, Lluch A, et al. Circulating HER2 extracellular domain and resistance to chemotherapy in advanced breast cancer. *Clin Cancer Res* 2000;6:2356-62.
97. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817-26.
98. Routh JC, Leibovich BC. Adenocarcinoma of the prostate: epidemiological trends, screening, diagnosis, and surgical management of localized disease. *Mayo Clin Proc* 2005;80:899-907.
99. Kutikov A, Guzzo TJ, Malkowicz SB. Clinical approach to the prostate: an update. *Radiol Clin North Am* 2006;44:649-63, vii.
100. Ward JF, Slezak JM, Blute ML, Bergstralh EJ, Zincke H. Radical prostatectomy for clinically advanced (cT3) prostate cancer since the advent of prostate-specific antigen testing: 15-year outcome. *BJU Int* 2005;95:751-6.
101. Keating NL, O'Malley AJ, Smith MR. Diabetes and cardiovascular disease during androgen deprivation therapy for prostate cancer. *J Clin Oncol* 2006;24:4448-56.
102. Wang MC, Papsidero LD, Kuriyama M, et al. Prostate antigen: a new potential marker for prostatic cancer. *Prostate* 1981;2:89-96.
103. Bradford TJ, Tomlins SA, Wang X, Chinnaiyan AM. Molecular markers of prostate cancer. *Urol Oncol* 2006;24:538-51.

104. Freedland SJ, Partin AW. Prostate-specific antigen: update 2006. *Urology* 2006;67:458-60.
105. Delongchamps NB, Singh A, Haas GP. The role of prevalence in the diagnosis of prostate cancer. *Cancer Control* 2006;13:158-68.
106. Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤ 4.0 ng per milliliter. *N Engl J Med* 2004;350:2239-46.
107. Thompson IM, Ankerst DP, Chi C, et al. Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *J Natl Cancer Inst* 2006;98:529-34.
108. Bracarda S, de Cobelli O, Greco C, et al. Cancer of the prostate. *Crit Rev Oncol Hematol* 2005;56:379-96.
109. Peter J, Unverzagt C, Krogh TN, Vorm O, Hoesel W. Identification of precursor forms of free prostate-specific antigen in serum of prostate cancer patients by immunosorption and mass spectrometry. *Cancer Res* 2001;61:957-62.
110. Peyromaure M, Fulla Y, Debre B, Dinh-Xuan AT. Pro PSA: a "pro cancer" form of PSA? *Med Hypotheses* 2005;64:92-5.
111. Mikolajczyk SD, Catalona WJ, Evans CL, et al. Proenzyme forms of prostate-specific antigen in serum improve the detection of prostate cancer. *Clin Chem* 2004;50:1017-25.
112. Sokoll LJ, Wang Y, Feng Z, et al. [-2]proPSA Improves Prostate Cancer Detection: an EDRN Validation Study. *In: Early Detection Research Network (EDRN) 5th Scientific Workshop, Bethesda, MD, 2008.*
113. Stephan C, Jung K, Lein M, Diamandis EP. PSA and other tissue kallikreins for prostate cancer detection. *Eur J Cancer* 2007;43:1918-26.
114. Lester J. Breast cancer in 2007: incidence, risk assessment, and risk reduction strategies. *Clin J Oncol Nurs* 2007;11:619-22.
115. West-Norager M, Kelstrup CD, Schou C, et al. Unravelling in vitro variables of major importance for the outcome of mass spectrometry-based serum proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007;847:30-7.
116. Hsieh SY, Chen RK, Pan YH, Lee HL. Systematical evaluation of the effects of sample collection procedures on low-molecular-weight serum/plasma proteome profiling. *Proteomics* 2006;6:3189-98.

117. Tibshirani R, Hastie T, Narasimhan B, et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* 2004;20:3034-44.
118. Franks F. *Biophysics and biochemistry at low temperature*. London: Cambridge University Press; 1985.
119. McLerran D, Grizzle WE, Feng Z, et al. Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias. *Clin Chem* 2008;54:44-52.
120. Pusztai L, Mazouni C, Anderson K, Wu Y, Symmans WF. Molecular classification of breast cancer: limitations and potential. *Oncologist* 2006;11:868-77.
121. Schuurman AG, van den Brandt PA, Goldbohm RA. Exogenous hormone use and the risk of postmenopausal breast cancer: results from The Netherlands Cohort Study. *Cancer Causes Control* 1995;6:416-24.
122. Espinosa E, Redondo A, Vara JA, et al. High-throughput techniques in breast cancer: a clinical perspective. *Eur J Cancer* 2006;42:598-607.
123. Paulson JC, Blixt O, Collins BE. Sweet spots in functional glycomics. *Nat Chem Biol* 2006;2:238-48.
124. Ohtsubo K, Marth JD. Glycosylation in cellular mechanisms of health and disease. *Cell* 2006;126:855-67.
125. Drake RR, Cazares, L.H. and Semmes, O.J. Mining the Low Molecular Weight Proteome of Blood for Disease Specific Biomarkers. *Proteomics* 2007;In press.
126. Fung ET, Yip TT, Lomas L, et al. Classification of cancer types by measuring variants of host response proteins using SELDI serum assays. *Int J Cancer* 2005;115:783-9.
127. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003;4:24.
128. de Noo ME, Tollenaar RA, Ozalp A, et al. Reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectrometry. *Anal Chem* 2005;77:7232-41.
129. Shin S, Cazares L, Schneider H, et al. Serum biomarkers to differentiate benign and malignant mammographic lesions. *J Am Coll Surg* 2007;204:1065-71.

130. Boja ES, Fales HM. Overalkylation of a protein digest with iodoacetamide. *Anal Chem* 2001;73:3576-82.
131. Sparbier K, Wenzel T, Kostrzewa M. Exploring the binding profiles of ConA, boronic acid and WGA by MALDI-TOF/TOF MS and magnetic particles. *J Chromatogr B Analyt Technol Biomed Life Sci* 2006;840:29-36.
132. Sechi S. *Quantitative Proteomics by Mass Spectrometry*, Vol. 359, p. 17-35. Totowa: Humana Press; 2007.
133. Zhang K, Wu S, Tang X, Kaiser NK, Bruce JE. A bifunctional monolithic column for combined protein preconcentration and digestion for high throughput proteomics research. *J Chromatogr B Analyt Technol Biomed Life Sci* 2006.
134. Gaevskaia VA, Azhitskii G. [Isoelectric fractions of healthy human serum albumin and their ability to bind bilirubin]. *Ukr Biokhim Zh* 1978;50:735-8.
135. Brandt E, Petersen F, Ludwig A, et al. The beta-thromboglobulins and platelet factor 4: blood platelet-derived CXC chemokines with divergent roles in early neutrophil regulation. *J Leukoc Biol* 2000;67:471-8.
136. Lowenthal MS, Mehta AI, Frogale K, et al. Analysis of albumin-associated peptides and proteins from ovarian cancer patients. *Clin Chem* 2005;51:1933-45.
137. Wang H, Zhang M, Bianchi M, et al. Fetuin (alpha2-HS-glycoprotein) opsonizes cationic macrophage-deactivating molecules. *Proc Natl Acad Sci U S A* 1998;95:14429-34.
138. Kundranda MN, Henderson M, Carter KJ, et al. The serum glycoprotein fetuin-A promotes Lewis lung carcinoma tumorigenesis via adhesive-dependent and adhesive-independent mechanisms. *Cancer Res* 2005;65:499-506.
139. Macintyre E, Willerford D, Morris SW. Non-Hodgkin's Lymphoma: Molecular Features of B Cell Lymphoma. *Hematology Am Soc Hematol Educ Program* 2000:180-204.
140. Sampietro J, Dahlberg CL, Cho US, et al. Crystal structure of a beta-catenin/BCL9/Tcf4 complex. *Mol Cell* 2006;24:293-300.
141. Fuerer C, Homicsko K, Lukashev AN, Pittet AL, Iggo RD. Fusion of the BCL9 HD2 domain to E1A increases the cytopathic effect of an oncolytic adenovirus that targets colon cancer cells. *BMC Cancer* 2006;6:236.
142. Fodde R, Brabletz T. Wnt/beta-catenin signaling in cancer stemness and malignant behavior. *Curr Opin Cell Biol* 2007;19:150-8.

143. Brown AM. Wnt signaling in breast cancer: have we come full circle? *Breast Cancer Res* 2001;3:351-5.
144. Lin SY, Xia W, Wang JC, et al. Beta-catenin, a novel prognostic marker for breast cancer: its roles in cyclin D1 expression and cancer progression. *Proc Natl Acad Sci U S A* 2000;97:4262-6.
145. Green PH, Glickman RM, Riley JW, Quinet E. Human apolipoprotein A-IV. Intestinal origin and distribution in plasma. *J Clin Invest* 1980;65:911-9.
146. Khovidhunkit W, Duchateau PN, Medzihradzky KF, et al. Apolipoproteins A-IV and A-V are acute-phase proteins in mouse HDL. *Atherosclerosis* 2004;176:37-44.
147. Nishimukai M, Hara H. Enteral administration of soybean phosphatidylcholine enhances the lymphatic absorption of lycopene, but reduces that of alpha-tocopherol in rats. *J Nutr* 2004;134:1862-6.
148. Barber NJ, Barber J. Lycopene and prostate cancer. *Prostate Cancer Prostatic Dis* 2002;5:6-12.
149. Liu X, Allen JD, Arnold JT, Blackman MR. Lycopene inhibits IGF-I signal transduction and growth in normal prostate epithelial cells by decreasing DHT-modulated IGF-I production in cocultured reactive stromal cells. *Carcinogenesis* 2008.
150. Peters U, Leitzmann MF, Chatterjee N, et al. Serum lycopene, other carotenoids, and prostate cancer risk: a nested case-control study in the prostate, lung, colorectal, and ovarian cancer screening trial. *Cancer Epidemiol Biomarkers Prev* 2007;16:962-8.
151. Lu QY, Hung JC, Heber D, et al. Inverse associations between plasma lycopene and other carotenoids and prostate cancer. *Cancer Epidemiol Biomarkers Prev* 2001;10:749-56.
152. Almushatat AS, Talwar D, McArdle PA, et al. Vitamin antioxidants, lipid peroxidation and the systemic inflammatory response in patients with prostate cancer. *Int J Cancer* 2006;118:1051-3.
153. Qin X, Swertfeger DK, Zheng S, Hui DY, Tso P. Apolipoprotein AIV: a potent endogenous inhibitor of lipid oxidation. *Am J Physiol* 1998;274:H1836-40.
154. Guo YL, Colman RW. Two faces of high-molecular-weight kininogen (HK) in angiogenesis: bradykinin turns it on and cleaved HK (HKa) turns it off. *J Thromb Haemost* 2005;3:670-6.

155. Schmaier AH. The kallikrein-kinin and the renin-angiotensin systems have a multilayered interaction. *Am J Physiol Regul Integr Comp Physiol* 2003;285:R1-13.
156. Daly ME, Makris A, Reed M, Lewis CE. Hemostatic regulators of tumor angiogenesis: a source of antiangiogenic agents for cancer treatment? *J Natl Cancer Inst* 2003;95:1660-73.
157. Drake RR, Cazares L, Semmes OJ. Mining the low molecular weight proteome of blood. *Proteomics Clin. Appl.* 2007;1:758-68.
158. Peracaula R, Tabares G, Royle L, et al. Altered glycosylation pattern allows the distinction between prostate-specific antigen (PSA) from normal and tumor origins. *Glycobiology* 2003;13:457-70.
159. Tabares G, Radcliffe CM, Barrabes S, et al. Different glycan structures in prostate-specific antigen from prostate cancer sera in relation to seminal plasma PSA. *Glycobiology* 2006;16:132-45.
160. Zhang H, Li XJ, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* 2003;21:660-6.
161. Zhang H, Liu AY, Loriaux P, et al. Mass spectrometric detection of tissue proteins in plasma. *Mol Cell Proteomics* 2007;6:64-71.
162. Faca V, Coram M, Phanstiel D, et al. Quantitative analysis of acrylamide labeled serum proteins by LC-MS/MS. *J Proteome Res* 2006;5:2009-18.
163. Cox DM, Zhong F, Du M, et al. Multiple reaction monitoring as a method for identifying protein posttranslational modifications. *J Biomol Tech* 2005;16:83-90.

VITA

IZABELA DEBKIEWICZ KARBASSI

Eastern Virginia Medical School

Department of Microbiology and Molecular Cell Biology

700 W. Olney Road

Norfolk, VA 23507

Old Dominion University

Hampton Boulevard

Norfolk, VA 23529

Education**Master of Science in Microbiology**

University of Rochester School of Medicine and Dentistry Rochester, NY (2001-2003)

M.S. Thesis: "*Protein-protein interaction studies of HPV-16 E6*"

Bachelor of Arts in Biochemistry and Molecular Biology

Boston University, Boston, MA (1997-2001)

Minor: Business Administration