BIOINFORMATICS ANALYSIS OF HUMORAL IMMUNE RESPONSE AND PROTEIN MICROARRAY FOR BIOMARKER DISCOVERY

by

Huy Q. Vuong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2011

Doctoral Committee:

Professor David M. Lubman, Co-Chair
Associate Professor Kerby Shedden, Co-Chair
Professor Megan Lim
Associate Professor Alexey Nesvizhskii
Associate Professor Subramaniam Pennathur

DEDICATION

To my parents Ha Vuong and ThuThuy Nguyen

ACKNOWLEDGMENTS

From the bottom of my heart, I owe my deepest gratitude to my advisor, Dr David M. Lubman, one of the most wonderful and smartest scientists I know. His generosities in providing constant support, encouragement, invaluable advice and infinite patience have guided me through my graduate studies. Without him, this dissertation would not have been possible and I would not become a better scientist today. I would also like to sincerely thank my committee members, Dr Kerby Shedden, Dr Megan Lim, Dr Alexey Nesvizhskii and Dr Subramaniam Pennathur, for being an active part of my committee as well as providing me with continuous encouragement and expert advice on my work. I am especially grateful to Dr Shedden for his inspiration and teaching of statistical concepts that have been instrumental in my PhD study. I also thank Dr Lim, Dr Nesvizhskii and Dr Pennathur for the fruitful discussion and expert advice in proteomics that enables me to complete my dissertation and to develop an understanding of the subjects.

I would also like to thank Professor Daniel Burns, Professor Margit Burmeister, Professor Gil Omenn, and Professor Brian Athey whose leadership roles in Bioinformatics Program could not keep them from generously giving me valuable advice about academic life. I also like to thank Julia Eussen, Alex Terzian and all support staff in the Bioinformatics Program who have helped me through various administrative

processes with professional disposition. I am heartily thankful to my friends, my colleagues in the Bioinformatics Program, especially Junguk Hur, Conner Sandefur, Arun Manoharan and all other members who have provided me with valuable experiences both within and outside of work.

I would like to thank all past and present Lubman group-members for all of their help. It is a pleasure to work in a dynamic team of post-docs and graduate students who provide me with continuous support and invaluable discussion. Special thanks to Dr. Evelyn Kim, Dr.Yanfei Wang, Dr. Tasneem Patwa, Dr Shun Feng, Dr Xiaoping Ao for guidance during my graduate studies and collaborations. I would also like to thank Dr. Dr. Xiaolei Xie, Dr. Chen Li, Dr. Yashu Liu and Dr. Jintang He for their support and collaborations.

I would like to make a special reference to my parents, Ha Vuong and ThuThuy Nguyen and my brother Quyen Vuong for their love and encouragement during my academic life. The completion of my graduate career would not have been possible without their sacrifices and inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FDR	False discovery rate
IPA	Ingenuity pathways analysis
NL	Neutral loss
TiO_2	Titanium dioxide
TPP	Trans-Proteomic Pipeline
ZrO_2	Zirconium dioxide
PTM	Post translational modification
RP-HPLC	Reverse phase high-performance liquid chromatography
COPA	Cancer outlier profile analysis
OS	Outlier sums
ORT	Outlier robust t-test
MOST	Maximum ordered subset t-statistics
LSOSS	Least sum of ordered subset square t-statistic
PDAC	Pancreatic ductal adenocarcinoma
CF	Chromatofocusing
ID	Inner diameter

ABSTRACT

Early detection is the best defense which could significantly improve the cancer survival rate in several cancers including melanoma and pancreatic cancer. A promising approach to discover biomarkers for early detection involves using the humoral immune response against tumor proteins. Together with advances in proteomics, in particular a high throughput protein microarray platform, humoral immune response studies have enabled a breakthrough in developing global screening of the highly complex plasma proteome for biomarkers for early detection. In this dissertation, we attempt to integrate proteomics and bioinformatics approaches to analyze signals from protein microarray data for the reliable identification of differentially expressed proteins under different biological conditions. First, we present a study comparing outlier-based to traditional mean-based approaches in differential expression analysis with applications in protein microarray data in heterogeneous diseases. Our investigation uses a glycoproteomics dataset from a melanoma study, an original simulation-based approach to benchmarking, and a new data visualization technique to assess the potential for methods that explicitly target heterogeneous patterns of differential expression to give improved performance relative to traditional approaches based on group-wise comparison of means. Results include identifications of 1 significant feature using outlier statistics and 15 significant features using t-statistics from a melanoma dataset of 43 samples and 47 features. Next, we apply the outlier strategy to a protein microarray dataset from a pancreatic cancer

study with sera from 37 pancreatic cancers, 24 chronic pancreatitis and 23 normal to identify protein biomarkers that are differentially expressed in only a subset of cancer samples. Three protein markers exhibiting outlier patterns exclusive to cancer sera and no outliers in normal sera are identified by mass spectrometry and confirmed by a follow-up study with an independent dataset. The next study presents the application of a label-free quantitation approach for measuring changes in protein abundance level associated with phosphorylation, a major mechanism of tumorigenesis. Results include identifications of differentially expressed post-translational modified proteins such as phosphorylated lamin-A/C, isoforms-A and GTPase-activating protein binding protein-1 in pancreatic cancer. Together, this dissertation contributes two approaches to biomarker discovery using protein microarrays and the humoral immune response and LC-MS/MS and label-free quantitation.

CHAPTER 1

INTRODUCTION

1.1 CANCER BIOMARKERS

Cancer is responsible for 25% of all deaths in the United States (US) and is second only to cardiovascular disease as the leading cause of death. One in three people in the US will develop cancer during their lifetimes[1]. These noteworthy cancer statistics are taken from the American Cancer Society (ACS) which recommends prevention and early detection as the best weapons against cancer. The crucial importance of early detection cannot be stressed enough, exemplified by a 90/10 survival curve often observed in ovarian cancer[2]. If ovarian tumors are discovered in early stage, the survival rate after diagnosis is almost 90% by means of surgery. Meanwhile, the survival rate sharply drops to 10% if the tumors are discovered in later stages when the cancers are well established and already metastasizing. In fact, it is a major goal of the National Cancer Institute (NCI) in initiating the Early Detection Research Network (EDRN) to bring together dozens of institutions to help evaluate new ways of testing cancer in its earliest stages[3].

For years, scientists have turned to tumor markers, which are substances found in abnormal amounts in the blood, urine or tissues of cancer patients, as the key to early detection, diagnosis and management of several types of cancers. The most well studied tumor marker is the p53 protein whose detection in serum correlates strongly with cancer as demonstrated by DeLeo et al. [4]. Other tumor markers such as NY-ESO-1[5],

her2/neu[6] and TA90[7] have been shown to reflect the stage (extent) of neuroblastoma, breast AND ovarian and melanoma, respectively. Tumor marker levels may be measured before treatment to help physicians estimate the stage of diseases and plan the appropriate therapy or after treatment to assess the patient response to therapy. To date, scientists have identified a number of tumor markers for several types of cancer with the help of recent advances in proteomics technologies[8]. For example, a novel proteomics technology called Nucleic Acid Protein Programmable Array (NAPPA) has recently been employed for identifying a potential panel of 28 autoantibody biomarkers for the early detection of breast cancer[9]. Although tumor markers for every type of cancer are not yet identified, five clinical trials have moved forward to test and validate biomarkers for early detection of liver, bladder, lung, mesothelioma and prostate cancer as a result of EDRN's work[10].

Body fluids such as sera and plasma are rich sources of cancer biomarkers since cancer cells are known to secrete tumor specific proteins into the blood. Several protein based cancer biomarkers are detected from sera or plasma of cancer patients. Unfortunately, those biomarkers are usually low abundance proteins whose concentration is at the lower end of the abundance range of known plasma proteins (spanning nine orders of magnitude)[11]. Moreover, blood serum is a complex system where the top three major protein constituents of serum including albumin, immunoglobulins and fibrinogens make up 82% of the total serum proteins, leaving only 1% to be regulatory proteins that could be candidates for cancer biomarkers [12]. Therefore, mining the plasma proteome for biomarkers is still a grand challenge in cancer research [13].

1.2 HUMORAL IMMUNE RESPONSE

Human immunity is defined by complex system of biological structures and processes designed to prevent diseases by detecting and killing pathogens and tumor cells within the body[14]. There are two arms of the immune system, innate immune and adaptive immune, and each arm has both humoral and cellular components. The humoral immune response is so named because it involves substances called antibodies found in the humors (body fluid). These antibodies are produced by white blood cells called B cells. When encountering foreign pathogens, B cells differentiate into plasma cells which can secrete antibodies to bind and destroy the pathogens. Antibodies are an important part of immunotherapy as a preferred treatment to certain cancers when standard cancer therapies, including chemotherapy and radiotherapy, are not effective. Antibody based vaccines have been successfully used in clinical studies for treatment of melanomas [15] and other diseases such as HIV[16]. Moreover, antibodies are an important class of biomarker for cancers, autoimmune and other diseases[17]. They are readily accessible in biological fluids such as serum, plasma, tears and saliva. They are remarkably stable with long half-lives in complex solution, such as human serum or plasma and are detectable by simple and rapid assays. More importantly, antibodies usually appear as soon as the antigens are present and because of the high specificity between antibody-antigen recognition, autoantibodies biomarkers can serve as reliable predictors of cancers, autoimmune and other diseases.

Not only does the humoral immune response protect our body from pathogens, it also guards against other diseases such as cancer. The connection between humoral immune response and cancer has a long and complex history beginning in 1909 with a

German immunologist Paul Ehrlich who won the Nobel Prize in physiology. He was the first scientist to propose that the human immune system can identify and eliminate nascent tumor cells. About 50 years later, Lewis Thomas and Frank MacFarlane Burnet continued Ehrlich's work and coined the term "immunosurveillance," which described the human immune system as a continuous monitoring process to alert against transformed tumor cells. However, the concept remained controversial due to the lack of supporting experimental evidence until a publication in Nature by Robert Schreiber and Lloyd Old in 2001 entitled "IFN-gamma and lymphocytes prevent primary tumor development and shape tumor immunogenicity"[18]. Using genetically engineered mice, Schreiber and Old unambiguously proved that the immune response can not only prevent tumors from developing but also plays a role in changing tumor behavior. Consequently, they came up with the concept of "immunoediting" which included this new role of immune response and in effect extended the concept of "immunosurveillance". Similar studies led to significant advances in the field of tumor immunotherapy by taking advantage of the immune response in cancer detection and treatment.

Ample evidence exists for tumor antigens eliciting humoral immunity for cancers at a very early stage when there are still no recognizable symptoms. Identification of tumor antigens that are expressed at the early stage of cancer and could drive the humoral response, therefore, leads to a simple non-invasive detection strategy [19]. Many cancer biomarker studies have identified a variety of tumor antigens in the serum of cancer patients that binds to specific autoantibodies and triggers the humoral response [4, 8, 20-21]. A number of these tumor antigens are being validated in clinical settings and some are promising candidates for antibody based cancer vaccines[9]. Examples of popular

identified antigens are Cyclin B1, LAMR1 and p53 [19]. The greatest success so far is the identification of antibody based humoral immunity targeting cancer stem-like cells and tumor initiating cells in vitro, based on which antibody therapies are also efficient in vivo [20].

The humoral immune response is a promising approach to fight against cancer because of several remarkable characteristics of antibodies which make them wonderful diagnostic and prognostic markers of disease [22]. Antibodies usually recognize tumor antigens as soon as they are present; are easily detected in human body fluids and persist a long time after infection or immunization. Diagnosis of cancer at an early stage is the most effective solution to improve the patient survival rate. However, early stage tumors are usually small and low abundance in protein concentration, making it difficult to detect by traditional approaches (e.g. using mass spectrometry). The advantage of using humoral response is that once there is a humoral immunity that can recognize the tumor antigens, even a limited amount of antigen can trigger a biological amplification of associated antibodies, which can be easily detected. In addition, humoral immune response provides not only an active defense against infection, but also a passive memory to store the information about a pathogen.

Harnessing the humoral immune response to tumor antigen is still in the early stage of development. An autoantibody test with high sensitivity and high specificity in directing against certain cancer types is in development. Thus far, applications of humoral immune response are primarily targeted to distinguish early stage from advanced stage cancer. The potential of humoral immune response to cancer is large, however, and may include: identification of cancer biomarkers for diagnosis, classification and

monitoring of response, determination of the impact of response on cancer progression and elucidation of mechanisms for immunotherapy development [23].

Several challenges must be overcome to explore the full potential of humoral immune response as a detection method for early cancers. One set of challenges arises from the inherent complexity of the cancer proteome. Proteins in the cancer proteome have a vast dynamic range, exceeding 10^{10} orders of magnitude, coupled with a plethora of isoforms and disease heterogeneity [13]. Moreover, many reliable cancer biomarkers are low abundance proteins and their concentrations are usually elevated only in advanced forms of disease [8]. Given that the humoral immune response is heterogeneous, in part because two individuals with the same cancer types may generate humoral immune responses to two different antigens, a panel of tumor associated antigens is needed for a robust test. Thus, validation of biomarkers discovered from several independent studies is critical for the selection of the best combination of markers [13]. The challenge that remains is how to reliably use autoantibodies for detecting cancer, especially in early stage when clinical appearance has not occurred [23].

1.3 PROTEOMICS APPROACHES IN HUMORAL IMMUNE RESPONSE

Despite recent advances in proteomics and microarray technology, the application of cancer proteomics to tumor immunology is still in its infancy due to limited understanding of the complex array of interaction involved in the immune response to cancer. Promising proteomics techniques have allowed the identification of tumor antigens and antibodies on a proteome-wide basis, such as cell fractionation, protein microarrays, nanospray mass spectrometry, and protein expression profiling of tumors in

vivo [23]. The three mainstream proteomics technologies being used to study humoral immune response in cancer are serological expression cloning (SEREX), serological proteome analysis (SERPA) and protein microarrays[24].

One of the first to adopt serological analysis to study humoral immune response, serological expression cloning (SEREX) uses cancer patient sera as probes against tumor antigens derived from tumor cells or tumor cell lysates to discover novel autoantigens. SEREX is unique in that it uses patients' autologous tumors as mRNA sources and a prokaryotic system (e.g. E.coli) for the expression of recombinant proteins to identify antigens that react with autoantibodies present in the autologous patient serum. It has been used to successfully identify novel tumor biomarkers in several cancers including NY-ESO-1 protein in squamous cell carcinoma[25], CAGE-1 in lung cancer[26] and GLEA1, GLEA2 and PHF3 in glioblastoma[27]. Many of these tumor biomarkers are documented in the Cancer Immunome Database[28] at http://ludwig-sun5.unil.ch/CancerImmunomeDB/. Unfortunately, SEREX has many disadvantages such as the cDNA library preparation is both time consuming and laborious, the use of autologous serum and tissues from the same patient yields many false positive detection and the use of a prokaryotics system prevents detection of antigens with specific post translational modifications[24].

A second proteomics approach to study humoral immune response is serological proteome analysis (SERPA) which is also based on serological analysis. In principle, SERPA combines two conventional proteomics approaches, namely, two dimensional gel electrophoresis (2-DE) to separate a complex mixture of proteins extracted from cell cultures or tumors and western blotting to detect immunogenic proteins in comparative

analysis using sera from the two populations (healthy and cancer)[24]. Compared with SEREX, SERPA is less time consuming and less laborious as it does not require preparation of a cDNA library for the extracted mRNA from tumor. In addition, SERPA is better suited for the detection of possible post translational modification and protein isoforms. However, SERPA has one drawback due to the detection limit of 2-DE. Low abundance proteins, transmembrane proteins which are hydrophobic and insoluble and proteins with extreme isoelectric points are usually difficult to detect by 2-DE. Recent advances in mass spectrometry technology with more sensitivity and specificity are replacing 2-DE for identification of immunogenic proteins.

A third proteomics technology used for the identification of tumor antigens involves combinatorial phage display. This powerful approach relies on the use of bacteriophages to display the antigens that binds to one of the autoantibodies in patient serum. Rather than immunoblotting as with SEREX and SERPA, phage display expressing tumor antigens are screened and amplified by a process called “bio-spanning,” which involves successive rounds of immunoprecipitation of phage libraries using patient serum. Bio-spanning enables rapid detection of tumor antigens and requires only a minute amount of serum. Phage display technology was used successfully to identify several tumor antigens in cancers including prostate cancer[29] and breast cancer[30]. Due to the use of a bacteria system, however, phage surface display suffers from the same drawback as SEREX does: lack of mammalian post translational modifications. In addition, antigens expressed by phage display may not be in the native conformation and each selected phage clone must be individually sequenced, therefore increasing the time and labor cost[23].

1.4 HIGH THROUGHPUT PROTEIN MICROARRAY

High throughput protein microarray has emerged as the most promising proteomics technology to harness the humoral immune response for biomarker discovery. Characterized by unique features such as high throughput, miniaturized and capable of parallel analysis, protein microarray has become an invaluable and indispensable platform for global analysis of the proteome[31]. Publications involving the use of protein microarray technology have increased 15-fold from only 10 articles in 2000 to 152 articles in 2011 (obtained through a simple PubMed search for the term "protein microarray" in the title or abstract. Several excellent reviews on protein microarray technology were recently published [32-35].

Protein microarrays, also known as protein chips, are a miniaturized, parallel assay system containing small amounts of purified proteins in high density array format[31]. Arrays can be made by fixing a membrane on a microscopic glass slide surface using a standard contact or a noncontact microarray printer. While contact printing mechanisms are more robust and cheaper, non-contact printers are much more reproducible and reliable in scientific studies[36]. Depending on the types of applications, a variety of slide surfaces can be used including the popular choices: nitrocellulose, gel-coated, aldehyde- and epoxy-derivatized glass surfaces. A typical microarray is the size of a microscopic slide and can contain hundreds to many thousands of spots depending on the array format and arrayed protein spots diameter. Each spot can contain a set of "bait" molecules ranging from antibodies, cell or phage lysates, recombinant proteins or peptides, drugs, or nucleic acids[35]. The array is hybridized with either a probe (labeled antibody or ligand)

or an unknown biological sample such as a cell lysate or serum sample. The reaction signals between the probe and the target are usually measured and recorded by fluorescent or radioisotope labeling.

The field of protein microarray is under rapid development in a variety of applications to systems biology. There are four major types of protein microarrays: 1) proteome microarrays 2) antibody microarrays 3) reverse phase microarrays and 4) lectin microarrays. Each type of microarray is capable of global and high throughput analysis for systems biology. Proteome microarrays, as the name suggests, contain the majority or all of the open reading frame coded proteins of an organism. It is mainly used for unbiased global discovery study. There are two main classes of proteome microarrays: traditional protein expression based and in vitro transcription and translation (IVTT) based. Expression based proteome microarrays from organisms such as human[37], yeast[38] and Escherichia coli[39] have been developed for discovery based expression profiling. As compared with the expression based proteome microarrays, IVTT based proteome microarrays are less expensive and simpler to fabricate without the need of protein purification and expression. Therefore, IVTT based microarrays are widely used in serum biomarker studies. For example, an IVTT based approach termed NAPPA (Nucleic Acid Programmable Protein Microarray)[40] used to track the humoral response in proteome scale and successfully identified a potential panel of 28 autoantibody biomarkers for the early detection of breast cancer[9].

Before proteome microarrays introduction, antibody microarrays and reverse phase microarrays were the two main uses of protein array technology for proteome discovery and clinical diagnostics[31]. Antibody microarrays are a type of forward phase arrays in

which multiple antibodies are arrayed on glass or membrane surfaces. Each array is then processed with one test sample to see if the test sample has proteins reactive with all the arrayed antibodies. Test samples are usually sera, plasma or cellular lysates for humoral immune response profiling in disease related studies. For example, in a study of LoVo colon carcinoma cells, Sreekumar et al. [41] created a high density antibody microarray with 146 distinct antibodies and were able to obtain proteins differentially expressed in colon carcinoma and also involved in radiation-induced up-regulation of apoptotic regulators. The main difficulty with antibody microarray is the limited availability of antibodies with high specificity and affinity. Haab et al. [42] reported that only 20% of the antibodies available are specific enough to use in an antibody microarray. One of the long term goals of The Human Proteome Atlas Project is to have a validated specific antibody for each protein in the human proteome by 2014[31]. Despite the challenges, many commercial antibody arrays are available and the growing list includes RayBiotech for cytokine detection, Kineuxs for phosphorylation event monitoring and Clontech with its Ab Microarray 500/507 system[32]. Together, these academic and industrial antibody microarrays advanced the field of biomarker discovery in human plasma proteome.

In contrast with antibody microarrays, reverse phase microarrays allow test samples such as recombinant proteins or natural proteins extracted from cellular lysates to print directly on the slides. As a result, thousands of samples can be analyzed simultaneously. The reverse phase microarray is hybridized with a single specific antibody to detect protein expression across many samples. While the antibody microarray has better specificity, reverse phase microarrays are more robust, more sensitive and higher throughput. In addition, since reverse phase microarrays are utilizing cell lysates, they

provide information for the analysis of protein post translational modifications. The most successful use of reverse phase microarrays is in early screening tests in cancer patients. For examples, Paweletz et al. [43] used reverse phase microarrays in the study of the progression of prostate tumors and discovered an association between significant increases in phosphorylated Akt levels and tumor development, resulting in biomarker candidates for early detection of prostate cancers. The biggest challenge in reverse phase microarray is the lack of a standard protocol. There are still several different types of array platforms and in each platform, there are different protein contents. As a result, data from different laboratories are challenging to compare and interpret. In addition, the high throughput nature of protein microarrays generates an enormous amount of data that has to be both analyzed by bioinformatics tools and validated by statistical analyses.

1.5 BIOINFORMATICS APPROACHES IN HUMORAL IMMUNE RESPONSE

With the advent of multiplex, high throughput technologies in the "omics" world, researchers are beginning to appreciate a systematic multi-omics research method which can provide a detailed picture of complex disease processes. This multi-omics strategy provides an information-intensive approach to understand the many layers involved in producing the phenotype of a complex disease, and ultimately play an important role in a holistic understanding of the mechanism underlying the disease phenotype. Now, with the large amount of data available to researchers, the need for developing bioinformatics tools to synthesize, store, understand and translate biological data into clinical application is well recognized. Bioinformatics, being an interdisciplinary subject, successfully integrated "omics" technologies such as genomics, proteomics and much more. An

example of this integration is how several bioinformatics techniques and methods in differential expression analysis can be applied to both genomics and proteomics data, in particular DNA microarray and protein microarray. In fact, in a study of characteristics of expression profiles of protein microarray in humoral immune response, it has been shown that the noise models, normalization, variance estimation and differential expression analysis techniques from DNA microarray data analysis can be applied to protein microarrays[21]. The remaining challenge is to identify which of these DNA microarray analyses are relevant and where and when to adapt them to protein microarray analyses[44].

1.5.1 Normalization

Normalization is a preprocessing step in the standard microarray data analysis pipeline and involves minimizing technical variation introduced during the experimental process[45]. Although protein microarrays do share some mechanistic aspects with the traditional DNA microarrays, they are quite different in several other respects. One noteworthy example is that unlike DNA microarrays, printed spots (features) in protein microarrays usually do not contain equal amount of proteins from spot to spot. As a result, differences in measured intensity between spots in protein microarrays may not reflect biological activities but rather an aspect of microarray construction; hence the need for normalization to the total amount of protein per spots[44] which can be obtained by hybridizing a protein microarray with a labeled universal protein marker such as Sypro Ruby Blot staining (Molecular Probes, Eugene, OR, USA)[46].

Another popular alternative for intensity normalization is the so called variance stabilizing normalization (vsn) which involves glog2 (generalized logarithm)

transformation, a variant of log transformation which can make the variance approximately independent of the mean intensity [47]. Vsn normalization has been widely used in protein microarrays to remove heteroskedasticity of intensity data [48-49]. Sundaresh et al. [21] have applied this method to the malaria dataset and found that the correlation between standard deviation and mean intensity after vsn treatment (r=-0.07) is much lower than those before vsn treatment (r=0.86). In another study, Luevano et al. [49] have claimed that the vsn method can also remove non-specific noise effects and effectively calibrate the data; therefore, it is a recommended preprocessing method for downstream statistical analyses.

1.5.2 Differential expression analysis

Differential expression analysis is used in both clinical genomics and proteomics research to identify genes or proteins that are differentially expressed between two biological conditions. A traditional approach for data analysis uses a comparison of means such as Student's t-statistics to compare the expression level of two groups, e.g. a normal "healthy" group and a disease group, or disease groups with different degrees of severity. To perform the analysis, a statistic is computed for each molecular feature (e.g. a gene or a protein), and the ensemble of such statistics is assessed using a statistical framework such as type-I/type-II error rates, sensitivity/specificity, or false discovery rates. This traditional mean-based approach and its variation have been widely used in numerous studies to successfully identify gene changes in microarray data [50-52]. In the context of protein microarray, Sundaresh et al. [21] demonstrated that approaches based on Student t-test such as Bayes regularized t-test or Cyber-T [53], can also be used to identify significantly expressed antigens targeted by immune responses in malaria. In

another protein microarray study, Zhong et al. [54] used Student t-test to successfully identify multiple non–small cell lung cancer–associated antibodies.

Another modified differential expression analysis recently raised in popularity is based on differential outlier analysis. The underlying assumption of this outlier-based approach is that in heterogeneous diseases such as cancers, only a subset of high risk samples exhibit altered expression of a particular protein[55]. A number of methods have been developed to detect so-called "cancer outlier genes" or genes expressed in only a subset of cancer samples. Methods for cancer outlier profile analysis include the COPA approach of Tomlins et al.[56], the outlier sum (OS) test[55], the outlier robust t-test[57], the MOST method[58], the LSOSS method[59], distribution based outlier-sum statistics[60] and others. The overall goal of this thesis is to focus on the challenge of analyzing protein microarray data by exploring the potential impact of adapting standardized analysis techniques from DNA microarray.

1.6 DISSERTATION OUTLINE

In this dissertation, we attempt to integrate proteomics and bioinformatics approaches to analyze signals from protein microarray data for the reliable identification of differentially expressed proteins under different biological conditions. In Chapter 2, we present a study comparing outlier-based to traditional mean-based approach in differential expression analysis with applications in protein microarray data in heterogeneous diseases. Our investigation uses a glycoproteomics dataset from a melanoma study, an original simulation-based approach to benchmarking and a new data visualization technique to assess the potential for methods that explicitly target

heterogeneous patterns of differential expression to give improved performance relative to traditional approaches based on group-wise comparison of means. In Chapter 3, we apply the outlier strategy to a protein microarray dataset from pancreatic cancer study with 37 pancreatic cancer sera, 24 chronic pancreatitis sera and 23 normal sera to identify protein biomarkers that are differentially expressed in only a subset of cancer samples. Three protein markers which exhibit outlier patterns exclusive to cancer sera and no outliers in normal sera are identified by mass spectrometry and confirmed by a follow up study with an independent dataset. Chapters 4 presents the application of a label-free quantitation approach for measuring changes in protein abundance level associated with phosphorylation and glycosylation respectively. Results are the identifications of differentially expressed post translational modified proteins such as phosphorylated lamin A/C, isoform A and GTPase activating protein binding protein 1 in pancreatic cancer. Chapter 5 summarizes overal conclusions from the study and presents ideas for future studies.

CHAPTER 2

OUTLIER-BASED DIFFERENTIAL EXPRESSION ANALYSIS IN PROTEOMICS STUDIES

2.1 INTRODUCTION

Differential expression analysis is a mainstay of clinical genomics and proteomics research and is used to identify genes or proteins that are differentially expressed between two conditions. A traditional approach for data analysis uses a comparison of means such as Student's t-statistic to compare the expression level of two groups, e.g. a normal "healthy" group and a disease group, or disease groups with different degrees of severity. To perform the analysis, a statistic is computed for each molecular feature (e.g. a protein), and the ensemble of such statistics is assessed using a statistical framework such as type-I/type-II error rates, sensitivity/specificity, or false discovery rates.

Statistics based on mean values, such as Student's t-statistic, perform well when all samples in a group share a common mean, with approximately symmetric variation around the mean. We consider this a "homogeneous situation". In heterogeneous diseases such as some forms of cancer including breast [61], lung [62], prostate [63] and melanoma cancer [64], only a subset of the high risk samples exhibit altered expression of a particular protein, resulting in a skewed distribution. While such skew will shift the mean to some degree, the sample mean may not be the most effective way to identify such a pattern. Moreover, a differentially expressed feature could be up-regulated in some samples, down-regulated in other samples and normally-expressed in others. In this

scenario, the mean expression of this gene or protein could be similar among groups and thus will not be detected using any mean-based approach.

One way to account for this heterogeneity and to improve the detection of differentially expressed genes or proteins is to adopt a modified differential expression statistic that it is more sensitive to heterogeneous patterns of differential expression. To this end, a number of methods have been developed to detect so-called "cancer outlier genes" or genes expressed in only a subset of cancer samples. Methods for cancer outlier profile analysis include the COPA approach of Tomlins *et al.* [56], the outlier sum (OS) test [55], the outlier robust t-test [57], the MOST method [58], the LSOSS method [59], distribution based outlier-sum statistics [60] and others. Compared to the traditional t-statistic, outlier-based methods have the potential to detect a greater number of differentially-expressed genes in heterogeneous data sets, at a lower false discovery rate. However these methods are less powerful than approaches based on t-statistics when the differential expression is present throughout the distribution, or is concentrated in the center of the distribution, as opposed to being concentrated in the tails. Figure 2-1 illustrates this difference by showing quantile functions corresponding to two distributions that are different at all quantile values (Figure 2-1a), and quantile functions corresponding to two distributions that differ only in the upper quantiles, or in the right tail (Figure 2-1b).

Focusing on the OS approach to outlier analysis, we used three complementary approaches to better understand the circumstances in which outlier-based differential expression approaches have the potential to outperform traditional approaches based on mean differences (such as Student's t-statistic). First, we used a simulation strategy in

which the strength of the outlier-pattern of differential expression and the strength of differential expression in the center of the distribution can be independently varied. This allowed us to identify the transition point where the outlier pattern is sufficiently strong for the outlier-based methods to perform best. We then explored a graphical diagnostic that summarizes the patterns of differential expression in a dataset. This diagnostic can be used to reveal the relative amounts of outlier-like versus "central" differential expression in a dataset. Finally, we used a proteomics dataset of serum samples from melanoma patients [65] to contrast the evidence for differential expression obtained using outlier-based versus mean-based approaches to differential expression analysis.

Our simulation studies suggest that approaches based on means are most powerful when the differential expression is strongest in the center of the distribution, or is equally strong at all quantile points. The outlier-based approaches are most powerful when the differential expression is concentrated in the tails of the distribution. Applying our graphical diagnostic to the melanoma data set revealed protein fractions that are differentially expressed primarily in the distribution centers, and protein fractions that are primarily differentially expressed in the distribution tails. However, even the latter proteins still showed shifted mean values, and in this moderate-sized dataset, the outlier approach did not identify any differential expression that was not also captured by the t-statistics (at a fixed significance level). However, inspection of distribution patterns did reveal an outlier-like pattern in several of the differentially expressed proteins, which may aid in understanding their mechanistic roles, or help to better define their utility as biomarkers.

2.2 MATERIALS AND METHODS

2.2.1 Outlier sum

The outlier sum (OS) statistic is intended to detect a difference between two statistical distributions that is concentrated in one or both tails of the distributions. In terms of data, the difference in the tails results in the presence of "outliers" in one of the two sets of samples being compared. Outlier sum analysis contrasts a reference set of samples to a second set of samples in which outliers may be present. The reference set can be chosen as the one corresponding to lower risk subjects (e.g., subjects who are healthy, have benign lesions, or have slowly progressing disease), with outliers assessed in the contrasting higher risk group. Alternatively, we can consider the reference set to be the higher risk set, based on the idea that the greater part of the low-risk set may already be moving toward a more adverse state, with a smaller fraction of the low-risk set (the outliers) not having yet made this transition.

For a given molecular feature, an outlier is defined as an observed value that is greater than the 75th percentile plus the interquartile range (IQR), or less than the 25th percentile minus the IQR. A high OS score means that either a large number of outliers is present (in the non-reference set relative to the reference set), or that a few strong outliers are present. The outlier score for molecular feature i is defined as max($|W_i|$, $|Wi'|$), where:

$$W_i = \sum_{j \in C2} x'_{ij} . I[x'_{ij} > q_{75}(i) + IQR(i)]$$

$$W'_i = \sum_{j \in C2} x'_{ij} . I[x'_{ij} < q_{25}(i) - IQR(i)]$$

Note that x'_{ij} denotes normalized expression value of molecular feature i in sample j, C2 denotes the non-reference set in which we are assessing for outlier samples, $q_{25}(i)$, $q_{75}(i)$, and IQR(i) represent, respectively, the 25^{th} percentile, the 75^{th} percentile, and the interquartile range for molecular feature i.

2.2.2 Simulation study

Previous simulation studies evaluating outlier-based differential expression methods have used Gaussian mixtures [55, 60, 66], or t-distributions [60] to produce synthetic data with outlier-type patterns of differential expression. To understand the operating characteristics of outlier-based differential expression analysis in more detail, we used a simulation approach in which the strength of differential expression in the tails and the strength of differential expression in the center of the distribution can be independently varied.

Our simulation studies are defined in terms of the quantile function Q(p) of the distribution of expression values. The defining property of the quantile function is that a value drawn from the distribution has probability p of being less than or equal to Q(p). The simplest form of differential expression is a constant shift of the quantile function, shifting all quantiles, as well as the mean, by the same amount (Figure 2-1a). To produce outlier-like differential expression (Figure 2-1b), we shifted the quantile function with a "hinge function" of the form $H(p; k, k_0) = k*(p-p_0)*I(p>=p_0)$, where I() is the indicator function that is equal to one when $p>=p_0$ and zero otherwise. The slope parameter k controls how strongly H deviates from a constant function. When we compare a quantile function $Q_1(p)$ to a hinge-shifted quantile function $Q_2(p) = Q_1(p) + H(p; k, k_0)$, we find that the quantiles agree up to the p_0^{th} quantile, but the quantiles for $p>p_0$ are greater under

Q_2 compared to Q_1. This represents an ideal setting for outlier-based analysis, with the advantage becoming stronger as the slope parameter k increases. We note that this construction maintains continuity of the quantile function, consistent with proteomics and genomics data that we have observed.

To aid in interpretation of the simulation studies, all comparisons were made under a fixed value for Cohen's-d effect size [67-68]. We make two types of comparisons – one in which $Q_1(p)$ and $Q_2(p)$ are constant shifts of each other (i.e. $Q_2(p) = Q_1(p) + k$), and one in which $Q_2(p)$ and $Q_1(p)$ are shifted by a hinge function. In the former case, the value of Cohen's d is fixed by setting the value of k. In the latter case, the value of Cohen's d is fixed by adjusting the values of p_0 for a given value of k. Since Cohen's d can be fixed by varying only the parameter p_0 of the hinge function, the slope parameter was available for us to change to control the strength of the outlier pattern. This gives us the ability to independently control the differential expression at the center of the distribution and in the tail of the distribution. In all our simulations, we used a normal distribution to determine the baseline quantile function $Q_1(p)$, but note that $Q_2(p)$ is not normal in the hinge-shifted case.

For each type of data distribution, we compared the power of the t-statistic and outlier sum approaches for detecting differential expression at the conventional type-I error rate of 0.05. This was done for various sample sizes (N=50, 100, 200) and Cohen's-d effect sizes (0, 0.1, 0.2, 0.3, 0.4). The power for detecting differential expression using t-statistics under normal populations for Q_1 and Q_2 was based on the normal approximation to the power function. All other power results were obtained using simulation. Specifically, the power of the outlier sum approach was estimated as the

proportion of simulation runs that reject the null hypothesis of no differential expression at the given significance level (0.05). Since the outlier sum statistic does not have a tractable null distribution, simulation under Q(p) was used to determine the decision threshold. We used 1000 replications for all simulation studies.

2.2.3 Graphical diagnostics

To summarize the pattern of differential expression in the i^{th} protein, we considered the difference $D_i = Q_{i1} - Q_{i2}$ between the estimated quantile functions Q_{i1} and Q_{i2}, for the two groups of samples being compared. For a fixed grid of probability points, the difference in quantile functions D_i was constructed for each protein, and summarized using principal components analysis (PCA). PCA captures the most important directions of variation in the D_i, relative to their mean.

For example, a constant principal component corresponds to constant translation between the quantile functions, producing equal levels of differential expression at all quantiles. A linearly increasing principal component corresponds to differential expression that changes linearly across the quantiles. In practice, these principal components are interpreted relative to the mean value of D, as demonstrated in the analysis of the melanoma data below. The proportion of variance explained by a given principal component indicates the extent to which a particular form of differential expression is present in a dataset. This in turn can be related to differences in statistical power for detecting differentially expressed proteins.

2.2.4 Case study: melanoma dataset

In melanoma, metastasis to sentinel lymph nodes signals a more advanced stage of melanoma, and sentinel lymph node biopsy is common for its prognostic value. A

melanoma dataset was used to explore the utility of serum autoantibodies as biomarkers to distinguish between "node-negative" and "node-positive" melanoma as described by Liu *et al.* [65]. Node-positive status indicates metastasis to a sentinel lymph node, while node-negative status indicates the absence of metastasis, and thus an earlier stage of cancer. The dataset consisted of 43 serum samples from patients with melanoma - 26 from node-negative melanoma and 17 from node-positive melanoma. Using methods described by Liu *et al.* [65], a panel of 47 glycoprotein fractions was extracted from a melanoma cell line to bind to (and enable detection of) autoantibodies in the patient serum samples. The dataset was presented as a matrix of signal intensity data with 43 columns corresponding to serum samples and 47 rows corresponding to cell line-derived protein fractions. We applied our graphical procedure to this dataset, and calculated t-statistics and OS statistics for each of the 47 glycoprotein fractions.

2.3 **RESULTS**

Figure 2-2 shows the results of power analysis based on simulation, in the setting where Q_2 and Q_1 are shifted relative to each other (or equivalently, D is constant). In this setting, the differential expression is homogeneous in that all quantiles, and the mean value, differ by the same amount. The Student t-method provides greater power than the OS method for all effect sizes and sample sizes considered.

Figure 2-3 shows selected results for the setting where Q_1 and Q_2 differ by a hinge function. The slope parameter k is plotted on the horizontal axis, and reflects the degree to which the differential expression is concentrated in the distribution tails. We thus

expect that the OS method will perform relatively well compared to the Student-t method as this parameter increases. The effect size indicated in each plot is Cohen's d.

The powers of both the OS and Student-t methods increase with increasing effect size. When the effect size is small, the OS method outperforms the Student-t method for all tested values of the slope (k), but when the power is somewhat greater, the Student-t method performs better for smaller values of the slope parameter. The first row of Figure 2-3 shows that when no differential expression is present, both methods maintain the correct type-I error rate. Taken together, Figures 2-2 and 2-3 indicate that when the data distribution is approximately homogeneous, the Student t-method outperforms the OS method. But when heterogeneity is present and the differential expression is stronger in the tails, the OS method can outperform the Student t-method. For a fixed level of strength of the outlier pattern (represented by the slope parameter k), the OS method performs increasingly better relative to the Student t-method as the power increases, either due to a larger sample size, or a larger effect size.

We applied our graphical diagnostic procedure to the melanoma dataset to capture the major patterns of differential expression in the 47 glycoprotein fractions. For each fraction, the estimated quantile differences D_i between the node-positive group and the node-negative group were constructed using the eleven deciles (i.e. 0, 0.1, 0.2 … 1) as probability points.

Figure 2-4 shows the results of applying principal components analysis to the D_i vectors. The mean value of D_i (across the 47 fractions) is shown as the broken line in both plots. The solid curves show the mean plus two different multiples of the first principal component (left plot) and of the second principal component (right plot). The

multiples used in the plots were the 5^{th} and 95^{th} percentiles of the corresponding principal component scores. These two dominant principal components explain 61% and 22% of the variance, respectively. The variation resulting from the dominant PC is approximately a vertical translation of the quantile function, in which all quantiles are shifted by roughly the same amount, but the slope is unaffected. The variation resulting from the second PC primarily affects the slope of the quantile function, shifting the pattern of differential expression from the center of the distribution to the tails. Thus we find that an outlier pattern of differential expression may play a role in these data, but it is not as prominent as a simple shifting pattern.

We next considered how the Student-t and OS statistics relate to patterns of differences in the quantile functions. The center panel of Figure 2-5 shows a scatterplot of the t-statistics and OS statistics in the 47 protein fractions. The two statistics are weakly positively correlated, but there are numerous fractions where one statistic is large while the other is not. Thus the statistics are capturing partially overlapping information, with the potential for either statistic to capture information that is complementary to the other.

The five panels of Figure 2-5 surrounding the central scatterplot show examples of quantile functions corresponding to five of the 47 fractions. When the Student t statistic is large but the OS statistic is close to zero, as in fractions 14 and 23, the pattern of differential expression is approximately a translation. When the OS statistic is large, the pattern of differential expression is more hinge-like, as in fractions 19 and 39. When both statistics are small (e.g. fraction 5), there is no differential expression of any type. These plots show that the OS and t-statistics are capturing complementary patterns of

differential expression, and that protein fractions showing both of these complementary patterns can be found in this data set.

Finally, we assessed the statistical evidence for differential expression in each fraction using the Student t-statistic and the OS statistic. Nominal p-values (not adjusted for multiple comparisons) were obtained using permutation analysis with 1000 permutation replications. We found 15 fractions to have nominal significance using the t-statistic, and 1 fraction to have nominal significance using the OS statistic. All fractions that were significant under the OS statistic were also significant under the t-statistic, indicating that in this dataset, the OS approach was unable to uniquely identify any significant fractions. We noted that the OS statistic p-values were non-monotonic functions of the OS statistic magnitudes, which can be explained by the strong dependence of the OS statistic's sampling distribution on the overall shape of the distribution, including the shape of the tails. In contrast, the t-statistic p-values were perfectly monotone in the t-statistic magnitudes.

2.4 DISCUSSION AND CONCLUSIONS

A standard model of cancer holds that cancer-related pathways are activated by expression of oncogenes [69-70]. However, changes in oncogene expression levels, and of their targets are not universal among individuals with the same cancer [61, 71]. This heterogeneity complicates efforts to identify cancer biomarkers for general use. For this and other reasons, regulatory agencies like the Food and Drug Administration often consider cancer biomarker assays as “high-risk”, and the regulatory path for biomarker-based diagnostic and prognostic tests is directed by this risk classification. Regulatory

approval of cancer diagnostic devices requires a deep understanding of the device's operating characteristics, especially in terms of false negative and false positive results. The presence of heterogeneity as explored here complicates efforts to understand these operating characteristics.

Here we have illustrated that when power considerations are favorable, the OS statistic has improved power relative to the Student t-statistic for identifying patterns of differential expression that are concentrated in one or both tails of a distribution. When the pattern of differential expression is present to an approximately equal degree at all quantiles of the distribution, or is concentrated in the center of the distribution, approaches based on means, like the Student t-statistic, can be more powerful than the OS approach. The potential advantage of the OS approach depends on the sample sizes and effect sizes being such that the t-statistic dominates the OS statistic only for a small range of strongly symmetric distributions. Using our graphical diagnostic approach, and inspecting the test results for the melanoma data set, it seems that power considerations for the melanoma data favor more traditional mean-based approaches. A further complicating factor for use of the OS statistic is the appearance of non-monotonic patterns between the test statistic magnitudes and the corresponding p-values. Nevertheless, by inspecting the pattern of differential expression in the protein fractions identified using Students t-statistic, we were able to identify several protein fractions showing a prominent hinge-like pattern of differential expression.

While outlier-based analysis approaches offer the potential to extract useful information from studies that yield minimal interesting results from conventional methods, these biomarkers are by definition limited in their predictive power in an

unselected population. As illustrated in the melanoma study, even markers that are identified using traditional approaches like the Student t-method may turn out to have a heterogeneous pattern of differential expression. Thus, we anticipate that while outlier-oriented statistics like the OS statistic may play a useful role, especially in larger studies, another important consequence of these efforts will result from the more widespread adoption of methods to characterize the detailed pattern of differential expression of candidate biomarkers identified through traditional approaches.

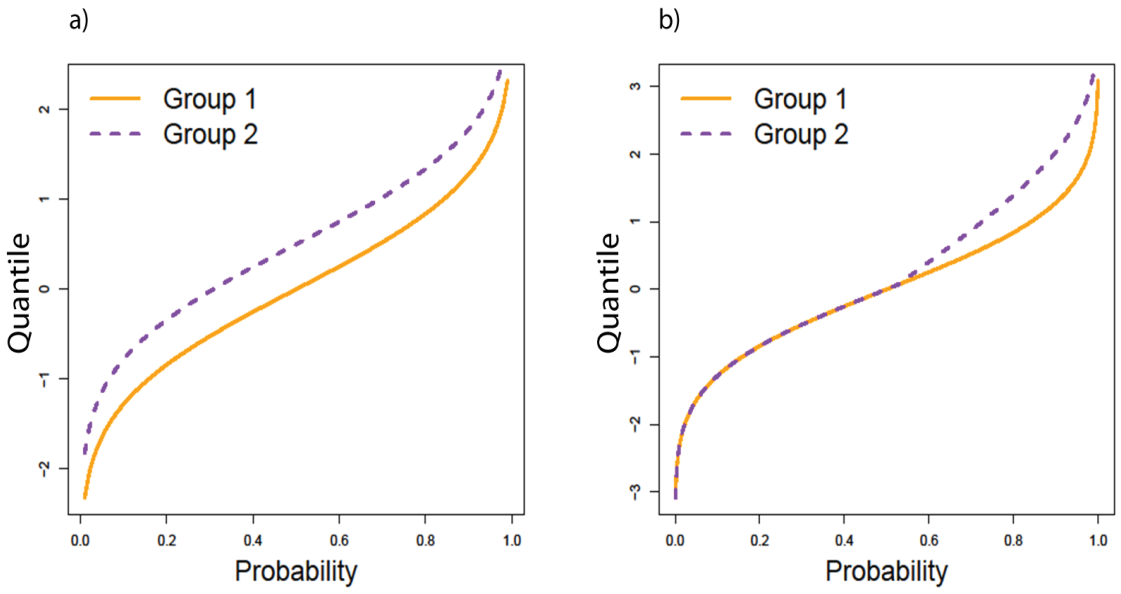


Figure 2-1: Quantile functions of illustrating differential expression with a constant shift of all quantiles (a) and with outlier-like differential expression that is present only in the right tail (b).

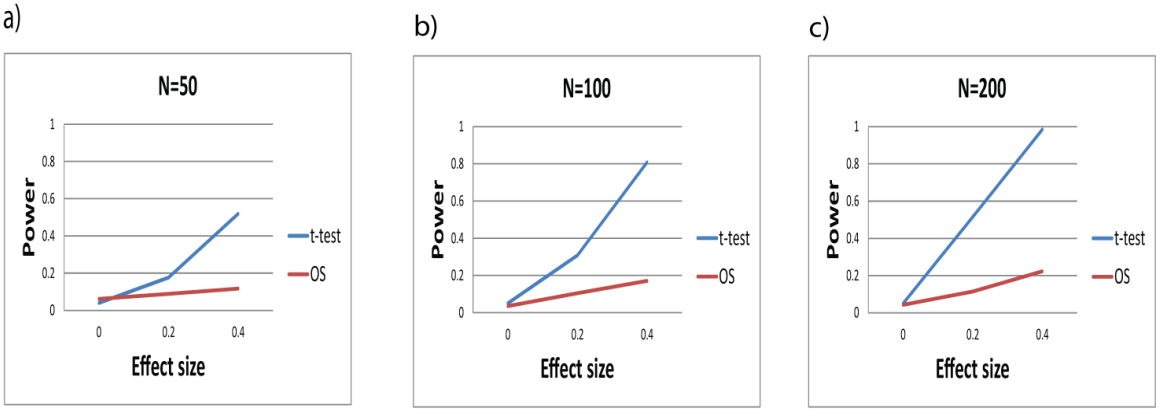


Figure 2-2: Power for Student t analysis and outlier sum analysis for homogeneous data.

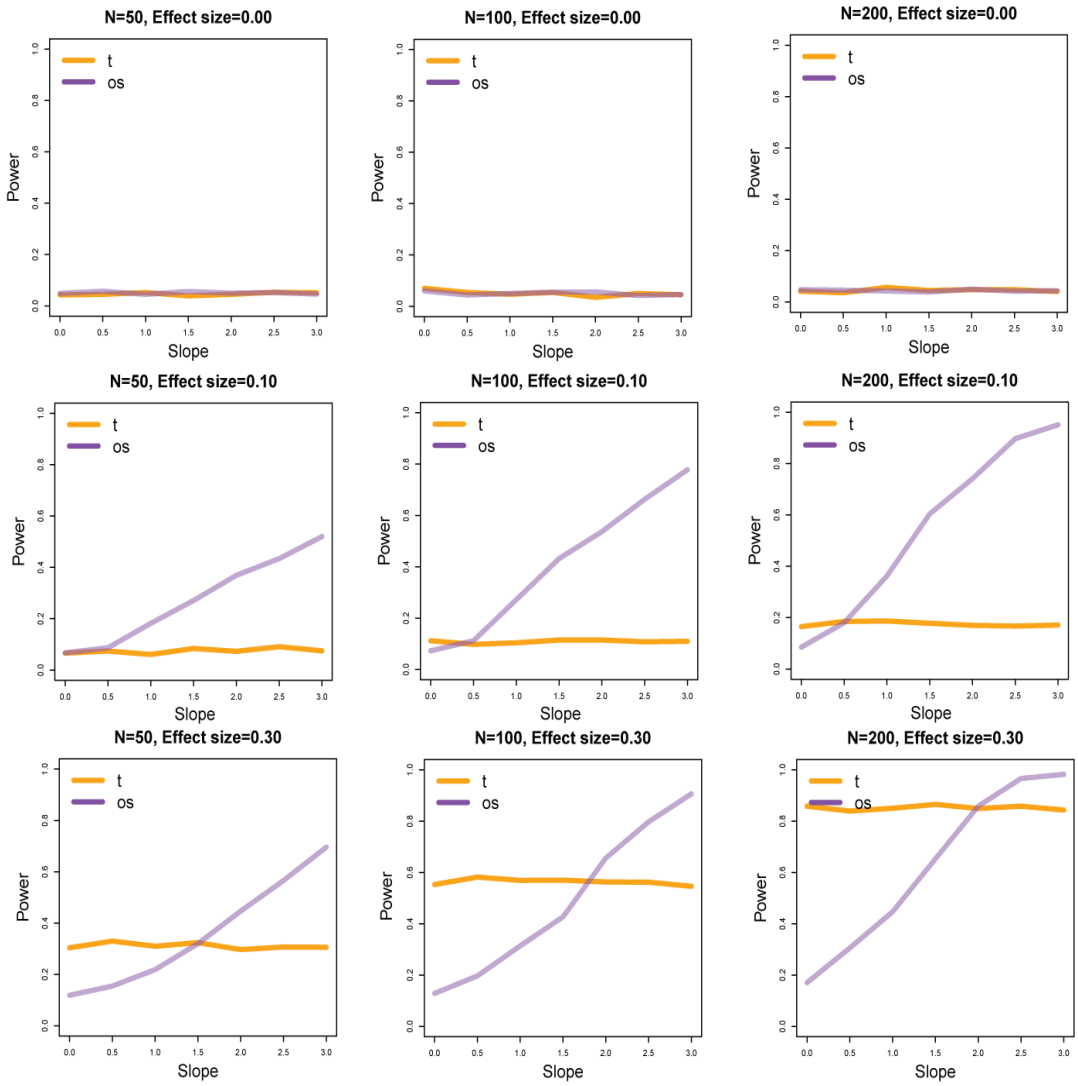


Figure 2-3: Power for Student t analysis and outlier sum analysis for heterogeneous data (rows 2 and 3) and for equally distributed data (row 1).

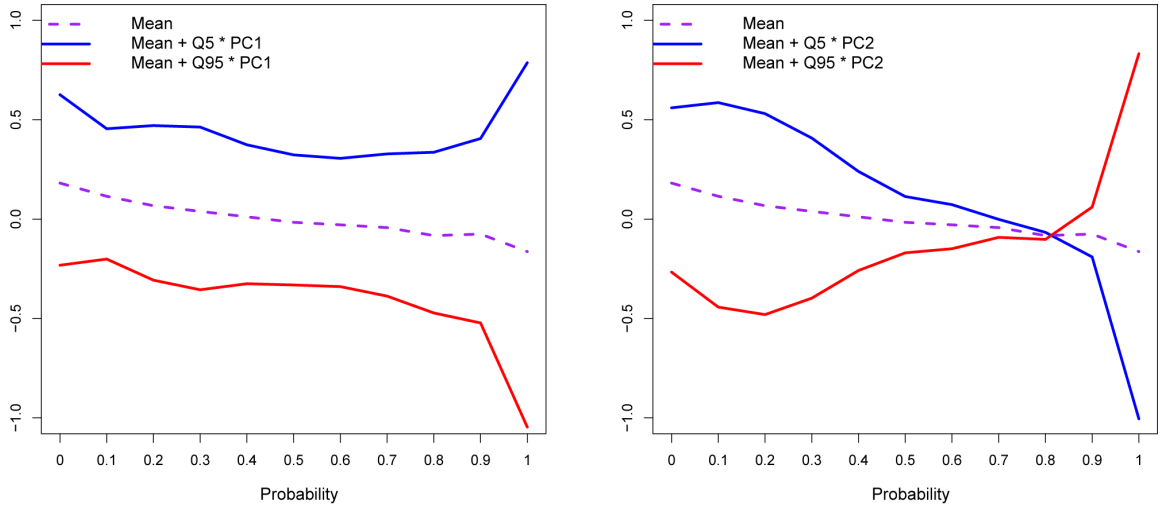


Figure 2-4: The mean of the quantile function difference Di over the 47 protein fractions is shown (broken line), along with plots of the mean plus two different multiples of the first principal component (left) and the second principal component (right).

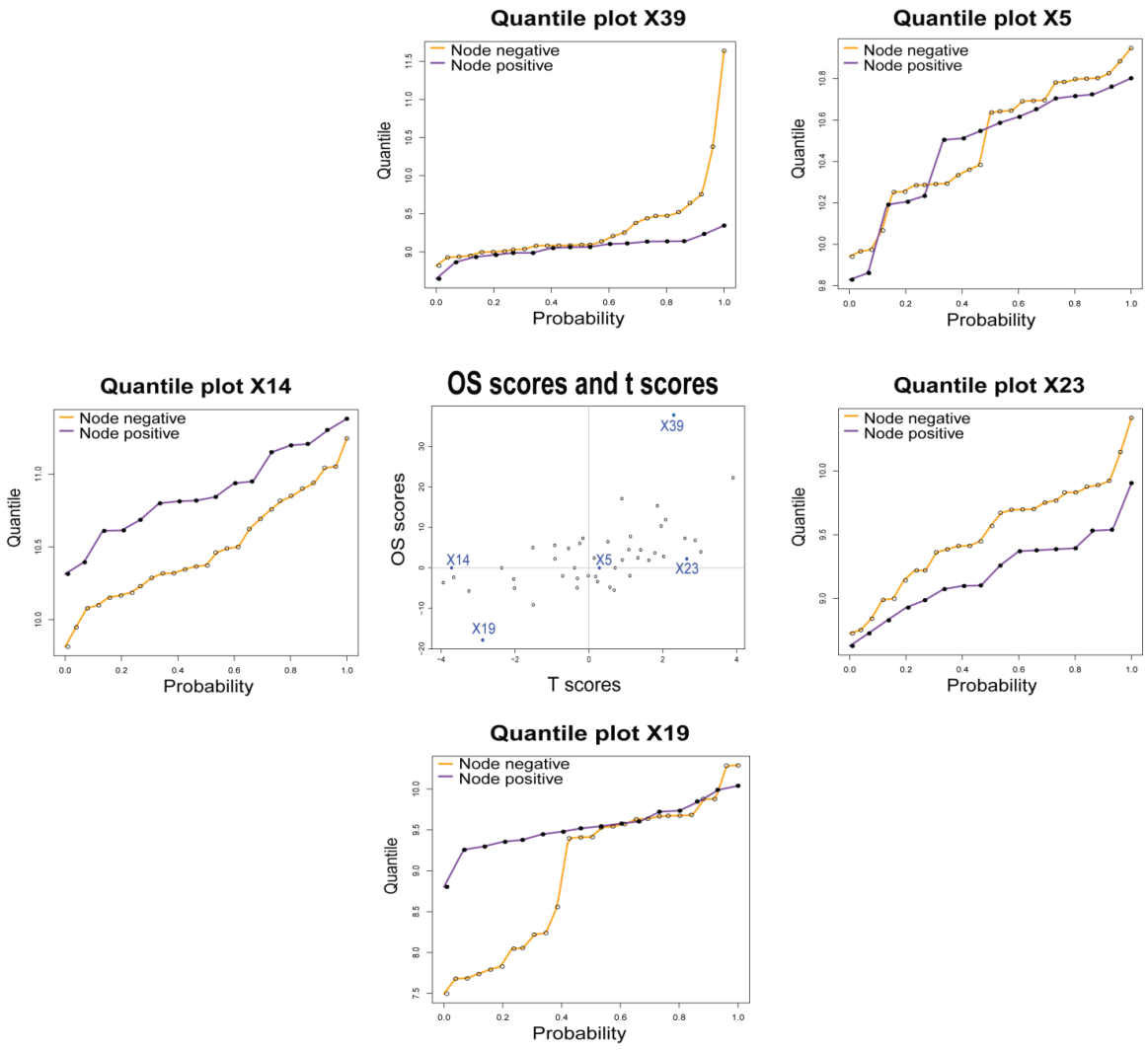


Figure 2-5: Scatterplot of OS statistic values (vertical axis) against t-statistic values (horizontal axis). Surrounding the central scatterplot are five examples of estimated quantile functions for the expression values of specific protein fractions.

CHAPTER 3

THE IDENTIFICATION OF AUTOANTIBODIES IN PANCREATIC CANCER PATIENT SERA USING A NATURALLY FRACTIONATED PANC-1 CELL LINE[1]

3.1 INTRODUCTION

Pancreatic adenocarcinoma is the 4th leading cause of cancer-related death in the United States [1]. Patients with pancreatic ductal adenocarcinoma (PDAC) have one of the poorest survival rates of any cancers. According to the American Cancer Society, for all stages of pancreatic cancer combined, the one-year relative survival rate is 20%, and the five-year rate is 4% [1]. These low-survival rates result from the failure to diagnose PDAC at an early stage when the possibility of a curative resection still exists. This is due to a variety of factors including the inaccessible location of the pancreas deep in the abdomen, late-presenting clinical manifestations (e.g., weight loss, or abdominal pain), and the early development of metastasis. Fewer than 10% of patients' tumors are confined to the pancreas when in most cases, diagnosis of 80% to 90% of PDAC cases are too late for surgical procedures to have a positive outcome[72]. Unfortunately there are not any available diagnostic tools that allow for detection of early stage pancreatic cancer. Although there has been an effort to find protein markers in serum, none have shown sufficient sensitivity and specificity for early diagnosis. This includes the commonly used

[1] Reprinted from Li, C., H.-Y. Kim, H. Vuong, T. Patwa, M. Pal, R.E. Brand, D.M. Simeone and D.M. Lubman; The identification of auto-antibodies in pancreatic cancer patient sera using a naturally fractionated Panc-1 cell line. Copyright (2010), with permission from IOS Press.

CA 19-9 test [73-74] which may be significantly increased in pancreatitis in addition to pancreatic cancer, and is not reliably elevated in early stage cancer.

There remains a need for the discovery of innovative serological biomarkers that effectively improve diagnosis and prognosis of human cancer. Antibody responses associated with the occurrence and progression of solid tumors have been identified in multiple cancer types [75-77]. The underlying mechanism of the auto-immune response is still not fully understood [78]. However, the known molecular changes that can induce auto-immune response include proteins expressed at an aberrant level, mutated gene products, and isoforms of proteins with abnormal post-translational modifications (PTMs) [79-81]. The immunogenic proteins are often found to be intracellular proteins whose functions are linked to the onset and growth of malignant tumors, such as oncoproteins HER-2/Neu and c-MYC [82-84] and tumor suppression proteins such as p53 [85].

Although research has suggested strong correlation between the presence of some autoantibodies and the process of tumorigenesis, the frequency of the appearance of autoantibodies in cancer patients varies i.e. elevated level of a specific autoantibody is always present in a variable subset of patients [76]. A mutation in the p53 gene elicits an autoimmune response in 4-30% of patients in several types of cancer [85]. Around 30% of patients with lung adenocarcinoma exhibit a humoral response to glycosylated annexins I and/or II whereas none of the noncancerous standards exhibit such a response [86]. In PDAC, autoantibodies to DEAD-box protein 48 were observed in 33.3% of pancreatic patient sera, while none of the patients with benign disease and healthy controls showed reactivity against the antigen [87]. MUC1 [88-89], p53 [90] and Rad51 [89] have also shown restricted immune reactivity in a subset of cancer samples. The typical frequency

of the detection of a particular autoantibody in a cancer type is 10-20% and may not be sufficient when used as a biomarker individually, but may be combined as a panel for improved performance [91-93].

Adding new cancer specific autoantigens to the existing biomarker repertoire is the impetus of developing new analytical and statistical techniques for auto-immune response studies. There are several approaches currently available for the identification of auto-antigens. One is targeting specific proteins or gene products known for their roles in cancer including p53, c-myc, and erB-2 [94]. This method only provides limited candidates for biomarkers. Recombinant protein microarrays produced from cDNA expression libraries has been used as a comprehensive antigen substrate to profile the autoimmune reactivity; however, it is unable to profile posttranslational modification (PTM) dependent antibody-antigen interactions [76]. The development of proteomic separation and identification techniques has benefited the discovery of autoantibody biomarkers, where proteins from tissue or cell lines are fractionated by gel or liquid based multidimensional separations, while maintaining the natural PTMs.

In this chapter, we present our use of liquid fractionation methods to produce microarrays for the humoral response experiment against a Panc-1 pancreatic cancer cell line. The methods involve separating intact proteins from cell lysates using two dimensions. A total cell fractionation can be performed using chromatofocusing separation in the first dimension where the proteins are fractionated according to pI. Each fraction is then separated in a second dimension by nonporous silica RP HPLC [95]. Using this method, isolated proteins in the liquid phase can be collected for spotting on coated glass slides [96]. The protein spots are probed for their humoral response by

exposing them to sera from cancer, chronic pancreatitis patients, and normal individuals. This method offers a means for comprehensive proteomic analysis of proteins from large numbers of purified proteins as expressed in cancer cells while maintaining their PTMs that are often critical to the humoral response [97]. The method can produce arrays with over a thousand spots and can produce large numbers of slides for testing the response against a large number of patients.

In order to account for the fact that a specific autoantibody is more often found in only a subset of the patients with the corresponding tumor, we attempt to apply two statistical methods, Wilcoxon rank-sum test and outlier-sum statistics [98], to find potential markers that show different types of immune reactivity patterns. Based on our results we perform a confirmation study of 5 potential markers for pancreatic cancer against recombinant proteins on a microarray-based format against samples from pancreatic cancer, pancreatitis, type 2 diabetes and normal controls. For three of the five proteins, a substantial number of samples from the cancer group show higher reactivity than the non-cancer samples.

3.2 EXPERIMENTAL SECTION

3.2.1 Chemicals

Methanol, acetonitrile, urea, thiourea, iminodiacetic acid, dithiothreitol (DTT), n-octyl-D-glucopyranoside (OG), glycerol, bis-tris, Trifluoroacetic acid (TFA), and PMSF (Phenylmethanesulfonyl fluoride) were purchased from Sigma (St. Louis, MO). Water was purified using a Milli-Q water filtration system (Millipore Inc., Bedford, MA) and all solvents were HPLC grade unless otherwise specified. Reagents were used in the purest

form commercially available. Polybuffer 74 and polybuffer 96 were purchased from GE Healthcare Bio-Sciences Corp. (Piscataway, NJ). 1x PBS and ultra pure DNase/RNase-free distilled water were obtained from Invitrogen (Carlsbad, CA).

3.2.2 Serum samples

As a discovery set, 86 serum samples were obtained at the time of diagnosis following informed consent using IRB-approved guidelines. Sera were obtained from patients with a confirmed diagnosis of pancreatic adenocarcinoma in the Multidisciplinary Pancreatic Tumor clinic at the University of Michigan Hospital. Inclusion criteria for the study included patients with a confirmed diagnosis of pancreatic cancer, the ability to provide written informed consent and the ability to provide 40 ml of blood. Exclusion criteria included inability to provide informed consent, patients actively undergoing chemotherapy or radiation therapy for pancreatic cancer and patients with other malignancies diagnosed or treated within the last 5 years. Sera were also obtained from patients with chronic pancreatitis who were seen in the Gastroenterology Clinic at University of Michigan Medical Center and from control healthy individuals collected at the University of Michigan under the auspices of the Early Detection Research Network (EDRN). Some pancreatic cancer samples were obtained under IRB approval from University of Pittsburg Medical Center and process similarly following EDRN guidelines. The mean age of the tumor group was 65.4 years (range 54-74 years) and the chronic pancreatitis group was 54 years (range 45-65 years). The sera from the normal subject group and the tumor group were similar in age and sex. The chronic pancreatitis group was sampled when there were no symptoms of acute flare of their diseases. All sera were processed using identical procedures. The samples were permitted to sit at room

temperature for a minimum of 30 minutes and a maximum of 60 minutes to allow the clot to form in the red top tubes, and then centrifuged at 1,300 x g at 4^{o}C for 20 minutes. The serum was removed, transferred to polypropylene, capped tubes in 1 ml aliquots, and frozen. The frozen samples were stored at -70^{o}C until assayed. All serum samples were labeled with a unique identifier to protect the confidentiality of the patient. None of the samples were thawed more than twice before analysis. In addition to the discovery set, another set of samples with no overlap with the discovery set was used in the confirmation experiment. The demographic and clinical information of the samples in the confirmation set are shown in Table 3-1.

3.2.3 Cell culture

The Panc-1 PDAC cell line was cultured in Dulbecco's modified Eagle medium supplemented with 10% fetal bovine serum, 100 units/mL penicillin and 100 units /mL streptomycin (Invitrogen, Carlsbad, CA). Upon reaching 80% confidence, the cells were washed twice in 10mL 1X PBS containing 4mM Na_3VO_4, 10mM NaF and one half of a protease inhibitor cocktail tablet. The sample was then solubilized in 300ul lysis buffer consisting of 7M urea, 2M thiourea, 100mM DTT, 0.5% biolyte ampholyte 3-10, 2% OG, 4mM Na_3VO_4, 10mM NaF and 1mM PMSF at room temperature for 30min, followed by centrifugation at 35000 rpm at 4^{o}C for 1 hour. The supernatant was stored at -80°C until further use.

3.2.4 Chromatofocusing (CF)

Prior to CF, a PD10 column (Amersham Biosciences) was used to exchange the cell lysate from the lysis buffer solution to the CF buffer solution according to the manufacturer's protocols. The start buffer consisted of 6M Urea, 0.2% OG, 25mM bis-

tris. The elution buffer solution was composed of 6M urea, 0.2% OG, and a 10 fold dilution of polybuffer 96 and polybuffer 74 in a ratio of 3:7. The pH of both buffer solutions (7.9, 4.0) was adjusted with saturated iminodiacetic acid. A CF column (weak anion exchange HPCF-1D prep column, 250mm x 4.6mm ID, Eprogen, Darien, IL) was preequilibrated with the start buffer solution and 13mg of the cell lysate was injected into the CF column with multiple injections. Fractionation was started after switching elution buffer and a stable baseline achieved. The pH fractions were collected in 0.3 pH intervals and pH was monitored using a flow-through on-line pH probe. UV absorption was recorded at 280 nm. When a pH of 4.0 was reached, elution buffer solution was switched to a 1M NaCl solution to wash the column followed by isopropanol to elute out strongly bound proteins from the column. The collected fractions were stored at -80°C.

3.2.5 Reverse Phase HPLC Separation

An ODS-1 (8 x 33 mm) column (Eprogen, Inc.) was used to separate the pH fractions of the Panc-1 cell line after CF. Solvent A was 0.1% TFA in water and solvent B was 0.1% TFA in acetonitrile. The gradient was run from 5% to 15% B in 1 minute, 15% to 25% in 2 minutes, 25% to 31% in 2 minutes, 31% to 45% in 10 minutes, 41% to 47% in 6 minutes, 47% to 67% in 4 minutes, 67% to 100% B in 3 minutes, and reduced to 5% B in 1 minute after maintaining 100% B for 1 minute. The flow rate was 1 mL/min and the column temperature was 65°C. UV absorption was monitored at 214 nm. The fractions were collected in 96-well plates and stored at -80°C.

3.2.6 Protein Microarrays

Approximately 30% of the total sample of the fractionated Panc-1 proteins obtained using 2D separation were transferred into 96-well printing plates (Bio-Rad) and

lypholized to dryness. The fractions were reconstituted in printing buffer which was composed of 62.5mM Tris-HCl (pH 6.8), 1% w/v sodium dodecyl sulfate (SDS), 5% w/v dithiothreitol (DTT) and 1% glycerol in 1X phosphate buffered saline (PBS). Reconstituted fractions in the printing plate were placed in a shaker overnight at 4°C. The fractions from the printing plate were spotted onto nitrocellulose slides using a non-contact piezoelectric printer (Nano-Plotter^{TM} NP2.1 GeSiM). Each spot contained 2.5nL of liquid of ~450μm diameter and the distance between spots was 600μm. Printed slides were dried on the printer deck overnight and stored in a refrigerator desiccated at 4°C if the slides were not used immediately.

3.2.7 Hybridization of slides

The printed slides were blocked in a solution of 1% BSA in PBS-T (0.1%) overnight. Each serum sample was diluted 1:400 in probe buffer which consisted of 1% BSA, 0.5mM DTT, 5mM magnesium chloride, 0.05% Triton X-100 and 5% glycerol in 1X PBS. The slides were hybridized in diluted serum for 2 hours using a mini-rotator at 4°C. After hybridization, slides were washed five times using probe buffer for 5 minutes each time, and then rehybridized with goat anti-human IgG conjugated with Alexafluor 647 (1μg/mL, Invitrogen, Calsbad, CA) for 1 hour at 4°C. The slides were washed five times again with probe buffer for 5 minutes each and dried. All slides were scanned using an Axon 4000B microarray scanner (Axon Instruments Inc., Foster City, CA).

3.2.8 Data acquisition and analysis

3.2.8.1 LC-MS/MS

The residual two-thirds of the sample in 96-well plates which was not used in microarray experiments were dried down to approximately 10μL and mixed with 10%

(v/v) ammonium bicarbonate, 10% (v/v) DTT, and 1:50 ratio (v/v) TPCK-treated trypsin (Promega, Madison, WI). The solution was incubated at 37°C overnight and the tryptic digestion was terminated by addition of 2.5% (v/v) of TFA. The digested peptide mixture was analyzed by nano-flow reverse-phase LC/MS/MS using the LTQ mass spectrometer with a nano-spray ESI ion source (Thermo, San Jose, CA). The samples were separated using a (0.1 x 150mm) capillary reverse phase column (Michrom Bioresources, Auburn, CA) with a flow rate of 5ul/min. An acetonitrile:water gradient method was used, starting with 5% acetonitrile which was ramped to 60% in 25 minutes and to 90% in another 5 minutes. Both solvent A (water) and B (acetonitrile) contained 0.3% formic acid. The electrospray voltage was 2.6kV, with a capillary temperature of 200°C and a capillary voltage of 4kV. The normalized collision energy was set at 35% for MS/MS. The MS/MS spectra obtained were analyzed using the Sequest feature of Bioworks 3.1 SR1, allowing only one missed cleavage during Swiss-Prot human protein database searching. To further validate data obtained from Sequest, Protein Prophet^{TM} and Peptide Prophet^{TM} software modified in house was used to provide a confidence level in identification of 95%. Since there might be more than one protein in a protein spot on the microarray slide, we compared proteins identified in adjacent fractions. If the spot that responded to the humoral response was unique and did not have an adjacent spot that lit up then the highest scoring protein based on LC/MS/MS analysis and Protein Prophet^{TM} and Peptide Prophet^{TM} was considered as the most likely identification. If more than one protein was identified in the spot, we performed mass spectrometry analysis on the adjacent spots. If the proteins were identified in the adjacent spots that did not respond then they were likely not to be the protein with the humoral response in our unique spot. However, if

adjacent spots also showed a humoral response then the protein present in all spots was considered as the most likely candidate.

3.3.8.2 Statistical Analysis

GenePix 6.0 software was used to grid all spots to determine the fluorescent intensities at wavelength 635nm and median local background intensities for each spot. Background subtracted intensities of the spots were taken into analysis if the foreground intensity was at least twice the background intensity (i.e. signal to noise ratio is greater than 2). The signal intensities from all the slides were normalized to minimize experimental slide-to-slide variation. The data for each individual sample in the columns was centered by the median and scaled by the interquartile range (IQR). Two types of statistical analysis were applied to the normalized data in search of biomarkers with up-regulated response in the cancer samples compared to the normal and pancreatitis samples. The non-parametric Wilcoxon rank-sum test was employed to identify fractions showing a universally increased reactivity in the cancer samples. The outlier-sum test was performed to select the fractions that react with only a subset of the samples in the cancer group.

3.2.8.3 Wilcoxon rank-sum test

Two pairwise Wilcoxon rank-sum tests were performed between cancer versus normal and cancer versus pancreatitis. The fractions with the lowest p-value and minimal correlation were combined in Receiver Operation Characteristic (ROC) analyses to determine their sensitivity and specificity in differentiating the sample groups. The Wilcoxon rank-sum tests and the ROC analyses are programmed in R.

3.2.8.4 Heatmap

The fractions with a p-value less than 0.02 in Wilcoxon rank-sum tests were clustered and shown in heatmaps. The $p<0.02$ threshold was determined to have proper numbers of fractions to show in the heatmaps. The heatmaps and dendrograms were drawn in R.

3.2.8.5 Outlier-sum statistics (OS)

The dataset was first standardized for each fraction by subsequently subtracting the median and dividing the median absolute deviation (MAD). The 75% quartile (q(75)) plus the interquartile range (q(75)+IQR) was used as a threshold. The data points beyond this threshold were defined as the outliers. The outlier-sum statistic is the sum of the values of these data points in the cancer groups. Fractions with outlier-sum statistics ranked top 5% and no outliers in the normal groups were considered to be differential. The overlapping fractions found in the pairwise comparisons between cancer/normal and cancer/pancreatitis were presented in a bar graph form (Figure 3-1) (made in R with COPA package).

3.2.8.6 Confirmation Using Recombinant Proteins

Recombinant proteins were purchased from Abnova Corporation (Taiwan) and Genway Biotech Inc., (San Diego, CA). The concentration of each recombinant protein was 10ug/mL. A piezoelectric non-contact printer (Nano-Plotter^{TM} NP2.1, GeSIM) was used to print all the recombinant protein arrays on ultra thin nitrocellulose slides (PATH slides, GenTel Bioscience). Each spotting event which resulted in 500 pL of solution being deposited was programmed to occur 5 times per spot to ensure that 2.5 nL of solution was deposited on each spot. Each recombinant protein was printed in triplicate and 14 identical blocks were printed on each slide. The slides were washed three times

with 0.1% Tween in PBS buffer (PBS-T 0.1) and then blocked with 1% bovine serum albumin (Roche) in PBS-T 0.1 for one hour. The blocked slides were dried by centrifugation and inserted into a SIMplex (GenTel Bioscience) multi-array device which divided each of the slides by 16 wells. The wells separated the neighboring blocks and prevented cross contamination. Serum samples were diluted ten times with PBS-T 0.1 containing 0.1% Brij. One hundred microliters of each diluted sample was applied to the recombinant protein arrays and the hybridization was performed in a humidified chamber for one hour. The 165 samples from different groups were perfectly balanced on each slide to eliminate bias from block-to-block variation and slide-to-slide variation. Two blocks on each of the slides were hybridized with two specific samples and used as control blocks for data normalization. The slides were then rinsed three times to remove unbound proteins. One ug/mL goat anti-human IgG conjugated with Alexafluor647 (Invitrogen, Carlsbad, CA) solution was used for detection. After a second one-hour hybridization with anti-human IgG, the slides were washed and dried again, then scanned with a microarray scanner (Axon 4000A). The program Genepix Pro 6.0 was used to extract the numerical data. The signals from different slides were normalized with the averaged signal of the control blocks on each slide.

3.3 RESULTS AND DISCUSSION

The proteins from Panc-1 human pancreatic ductal adenocarcinoma (PDAC) cell line were used as bait to study the humoral response in pancreatic cancer since the Panc-1 cell line is a good representative sample of human pancreatic cancer [99]. The overall analytical work flow is illustrated in Figure 3-2. The solubilized protein solution

extracted from Panc-1 cell line was fractionated using 2-D liquid separation methods as described consisting of chromatofocusing in the first dimension followed by nonporous reversed phase HPLC where intact proteins were collected as the final products. Fraction collection was performed where liquid eluent from each chromatographic peak was collected into 96-well plates. Each collected protein fraction was separated into two parts for further work. One portion was used for spotting the microarray plates and the other portion was used for protein identification based on LC-MS/MS. There were 1052 protein peaks obtained over a pH range of 8.0-4.0 spotted using the microarray device onto each nitrocellulose coated glass slide. Each slide was hybridized against a patient serum sample where the humoral response was profiled in 37 cancer serum samples, 24 pancreatitis serum samples, and 23 normal controls. Statistical analysis including non-parametric Wilcoxon rank-sum tests and outlier-sum statistics were then performed over this sample set to determine which proteins provided a significant response to patient sera. For the selection of identified proteins, a confirmation study using a second independent set of 168 serum samples was performed where five recombinant proteins were arrayed on nitrocellulose slides and probed with serum from a separate cohort of normal, pancreatitis, type 2 diabetes, and pancreatic cancer patients.

3.3.1 Microarray Result of Humoral Response

The heterogeneity of humoral response has been displayed in a substantial percentage of patients with increased antibody expression to disease related antigens, where only a subset of patients has an autoimmune response to a particular antigen. We herein assume that autoimmune markers show either an increased level of reactivity against most of the patient sera or an outlier pattern that exclusively appear in the cancer

group. Two statistical methods, Wilcoxon rank-sum test and outlier-sum test, were applied to the dataset to search fractions for autoantibody response.

3.3.2 Statistical Analysis

Rather than traditional t-test, Wilcoxon rank-sum test is preferred in several previous humoral response studies because the dataset do not always fit a Gaussian distribution. The test generates a list of fractions with significantly greater intensities in the cancer group (p-value set at <0.02) in the pairwise comparisons in cancer versus normal and cancer versus pancreatitis. Twenty nine fractions are selected in the cancer/normal pair and only seventeen pass the threshold in the cancer/pancreatitis pair. Figure 3-3 shows two heatmaps of these fractions after they are clustered using a hierarchical clustering algorithm. The clustering tree is added on top of the heatmaps. In the first heatmap/dendrogram, 65% (24 out of 37) of the cancer samples and only 17% (4 out of 23) normal samples are clustered on the left side with more blue bands which indicate increased reactivity with serum. Similarly, the left side of the second heatmap/dendrogram includes 70% (26 out of 37) of the cancer samples and 29% (7 out of 24) of pancreatitis samples. Most of the samples are clustered with their own groups while a portion of the samples are not. Several reasons can be taken into account for this result. The incorrectly clustered cancer samples may not contain the antibodies to some particular antigens in the Panc-1 cell line. Additionally, the non-cancer samples that incorrectly clustered with the majority of the cancer samples may be reactive to the co-eluted proteins in some fractions containing cancer related antigens.

In a 2-dimensional separation, a protein often appears in two or more subsequent fractions rather than one because of the limited chromatographic resolution and the post-

column diffusion. In Figure 3-3, it is worth noting that the fractions in the heatmap are often accompanied by their adjacent fractions (red circled) e.g. 7B1-7B3 and 4E4-4E11. The consecutive bands with a smooth reactivity profile are better candidates for further investigation and confer important information for protein identification.

The result of the Wilcoxon rank-sum tests can be transformed to a ROC curve which estimates the ability of the selected biomarkers to distinguish cancers from non-cancers. For each ROC curve, an area under the curve (AUC) value is reported, where 1.0 represents perfect separation of one group from the other and 0.5 represents a completely random result or no separation. In both cancer versus normal and cancer versus pancreatitis categories, the top-ranked fractions in Wilcoxon rank-sum tests exhibit AUC values of 0.70-0.72. To improve the AUC value, we combined the top three fractions with the lowest correlations in ROC analyses so as to avoid combining neighboring fractions. The AUC values of the ROC curves (Figure 3-4) for the three combined fractions are 0.813 and 0.792 for cancer versus normal and cancer versus pancreatitis respectively.

3.3.3 Outlier-sum statistics

Recently, outlier-based statistical methods, such as cancer outlier profile analysis (COPA) [100] and outlier sum (OS) [98] have been proposed as methods for searching cancer related genes with microarray techniques. The outlier-based analysis is able to detect a small number of significantly upregulated signals from microarray data in the disease group while the signal from the majority may not necessarily change. Since the majority of cancers have heterogeneous activation for different individuals, it appears that the application of this method using the “subset” idea where some cancers respond to the

humoral response and others do not respond may result in an improved performance for microarray data. Of the two outlier-based methods, COPA identifies pairs of biomarkers with mutually exclusive upregulated samples because it was designed to search for gene activation with a mutually exclusive mechanism, while the protein biomarkers in this work may not have the same feature. OS identifies outliers in a similar manner as COPA, but it calculates an outlier score for each individual. Therefore, OS is the preferred method in this study.

After OS analysis was applied to the dataset, we found 9 fractions (listed in Table 3-2) that ranked in the top 5% in both of the comparisons of cancer/normal and cancer/pancreatitis. The reactivity profiles for the top 3 fractions are shown in bar graphs in Figure 3-1. It shows that only subsets of the cancer group show increased reactivity against these fractions, while the signals from the other samples remain the same. The signals of subsets with increased reactivity in the cancer group are much higher than the range of the non-cancer groups, which remain close to the baseline. Such an outlier pattern for these fractions suggests that they may contain proteins that are only immunogenic for a subgroup of the cancer samples and not immunogenic for all the non-cancer samples. In clinical application, these fractions can provide information for accurate diagnosis as the immunogenic cancer samples distinguish themselves with a high signal.

The results from Wilcoxon rank-sum test and outlier-sum test are compared, where the lists of marker fractions do not overlap. We are more interested in the candidates given by OS test as these fractions exhibiting an outlier pattern exclusively in the cancer group would be more useful in diagnosis. The list of candidate fractions from

the OS test are thus identified with mass spectrometry and their identifications and performance were confirmed with the recombinant protein arrays.

3.3.4 Mass Spectrometry Identification

LC-MS/MS is used to identify the proteins in candidate fractions and their adjacent fractions. As expected, multiple proteins are identified in each of the fractions. The identified proteins are screened based on an assumption that their reactivity profile should be consistent with their appearance in the neighboring fractions. The resulting protein IDs are listed in Table 3-2 for each of the fractions.

3.3.5 Biomarker confirmation

Due to the large number of fractions from the 2-dimension separation, a set of only 84 serum samples was used to search for the fractions that could be potential biomarkers, where a larger set is usually required for confident biomarker discovery. It is also necessary to confirm the protein IDs identified for the candidate fractions. To confirm these potential markers, recombinant proteins were tested with a different sample set. Five commercially available recombinant proteins were selected for the confirmation experiment with 48 samples from the cancer group, 40 samples from pancreatitis group, 40 samples from normal group, and 37 samples from diabetes group. Type 2 diabetes samples are included since some pancreatic patients also develop this condition which might be responsible for the autoimmune reactivity.

In order to measure the autoantibody response that is elicited against the recombinant proteins correctly, care must be taken to avoid saturating the signal. Hence, the serum must be diluted sufficiently so that the amount of available autoantibody in the serum is lower than the binding capacity of the specific recombinant protein. Therefore,

a saturation curve was made using different dilutions of serum to hybridize against identical blocks of the recombinant proteins. The result of the saturation test showed that with ten-fold dilution, the recombinant proteins were not saturated and yielded a signal/background ratio of >5. Higher or lower dilution resulted in partial saturation or decreased signal intensity. A ten-fold dilution factor was therefore used in the current pre-confirmation experiment using recombinant proteins. The microarray data (background subtracted) was also adjusted by the average signal of control blocks on each slide and standardized for each recombinant protein.

In Figure 3-5, we show the plot of the distribution of the reactivity for each of the recombinant proteins against the sera. The recombinant protein that produces the best result is phosphoglycerate kinase 1 (PGK1), where 10 outliers out of 48 total samples are observed in the cancer group, while no outlier is present in the other three non-cancer groups. PGK1 protein is a kinase in the glycolytic pathway and can be up-regulated by HIF-1α in the cellular response to hypoxia to provide energy for tumor cell proliferation [101]. Genomics-based studies have found that it acts as a suppressor of proangiogenic factor such as VEGF and triggers metastasis due to its effect on the increased expression level of β-catenin, chemokine CXCR4 and CXCL12 [102-104]. At the protein level, PGK-1 has been found overexpressed in pancreatic cancer tissue versus adjacent controls and also elevated significantly in the sera of pancreatic cancer patients (19% strongly up-regulated, 50% weakly-moderately up-regulated) [105]. The performance of PGK-1 in the confirmation experiment using recombinant protein indicates that a lower percentage of patients elicit auto-response. In a future study, it would be interesting to see whether

there is a correlation between serum level of PGK-1 and autoantibody level and how the production of antibody affects the development of the cancer.

For both malate dehydrogenase (MDH1) and ADP-ribosylation factor interacting protein 2 (ARFIP2), there are 4 such outliers in the cancer group. The absence of outliers in the non-cancer group indicates that these 3 recombinant proteins are exclusively antigenic in cancer sera and could be tumor-associated. In Table 3-3, the performance of these 3 biomarkers used together to distinguish cancer is estimated. A cutoff equal to the highest signal in a certain non-cancer group is applied to define the reactive samples in the cancer group. The 3 recombinant proteins together distinguish more than 40% of the cancer samples from the normal and diabetes group, while only 29.2% from the pancreatitis group.

For annexin A2 (ANXA2), the cancer group only has one outlier that is above all the other groups. This is not consistent with the OS analysis which showed differential humoral response in the fraction where ANXA2 was identified in pancreatic cancer sera. It could be due to the use of the recombinant proteins, which may lack the required PTMs to induce a humoral response or the protein may not be in a form to induce a humoral response [88]. Also, since multiple proteins are identified in the fraction, the protein that is responsible for the observed humoral response may not be ANXA2. Heterogeneous nuclear ribonucleoprotein A2 (HNRPA2) produced a more unexpected result in the confirmation experiment where it showed a universal increase in the reactivity against diabetes samples, while the signals of the other three groups remained at the same level.

3.4 CONCLUSION

We have presented a study of the cancer-related humoral response on pancreatic adenocarcinoma using 2-dimensional separation and protein microarray techniques. After analyzing the data with two statistical tests, the fractions showing outlier patterns in outlier sum test were chosen for identification of the proteins and confirmation using recombinant proteins. In the confirmation experiment, 20.8% of the cancer samples demonstrated strongly elevated reactivity for PGK-1, while no proteins in the non-cancer groups were found to react. This result suggests that the autoantibody level of PGK-1 in the serum is useful as a diagnostic biomarker indicating the presence of cancer. Future study of the correlation between the protein level and autoantibody level of PGK-1 in cancer patients may provide a better understanding of the role of PGK-1 in cancer development.

Table 3-1: Demographic and clinical characteristics of the samples used in the experiment.

	Disease Groups			
Characteristics	**Cancer**	**Pancreatitis**	**Normal**	**Diabetes**
Age	**66.6**	**55.8**	**52.8**	**65.2**
Gender (Male)	**56.2%**	**62.5%**	**65.0%**	**35.1%**
Clinical	**Pancreatic Adenocarcinoma Stage I/II 20.8% Stage III/IV 79.2%**	**Chronic Pancreatitis (No acute symptoms)**	**Healthy**	**Type II Diabetes**

Table 3-2: List of differentiated fractions picked by OS analysis in both pairwise comparison between cancer versus normal and cancer versus pancreatitis. Information about the identified proteins in these fractions is also included.

Fraction	Access Number	Protein Name	Fraction pH	MW	Seq Cov %	Theoretical pI	Unique peptides
1B1	P62847	40S ribosomal protein S24	7.9-7.6	15414	28.59	10.79	3
1F11	P00558	Phosphoglycerate kinase 1	7.9-7.6	44587	18.21	8.3	5
1F6	Q15369	Transcription elongation factor B polypeptide 1	7.9-7.6	12466	17.74	4.74	2
3E5	P04406	Glyceraldehyde-3-phosphate dehydrogenase	7.0-6.7	35900	18.97	8.58	4
4D6	Q9Y6N5	Sulfide quinine oxidoreductase mitochondrial precursor	6.7-6.4	49929	11.62	9.18	4
5G2	Q06830	Peroxiredoxin 1 (Thioredoxin peroxidase 2)	6.1-5.8	22097	29.35	8.27	6
8C3	O95881	Thioredoxin domain containing protein 12 precursor	4.9-4.6	19194	37.98	5.25	5
9A9	Q99729	Heterogeneous nuclear ribonucleoprotein A/B	4.6-4.3	36590	9.15	9.04	3
11D5	Q8NC51	Plasminogen activator inhibitor 1 RNA-binding protein	IPA wash	44539	18.43	8.66	5

Table 3-3: Numbers of samples with reactivity above the cutoff (cutoff=highest signal in the non-cancer group) against recombinant proteins in the cancer group compared to 3 non-cancer groups. The numbers in the parentheses are the percentage of the positive reactors in the cancer category i.e. sensitivity at 100% specificity.

Recombinant proteins	Pair of sample groups		
	Cancer vs Normal	Cancer vs. Pancreatitis	Cancer vs. Diabetes
PGK1	10 (20.8)	10 (20.8)	12 (25)
PGK1 or MPH1	18 (37.5)	12 (25)	19 (39.6)
PGK1 or MPH1 or ARFIP2	22 (45.8)	14 (29.2)	21 (43.8)

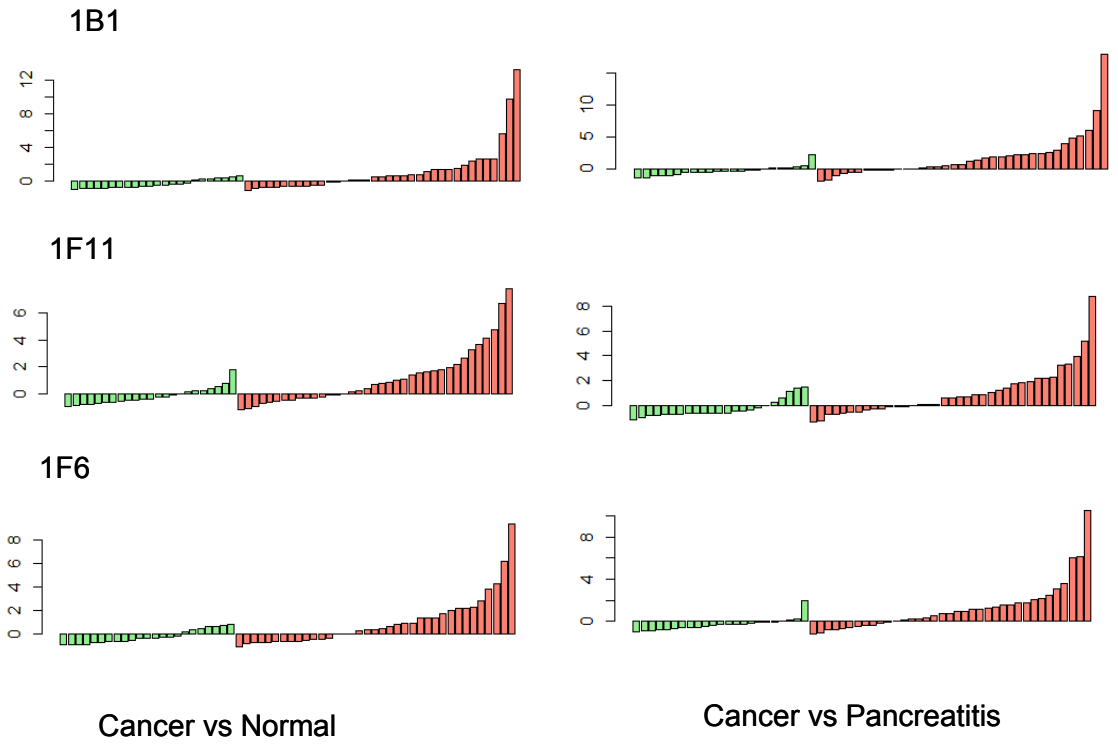


Figure 3-1: Colored bar graphs of three fractions found responded exclusively to some cancer sera in both pairwise comparisons between cancer versus normal and cancer versus pancreatitis. The y-axis is the normalized microarray signal for each sample.

(Reprinted from Li et al. The identification of auto-antibodies in pancreatic cancer patient sera using a naturally fractionated Panc-1 cell line. Copyright (2010), with permission from IOS Press)

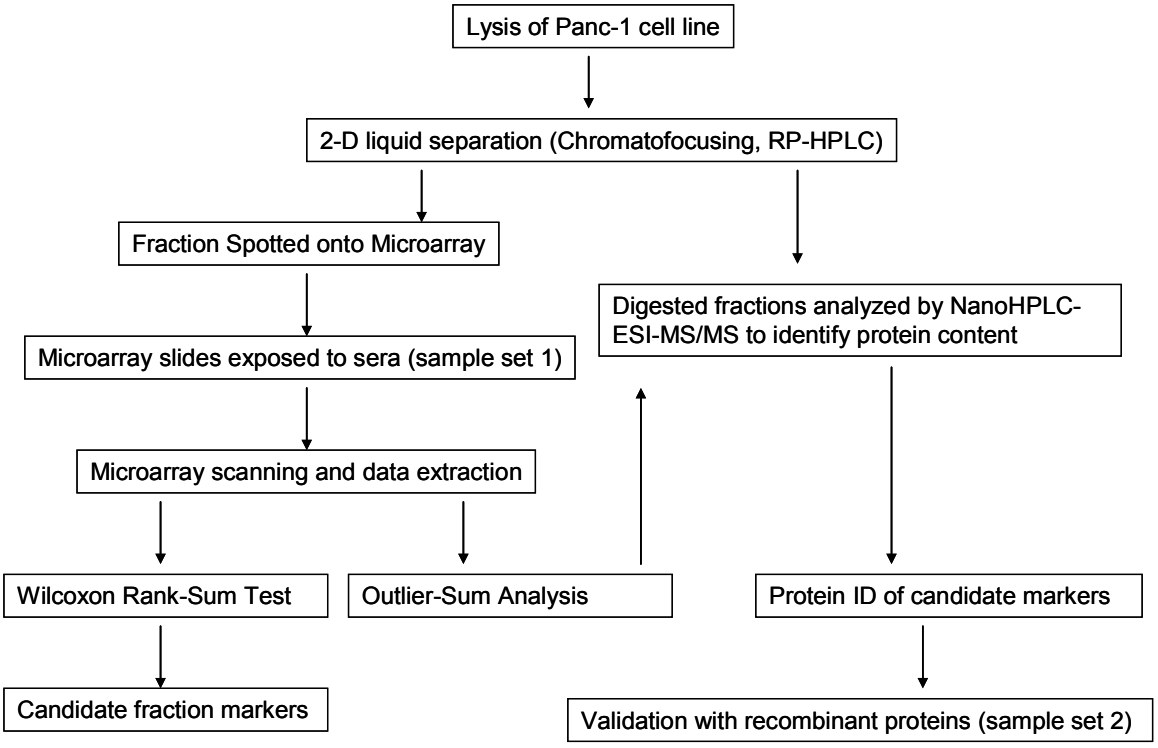
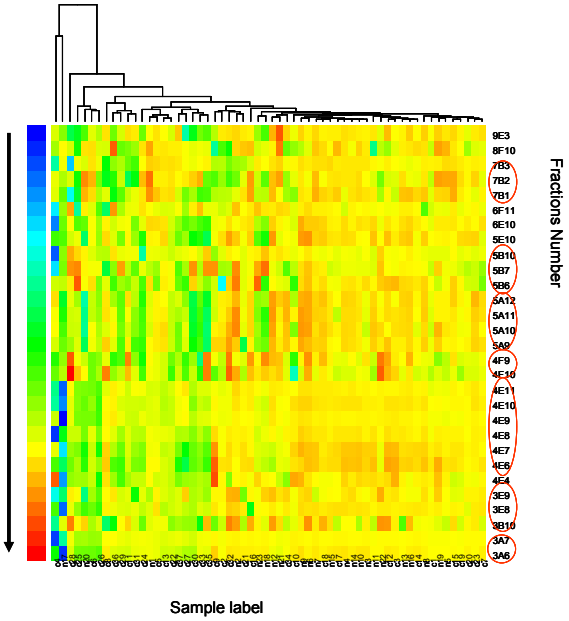


Figure 3-2: Flowchart of the experiment.

(Reprinted from Li et al. The identification of auto-antibodies in pancreatic cancer patient sera using a naturally fractionated Panc-1 cell line. Copyright (2010), with permission from IOS Press)

A) Cancer versus normal



B) Cancer versus pancreatitis

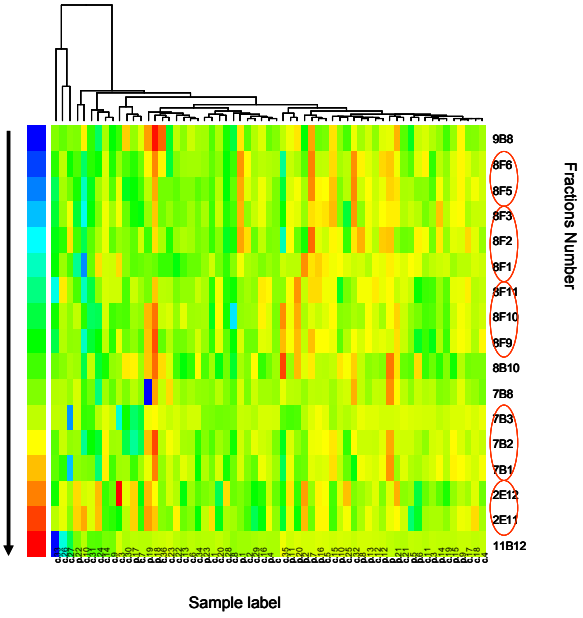


Figure 3-3: Heatmap with dendrogram of the microarray data. The colors of the bands indicate the normalized intensities of the microarray signal of the fractions. A total of 46 fractions with a p-value less than 0.02 in the pairwise comparisons using Wilcoxon rank-sum tests are shown in the heatmaps. The fractions appearing with their adjacent ones are red-circled. ((Reprinted from Li et al. The identification of auto-antibodies in pancreatic cancer patient sera using a naturally fractionated Panc-1 cell line. Copyright (2010), with permission from IOS Press)

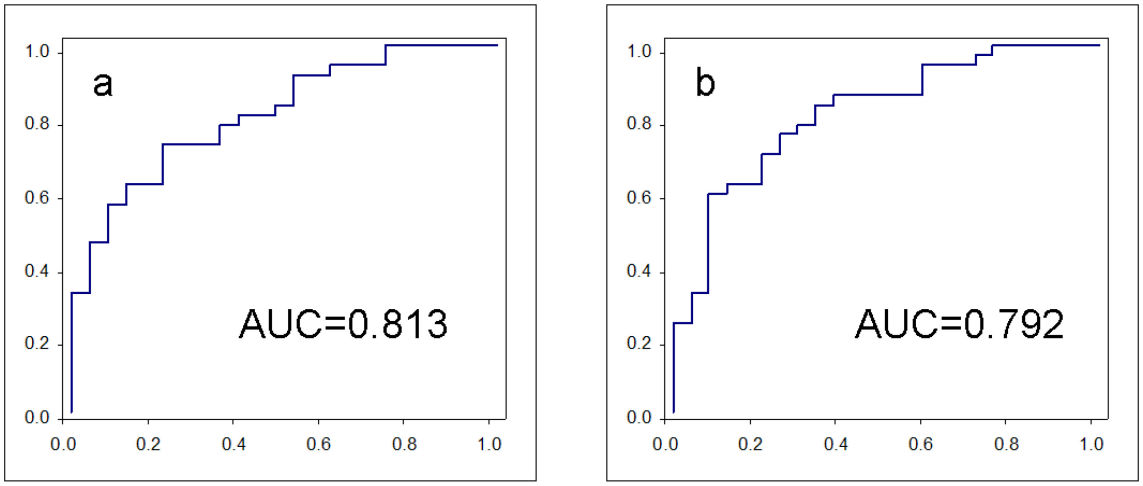


Figure 3-4: Combined ROC curves for the top-ranked fractions with lowest correlations in the Wilcoxon rank-sum tests a) cancer versus normal; b) cancer versus pancreatitis.

(Reprinted from Li et al. The identification of auto-antibodies in pancreatic cancer patient sera using a naturally fractionated Panc-1 cell line. Copyright (2010), with permission from IOS Press)

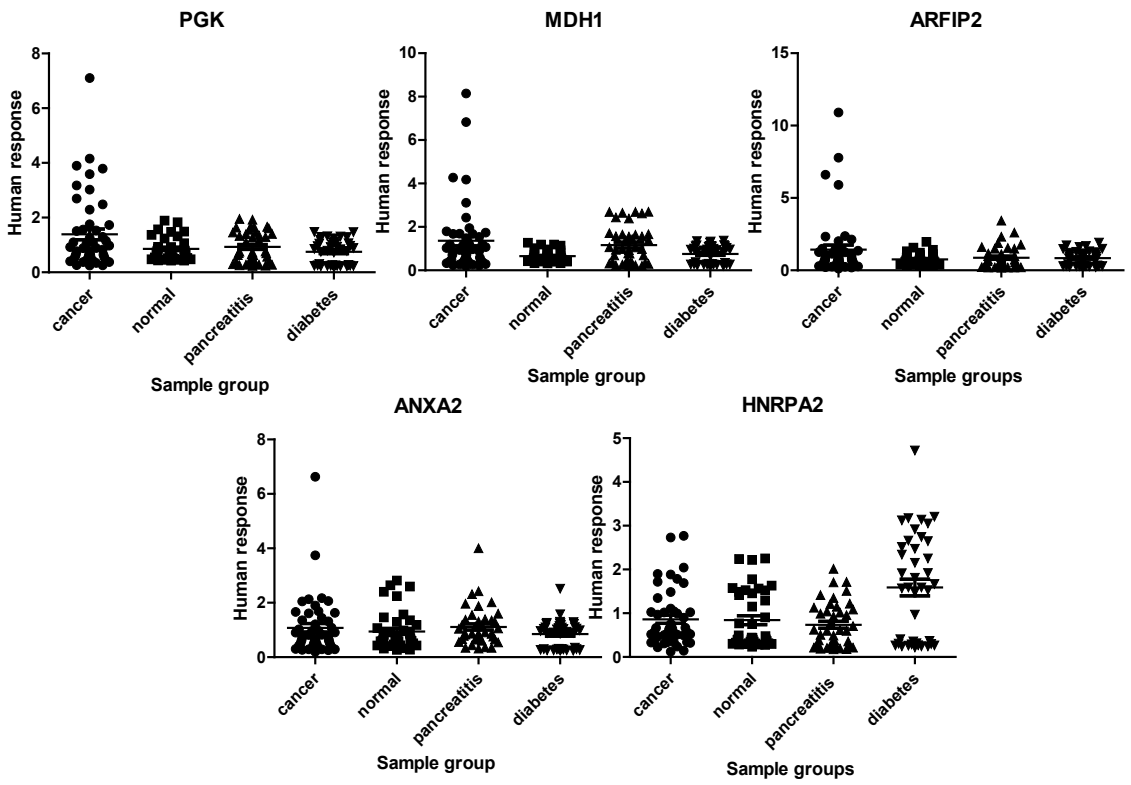


Figure 3-5: Distribution of the level of reactivity of five biomarker candidates examined in the confirmation experiment. The y axis represents the normalized fluorescent intensities of the autoantibody capture by the five recombinant proteins. The x axis represents 4 different disease groups: cancer, normal, pancreatitis and type 2 diabetes.

(Reprinted from Li et al. The identification of auto-antibodies in pancreatic cancer patient sera using a naturally fractionated Panc-1 cell line. Copyright (2010), with permission from IOS Press)

CHAPTER 4

A COMPARATIVE PHOSPHOPROTEOMIC ANALYSIS OF A HUMAN TUMOR METASTASIS MODEL USING A LABEL-FREE QUANTITATIVE APPROACH[2]

4.1 INTRODUCTION

Breast cancer is by far the most frequent cancer of women, with an estimated 192,370 new cases and 40,170 deaths in the United States in 2009[72]. The majority of cancer mortality is attributed to metastasis, which is the spread of tumor cells to a secondary site such as bone, lung, and liver. The multistep nature of metastasis poses difficulties in both design and interpretation of experiments to unveil the mechanisms causing the process. Studies on excised fixed human tissues are complicated by the variance of genetic background between individuals and by the cellular heterogeneity of a complex tissue mass [106]. Through in vivo selection of monoclonal cultures of the MDA-MB-435 breast tumor cell line we were able to characterize a pair of subclones (M-4A4 and NM-2C5) which differ in their ability to complete the metastatic process [107-109]. When orthotopically inoculated into athymic mice, both cell lines form primary tumors, but only M-4A4 is capable of metastasis to the lungs and lymph nodes. These cell lines constitute a valuable model for the study of cancer metastasis.

[2] Reprinted from Xie, X., S. Feng, H. Vuong, Y. Liu, S. Goodison, et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons

M-4A4 and NM-2C5 have been extensively compared using gene and protein expression analysis identifying a panel of differentially expressed genes and protein [106-108, 110-111]. However, because protein phosphorylation-mediated signaling networks regulate much of the cellular response to external stimuli, and dysregulation in these networks has been linked to multiple disease states including cancer [112], similar studies at the phosphoprotein level may add valuable biological insight to inhibit the metastatic process.

Although significant advances have been made over the past decade to enable the analysis and quantification of cellular protein phosphorylation events, comprehensive quantitative analysis of the phosphoproteome is still lacking. Several mass spectrometry (MS) -based quantification methods have been implemented for phosphoproteomics, including stable-isotope labeling through chemical modification of peptides with, for example, isobaric tags for relative and absolute quantitation (iTRAQ) [113] and stable-isotope labeling of amino acids in cell culture (SILAC) [114]. The well known limitations of label based-methods include increased complexity of the experimental protocols and the high cost of reagents.

In recent years, label-free quantitation methods have received increased attention as promising alternatives that automatically avoid some of the disadvantages of using stable isotope labeling methods. One approach is based on calculating extracted ion chromatogram ratios of peptides from separate LC-MS experiments and often includes an additional normalization step. Furthermore, the simple and straightforward spectral counting approach, in which total numbers of acquired MS/MS (MS2) spectra assigned to peptides are used as a read-out, transforms the frequency by which a peptide is identified

into a measure for peptide abundance. Spectral counts of peptides associated with a protein are then averaged into a protein abundance index [115]. This approach was recently employed as a semi-quantitative measure of phosphoprotein abundance [116-117]. Although conceptually simple, recent studies have demonstrated that spectral counting can be as sensitive as ion peak intensities in terms of detection range while retaining linearity [118].

Despite published examples of using spectral counting in quantitative phosphoproteomics, there are still challenges. In a relatively large-scale phosphorylation study, especially for phosphoserine and phosphothreonine, there is frequent and often overwhelming domination of phosphorylation-specific neutral losses (NL) in MS2 spectra. These peaks reduce the intensity of backbone b- and y-type ions that are critical for both phosphopeptide identification and precise site localization. To address this issue, a new data-dependent neutral loss (DDNL) MS/MS/MS (MS3) method that consists of additional fragmentation of the product of the precursor neutral loss in the form of a MS3 scan has been introduced. This approach (MS2 + MS3 scan) has now been widely adopted for phosphorylation identification analysis and is especially used on low mass accuracy mass spectrometers [119-120]. However, this strategy requires additional cycle time on the instrument and therefore reduces the number of spectra that can be measured in the same amount of time, so that the spectral counting method is often employed using the MS2-only scan. Sequence and phosphorylation site assignments are often manually validated for the phosphopeptides reported by the MS2 scan, where spectra are checked for the presence of neutral loss peak(s), coverage of the phosphorylation site by b- and y-

ions and alternative phosphorylation sites in the sequence matching the same spectrum [121]. This process is time-consuming and laborious.

In this chapter, we present a survey of phosphorylation profiles for an isogenic pair of human breast cancer cell lines and describe a general integrated framework for quantifying enriched phosphoproteins in the two cell lines by combining automatic validation of the MS2 + MS3 scan for phosphopeptide identification, with a subsequent MS2-only scan for spectral counting. The regulated phosphorylated peptides and sites identified by MS2 scan were validated by the MS2 + MS3 neutral loss method. Application of the label-free approach to this source material revealed a panel of differentially expressed phosphoproteins which implicate specific signaling pathways as being associated with distinct cellular phenotypes.

4.2 MATERIALS AND METHODS

4.2.1 Materials

Titanium dioxide (TiO_2) (3µm, 300Å) and zirconium dioxide (ZrO_2) (3µm, 300Å) were purchased from Glygen (Glygen, Columbia, MD). Protease inhibitor cocktail and phosphatase inhibitor cocktail were from Roche (Roche, Nutley, NJ). Sequencing grade modified trypsin was from Promega (Promega, Madison, WI). All other chemicals were from Sigma (Sigma, St. Louis, MO).

4.2.2 Cell culture

Human tumor cell lines M-4A4 and NM-2C5 were derived from the tumor cell line MDA-MB-435 as described previously [107-108]. Cell lines were maintained as subconfluent monolayer cultures in RPMI 1640 medium (Gibco-BRL, New York, NY)

supplemented with 10% fetal calf serum at 37°C under 5% CO_2 / 95% air. Cell lines were maintained in parallel cultures and harvested using non-trypsin cell dissociation as cultures reached ~75% confluency. Harvested cells were washed once in serum-free media and immediately snap-frozen in liquid nitrogen.

4.2.3 Enrichment of phosphopeptides

TiO_2 and ZrO_2 particles were pretreated with 30% ACN, 0.1% TFA and 50% ACN, 10% acetic acid (HAC) respectively by vortex for 15 min. After centrifuging at 25, 000 g for 5 min, the supernatant was discarded. The pellet was then treated with 100% ACN. Then TiO_2 and ZrO_2 beads were diluted as 20 mg/mL in 30% ACN, 0.1% TFA and 50% ACN, 10% HAC separately.

To make cell extracts, a lysis buffer (7 M urea, 2 M thiourea, 10% glycerol, 2% n-octyl G-D-glucopyranoside (OG), 100 mM DTT, protease inhibitor cocktail, phosphatase inhibitor cocktail) was directly added to frozen cell pellets and the lysates were vibrated at room temperature for 1 hr. Cellular debris and other insoluble materials were removed by centrifuging the mixture at 80, 000 g for 1 hr. After measuring protein concentration in each lysate, proteins were digested with trypsin with a ratio of 50/1 (w/w) overnight at 37°C.

For enrichment with TiO_2, 100 μg tryptic digest of the lysate was incubated with 50 μL TiO_2 beads (20 mg/mL). After incubation for 30 min with vibration, the TiO_2 beads were first washed with 300 μL 50% ACN, 6% TFA solution, followed by 300 μL 30% ACN, 0.1% TFA solution twice. The bound peptides were eluted with 100 μL 10% NH_4OH. After centrifugation, the supernatant was collected and lyophilized to dryness.

For enrichment with ZrO_2, 100 μg protein digests were diluted with 50% ACN, 10% HAC. The sample solution was mixed with 50 μL ZrO_2 beads suspension (20 mg/mL). The protocol for preparation of standard protein mixture digest was the same as the one used with TiO_2. The resulting solution was incubated for 30 min at room temperature. Then the ZrO_2 beads were firstly washed with 300 μL 50% ACN, 10% HAC solution, followed by two washes with 300 μL 10% HAC. The trapped phosphopeptides on ZrO_2 beads were eluted using 100 μL NH_4OH under sonication for 20 min. After centrifugation, the supernatant was collected and lyophilized to dryness.

4.2.4 Mass spectrometry

Dry peptides were suspended in 0.1% formic acid and loaded for LC-MS/MS analysis in a LTQ mass spectrometer. The nano-RPLC column (Nano Trap Column 5 μm 200Å Magic C18AQ 100 μm × 150 mm, Michrom Bioresources, Auburn, CA) was directly coupled to a LTQ linear IT MS from Thermo Scientific with a nanospray source. The LTQ instrument was operated in positive ion mode. The scan range of each full MS scan was *m*/*z* 400–2000. ACN gradients of 5–35% for 70 min at a flow rate 300 nL/min were applied for the separation of phosphopeptides. For the detection, the MS was set as a full scan followed by three data dependent MS2 events. For MS2 + MS3 scan, a subsequent MS3 event was triggered upon detection when a neutral loss of -49 or -32.7 (loss of H_3PO_4 for the +2 and +3 charged ions, respectively) was detected among the top 10 most intense ions in MS2. A dynamic exclusion window was applied which prevented the same *m*/*z* from being selected for 1 min after its acquisition. This entire LC-MS*n* system was controlled under Xcalibur software 2.0 (Thermo Scientific, Waltham, MA).

The MS2 and MS3 spectra were searched using SEQUEST (v0.27) against human IPI database v3.49 with the following parameters: peptide mass tolerance, 1.5 Da; MS2 and MS3 fragment ion mass tolerance, 1.4 Da; enzyme set as trypsin and allowance up to two missed cleavages; no static modification; dynamic modifications were methionine oxidation (+16 Da), phosphorylation on serine, threonine, and tyrosine (+80 Da); for MS3 data, besides the above modifications, variable modifications of -18 Da (elimination of phosphoric acid) on serine and threonine residues were also selected.

4.2.5 MS2 + MS3 scan data analysis

Enriched phosphopeptides were identified with automatic cross-validation of MS2 and MS3 spectra using the method of Jiang and coworkers [122]. Because the charge state of the precursor ion cannot be determined with low mass accuracy MS, more than one DTA file with different precursor charge states (commonly 2+ and 3+, respectively) were exported for one tandem spectrum. By combining MS2 spectra and corresponding neutral loss MS3, charge states of precursor ions can be determined from the m/z value of neutral loss: -49 indicated +2 charged precursor ions, while -32.7 was for +3 charged ions. Only a DTA spectrum with neutral loss peak of at least 50% of the base peak in intensity was considered. After removal of MS2/MS3 pairs with incorrect charge states, MS2 with no MS3, and MS2/MS3 pairs with neutral loss intensity less than 50% of the base peak in MS2 spectrum, the remaining MS2 and MS3 DTA spectra with specific precursor charge states were searched against the database, respectively. The top 10 hit peptides from a database search for a spectrum were considered. Then peptide identifications from a pair of spectra (MS2 and its corresponding MS3) were combined. Only peptides which were identified from both of the spectra (MS2 and MS3) were retained. The matched peptide

in a spectra pair with the highest Xcorr's score was defined as the top matched peptide for the spectra pair and selected for filter afterward.

For the determination of phosphorylation sites, Tscore was introduced as the sum of MS2 and MS3 PTM scores with the definition as -10 log*P(*total). For the phosphopeptide with two or more phosphorylation sites, Tscores of all candidate sequences with different phosphorylation site combinations for this phosphopeptide were calculated. Then the Tscore of a given site was computed by summing the Tscores of all candidate sequences containing this site. Phosphorylation sites with top *n* (equal to the number of possible phosphorylation sites) Tscores were considered as the most likely phosphorylation site localizations [122].

4.2.6 Spectral counting analysis

We counted the number of spectra observed for each peptide sequence in a mass spectrometry run [123]. To calculate a protein spectrum count, we summed the numbers for all of the peptides assigned to each protein in that run. We found this approach preferable to other methods such as parent ion peak height because it allowed us to simplify the analysis by combining all sites on a given protein [116]. Then we applied a normalized spectral abundance factor (NSAF) approach [124] to quantify phosphoprotein expression profiles. This approach, taking protein length as a normalization method, is an improvement over spectral counting methods in which protein abundance is quantitatively estimated based on the number of spectra acquired by MS2-only scan. The NSAF approach has at least the same, or better, capability to capture a wide dynamic range of protein expression ratios, and it can also identify significantly expressed proteins via simple statistical tests, such as the *t*-test, to compare the mean protein intensities of

two or more samples. The use of the *t*-test is applicable in this approach because it has been shown that the log transformation of the NSAF value is normally distributed. In addition, the NSAF approach has comparable sensitivity in identifying differentially expressed proteins as other approaches based on protein ratios.

After MS2-only scan analysis, the RAW data file was processed using SEQUEST and validated by Trans-Proteomic Pipeline (TPP). Spectral count data was then extracted from xml files using an in-house Perl script and output into Microsoft Excel files. In order to calculate the NSAF value, we applied the formula

$$(NSAF)_k = \frac{\left(\frac{Spc}{L}\right)_k}{\sum_{i=1}^{n}\left(\frac{Spc}{L}\right)_i}$$

where Spc is the number of spectral count, and L is the length in amino acid for k^{th} protein. The NSAF value was then natural log-transformed and subjected to independent two sample *t*-test using Microsoft Excel. A *t*-test *p* value of less than 0.05 was used to identify significant differentially expressed phosphoproteins.

4.2.7 Western blotting

Western blotting was performed using established methods [125] to confirm the phosphoprotein expression of isoform A of Lamin-A/C (LMNA, phospho Ser22) and Ras GTPase-activating protein-binding protein 1 (G3BP1, phospho Ser232). Briefly, equal amounts of isolated proteins from M-4A4 and NM-2C5 cell lysates were separated by 12% SDS-PAGE and then transferred to PVDF membrane using Transblot (Bio-Rad). After blocking for 1 h, the membrane was probed with rabbit polyclonal antibody against human phospho Lamin-A/C (Cell Signaling Technology, Boston, MA) or phospho G3BP1 (Abcam, Cambridge, MA) diluted in 1:1000 overnight. After incubation with

peroxidase-conjugated goat anti-rabbit IgG secondary antibody (Abcam) for 1 h, immunoblots were visualized with an enhanced chemiluminescent method kit (GE Healthcare, Piscataway, NJ). Densitometric analysis was performed.

4.3 RESULTS

Our phosphorylation profiling approach combined phosphopeptide enrichment using TiO_2 and ZrO_2 particles, multistage MS for phosphopeptide identification, and label-free spectral counting for quantitation. Extracted proteins from the human breast cancer cell lines M-4A4 and NM-2C5 were digested with trypsin, and the phosphopeptides were enriched on TiO_2 or ZrO_2 particles. The resulting peptide mixtures were analyzed by online LTQ linear IT MS with two consecutive stages of fragmentation. Automatic cross-validation by combining consecutive stage mass spectrometry data and the target-decoy database searching strategy was used to identify phosphopeptides. Quantitation of phosphoproteins in the two cell lines was achieved by the spectral counting method, with MS2-only scan implemented successively after each MS2 + MS3 scan. Sequence and site assignments were validated with the MS2 + MS3 neutral loss method. Western blotting was used to validate the altered expression of differentially expressed phosphoproteins (Figure 4-1).

4.3.1 Phosphoproteome

To provide adequate coverage of the phosphoproteome, six replicates of MS2 + MS3 scans for each sample were analyzed. More than 6,700 phosphopeptides were detected in 24 LC MS analyses. The six MS runs had very similar counts of identified phosphopeptide, as shown in Figure 4-2A. Only the peptides detected 3 or more times

within 6 replicates were considered for further analysis. After filtering with the criteria Rank'm = 1, ΔCn'm ≥ 0.1, and the Xcorr's ≥ 3.7, our analysis identified 425 phosphorylation sites on 160 unique proteins with a FDR less than 3% using the described stringent criteria in the two cell line samples. Of these, 65 sites (15.3%) had not been reported in the PhosphoSite Plus database as of September, 2009 (Supplemental material table 1). The representative MS2 and MS3 spectra of identified peptide VLGpSEGEEEDEALpSPAK assigned to protein DNA ligase 1 is shown in Figure 4-3.

In the M-4A4 cell line, 328 phosphorylated sites on 263 unique phosphopeptides were identified, and in the NM-2C5 cell line 345 phosphorylated sites on 264 unique phosphopeptides were identified (Figure 4-2B). Of these, we determined the distribution between individually identified sites to be 284 phosphoserine (pS), 43 phosphothreonine (pT), and 1 phosphotyrosine (pY) sites in M-4A4 cells and 297 pS, 46 pT, and 2 pY sites in NM-2C5 cells. In the Hunter and Sefton classic study using phosphoamino acid analysis, a relative abundance of 90%, 10%, and 0.05% for pS , pT, and pY was observed in proliferative, non-cancerous human cells [126]. The distribution of pS, pT, and pY sites was 86.6%, 13.1%, and 0.3% for M-4A4 cells and 86.1%, 13.3%, and 0.6% for NM-2C5 cells, a distribution markedly similar to the estimated phosphorylated amino acid content in the previous study (Figure 4-2C).

We observed phosphorylation sites on a wide variety of proteins. Figure 4-2D shows a Gene Ontology (GO) analysis of the phosphoproteome of M-4A4 and NM-2C5 cell lines. Almost half of the phosphorylation events occurred on nuclear proteins, whereas only one-third of all proteins in the IPI database are assigned as nuclear by GO [127], indicating that phosphorylation in these cells preferentially occurs in nuclear

proteins. As expected, proteins annotated as extracellular were significantly underrepresented in the phosphoproteome. In addition, proteins annotated as mitochondrial by GO were underrepresented, as were plasma membrane proteins.

TiO_2 and ZrO_2 have often been used to enrich phosphopeptides because of their strong interaction between phosphate groups on target molecules. We found that these two enrichment methods were complementary in identifying phosphopeptides, with fifty to sixty percent of identical phosphopeptides being enriched by both TiO_2 and ZrO_2 (Figure 4-4A). Moreover, more selective isolation of singly-phosphorylated peptides was observed with ZrO_2 compared to TiO_2, whereas TiO_2 preferentially enriched multiply-phosphorylated peptides (Figure 4-4B).

4.3.2 Quantitative phosphoproteomics

While MS3 scans followed by each MS2 scan will interfere with the spectral counting of peptides, we developed an approach to address this problem whereby one MS2-only scan is run successively after each MS2 + MS3 scan of the same sample with different sample injections and different method files. MS2-only scans in each sample were run 6 times. Hierarchical clustering analysis was employed to evaluate reproducibility between the 6 MS2-only scans. Correlation factors calculated using the spectral count of peptides showed very similar results between different MS2-only scans in the same sample, thus, the spectral count method is applicable in label-free shotgun proteomics (Figure 4-5). The spectral count was generated from MS2 raw data after TPP analysis. Only the peptides detected for 3 or more times in six MS2 runs for each sample were further analyzed. After normalization for protein length, the changed ratios and *p* value calculated by Student's *t*-test for each protein were recorded. As a result, 33

regulated proteins were identified with a *p* value less than 0.05 from the two enrichment methods.

Peptide sequence and site assignments reported from MS2-only scan were validated by a MS2 + MS3 neutral loss method. Only the proteins identified both from MS2-only scans and MS2 +MS3 scans were retained for further analysis. After validation, over 70 phosphorylated sites on 27 proteins were found to be differentially expressed, and 3 of them were present in both of the enrichment experiments with the same change trend. These included neuroblast differentiation-associated protein (AHAK), myosin-IXb, and protein NDRG1. Among the proteins we observed, 16 proteins were expressed at higher levels and 11 proteins were underexpressed in M-4A4 cells compared with NM-2C5 cells. The up-regulated protein group included lamin A/C, G3BP1, protein NDRG1, and myosin-IXb, and the down-regulated group included AHNAK, eukaryotic translation initiation factor 5B (EIF5B), serine/threonine-protein kinase 10 (STK10), and prostaglandin E synthase 3 (PTGES3). To provide the foundation of integrating MS2 + MS3 scans and MS2-only scans, we showed that the neutral loss peak for the peptide AEEDEILNRpSPR assigned to the protein calnexin was detected from both the MS2-only scan (Figure 4-6A) and the MS2 + MS3 scan (Figure 4-6B and 6C). Nevertheless, for the remaining 6 proteins which were not validated by the MS2 + MS3 neutral loss method, we discovered the neutral loss peak neither from MS2-only scans nor MS2 + MS3 scans. Furthermore, we evaluated the recurrence of the peptide AEEDEILNRpSPR in different MS2 + MS3 and MS2-only scans in the same sample. The very similar retention time of this peptide in 3 different MS2 + MS3 scans and 3 different MS2-only scans strongly supports the utility of this integrated quantitative method (Figure 4-7).

Although an increasing number of phospho-specific antibodies are emerging, the availability is still limited. We found 2 commercially available phospho-specific antibodies for proteinsfrom our list of differentially expressed proteins. We used these antibodies in Western blotting to verify the spectral counting quantification of LMNA and G3BP1 Good agreement between the two methods was achieved (Figure 4-8).

4.4 DISCUSSION

In this study, we have developed a novel strategy to identify 425 phosphorylation sites on 160 unique proteins with a FDR less than 3% in isogenic human breast cancer cell lines M-4A4 and NM-2C5. The approach uses TiO_2 and ZrO_2 particle enrichment of phosphopeptides and automatic validation by combining consecutive stage mass spectrometry data and target-decoy database searching. We have demonstrated the comparative application of the strategy by identifying 27 phosphoproteins that were deemed to be differentially expressed in the tumor metastasis model through spectral counting with cross-confirmation of MS2 + MS3 and MS2-only scans.

Although spectral counting approaches have been employed for the quantitative measure of phosphoprotein abundance by several groups, especially for differential profiling of phosphotyrosine-containing proteins in cancer tissues and cells [116], a generalizable quantitative spectral counting method for enriched phosphoprotein analysis including phosphoserine and phosphothreonine-containing proteins has been lacking. A remaining challenge is that MS2-only scans are often used to implement spectral counting in label-free quantitative proteomics. Many large-scale phosphorylation studies rely solely on MS2 scan and report error rates < 1% FDR. In those cases, data filtering is

accomplished by using high mass accuracy data for the precursors and/or higher cutoff values on scores derived from searching algorithms. However, for low mass accuracy mass spectrometers, the manual or automatic validation using the combination of MS2 and MS3 scan is often undertaken to assure accurate peptide and phosphorylation site assignment [120, 122, 128]. These reports indicate that cross-validation of phophopeptide assignment by MS2 and MS3 scans result in the high confidence in identification [120]. Therefore, we developed an integrated quantitative method combining spectral counting of MS2-only scans and phosphopeptide identification achieved with MS2 + MS3 scans. The phosphopeptide sequence and site assignments reported from MS2-only scans were validated by the MS2 + MS3 neutral loss method. Only those proteins identified by both MS2 + MS3 scans and with a *p* value less than 0.05 in statistical analysis with spectral counts, after normalization in MS2-only scans, can be considered for further analysis. Finally, 27 phosphoproteins met the criteria out of 33 proteins, where the latter group was found to have a *p* value less than 0.05 by MS2-only scan. The pitfall of this approach is that the stringent criterion might fail to detect peptides for which MS3 are not triggered because of low intensity (or absent) neutral loss fragment ions, such as phosphotyrosine-containing peptides.

In this study, we used the integrated strategy to identify 27 phosphoproteins that were differentially expressed in human cells with distinct metastatic phenotypes. Of these, 16 were up-regulated and 11 were down-regulated in metastatic M-4A4 cells relative to non-metastatic NM-2C5 cells. The expression change of LMNA and G3BP1 reported from spectral counting was validated with good agreement by Western blotting. Using the Ingenuity Pathways Analysis (IPA) we observed that the majority of the differentially

expressed proteins were highly interconnected and belong to two major intracellular signaling pathways.

Twelve of the 27 identified phosphoproteins were revealed to be interconnected through one signaling pathway (Figure 4-9A). LMNA and G3BP1 are involved in this pathway. They are all connected, directly or indirectly through 3 signaling hub proteins, v-myc myelocytomatosis viral oncogene homolog (c-myc), interferon gamma (IFNG), and retinoic acid, a signal molecule involved in cellular differentiation and response to extracellular stimuli. These factors are well-known to influence the behavior of breast cancer cells through multiple mechanisms, including progression [129], inflammation [130], proliferation, and apoptosis [131]. In another regulatory pathway, another 13 of the identified phosphoproteins are interconnected directly, or indirectly, through hepatocyte nuclear factor 4 alpha (HNF4A) or transforming growth factor beta 1 (TGFB1) , a signaling molecule that controls proliferation, differentiation, and other functions in many cell types (Figure 4-9B). The interconnection between proteins identified in this study with known cancer-associated factors implies a role for these proteins in cancer progression or metastasis.

Lamins are components of the nuclear lamina, a fibrous layer on the nucleoplasmic side of the inner nuclear membrane, which is thought to provide a framework for the nuclear envelope and may also interact with chromatin. Functional analysis of phosphorylation sites in human lamin A indicates the phosphorylation of T19 (Threonine), S22, S403, and S404 in controlling lamin disassembly, nuclear transport and assembly [132]. In a pathogenesis study, lamin A phosphorylation was reported to be associated with myoblast activation and involved in the pathogenic mechanism of Emery-

Dreifuss muscular dystrophy and limb girdle muscular dystrophy 1B [133]. Furthermore, studies have demonstrated that lamin A Ser404 is a nuclear target of Akt phosphorylation in C2C12 cells and implicated Akt phosphorylation of lamin A in the correct function of the nuclear lamina (Figure 4-9A) [134]. The irregularity of the nuclear envelope, whose framework is supported by lamin, has been observed to significantly correlate with lymph node metastases in breast cancers [135]. These findings suggest that the phosphorylation of LMNA might play a role in the decoration of the nuclear envelope in the cancer cell during metastasis.

G3BP1 is an hnRNA-binding protein and an element of the Ras signal transduction pathway. It is a DNA-unwinding enzyme which can unwind partial RNA/DNA and RNA/RNA duplexes in an ATP-dependent fashion. It binds specifically to the Ras-GTPase-activating protein by associating with its SH3 domain. In quiescent cells, G3BP1 is hyperphosphorylated on serine residues, and this modification is essential for its activity. G3BP1 harbors a phosphorylation-dependent RNase activity which specifically cleaves the 3'-untranslated region of human c-myc mRNA (Figure 4-9A) [136]. C-myc is a multifunctional oncogene and it plays a role in cycle progression, apoptosis and cellular transformation. Its overexpression is found during progression and distant metastasis of hormone-treated breast cancer [137]. It is possible that (de)phosphorylation of G3BP1 regulates interaction with c-myc, supporting a role for differential phosphorylation of this protein in the metastatic process. In addition, the growth factor heregulin beta 1 stimulation of breast cancer cells promotes phosphorylation of G3BP1 and increased the association of G3BP1 with GTPase-activating protein, again suggesting a role for G3BP1 in cancer progression [138].

4.5 CONCLUSION

This study describes a novel comparative phosphorylation strategy and application of this analysis to a human cell line model of tumor metastasis. The model consists of a pair of monoclonal cell lines derived from the same tumor source, but that have opposite metastatic propensity in murine xenograft models. LC-MS/MS based spectral counting analysis leads to the reliable identification of altered phosphorylation events in cells of distinct phenotype. This label-free, relative quantification of the phosphoproteome of complex samples enabled us to find new connections between the ability of cancer cells to establish metastasis in distant organs and altered expression levels of specific phosphorylated proteins. Biological interpretation of our data suggests that the phosphorylation of isoform A of lamin A/C and GTPase activating protein binding protein 1 may be involved in the metastatic behavior of human breast caner. Further investigations using this strategy hold promise for elucidating mechanisms involved in tumor progression and identifying novel therapeutic targets for potentially ameliorating the fatal spread of disease.

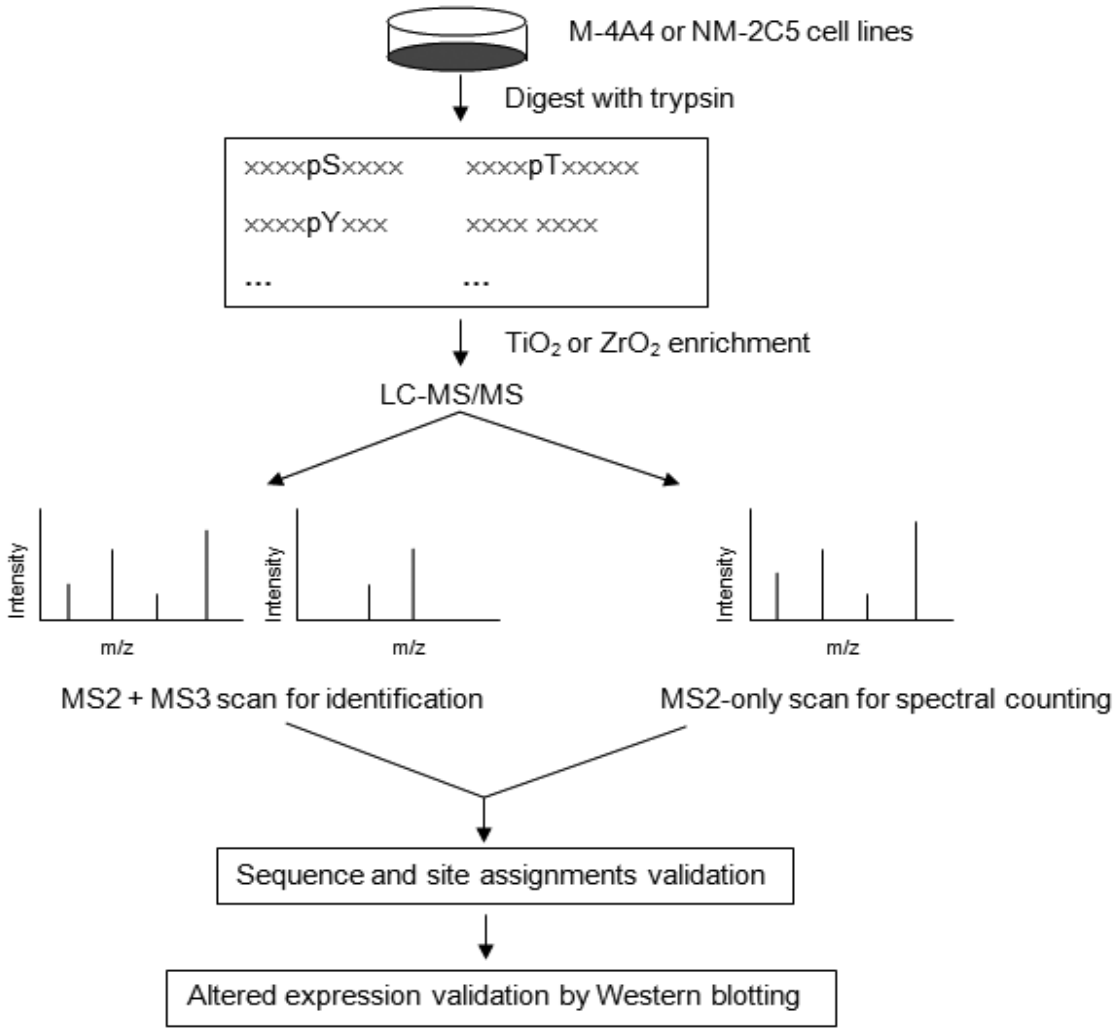
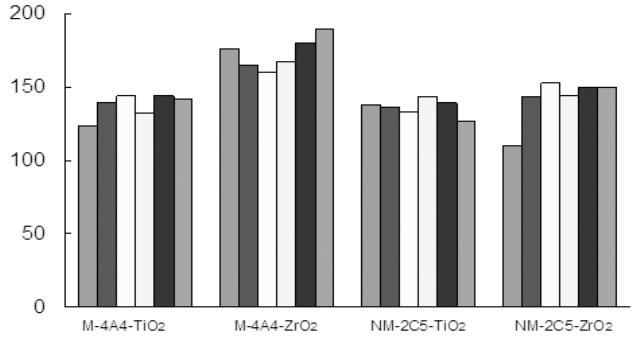


Figure 4-1: Schematic diagram illustrating the label-free quantitative analytical approach for protein phosphorylation profiling. Extracted proteins from human breast cancer cell lines, malignant M-4A4 and non-malignant NM-2C5, were digested with trypsin. Phosphopeptide mixtures with TiO_2 and ZrO_2 particles were analyzed by LC-MS/MS. Phosphopeptides were identified by automatic cross-validation of the MS2 + MS3 scan. Quantification of phosphoprotein expression in the tumor metastasis model was implemented by spectral counting with the MS2-only spectra. Then peptide sequence and site assignments were validated by MS2 + MS3 neutral loss method. The altered expression of selected phosphoproteins reported from spectral counting was validated by Western blotting. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons)

A



B

Cell lines	M-4A4			NM-2C5		
Enrichment method	TiO_2	ZrO_2	Total	TiO_2	ZrO_2	Total
Unique phosphopeptides	173	190	263	162	194	264
Phosphorylated sites	298	163	328	281	261	345

C

Cell lines	Phosphorylation sites		
	pSer	pThr	pTyr
M-4A4	284	43	1
NM-2C5	297	46	2

D

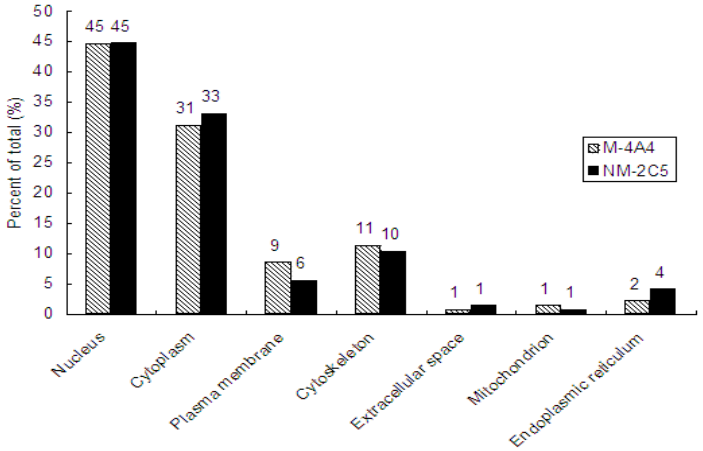


Figure 4-2: A. Summary of the phosphopeptide identification counts for M-4A4 or NM-2C5 cell lines with TiO2 or ZrO2 enrichment methods in six MS2 + MS3 replicate runs. B. Identified unique phosphopeptides and phosphorylation sites from 6 replicates of MS2 + MS3 scans in the M-4A4 and NM-2C5 cell lines; C. Distribution of phosphorylated sites by amino acid; D. Gene ontology analysis on cellular component of unique phosphoproteins. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons.

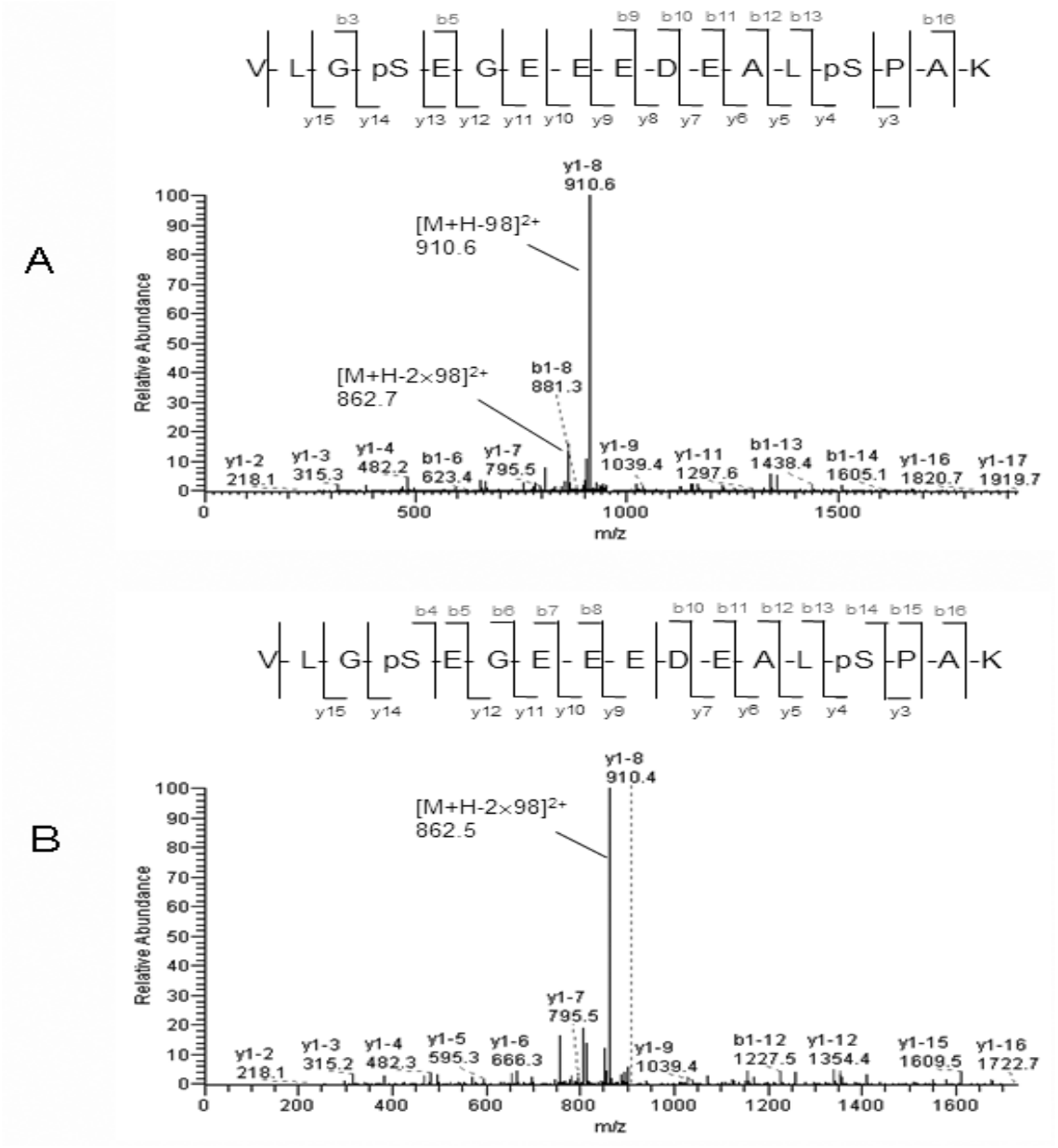


Figure 4-3: Representative spectra of a phosphorylated peptide identified by automatic cross-validation of MS2 and MS3 data-dependent neutral loss method. A. MS/MS spectrum of doubly charged and doubly phosphorylated peptide VLGpSEGEEEDEAL pSPAK assigned to protein DNA ligase 1. The peak at m/z 910.6 represents the doubly charged form of the selected precursor ion at *m/z* 959.8 with loss of one H_3PO_4 group, and at *m/z* 862.7 with loss of two H_3PO_4 groups. B. MS/MS/MS spectrum for the fragment ion at *m/z* 910.6 corresponding to neutral loss of the H_3PO_4 group. The peak at *m/z* 862.5 represents the doubly charged form of *m/z* 910.6 after loss of the H_3PO_4 group. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons.

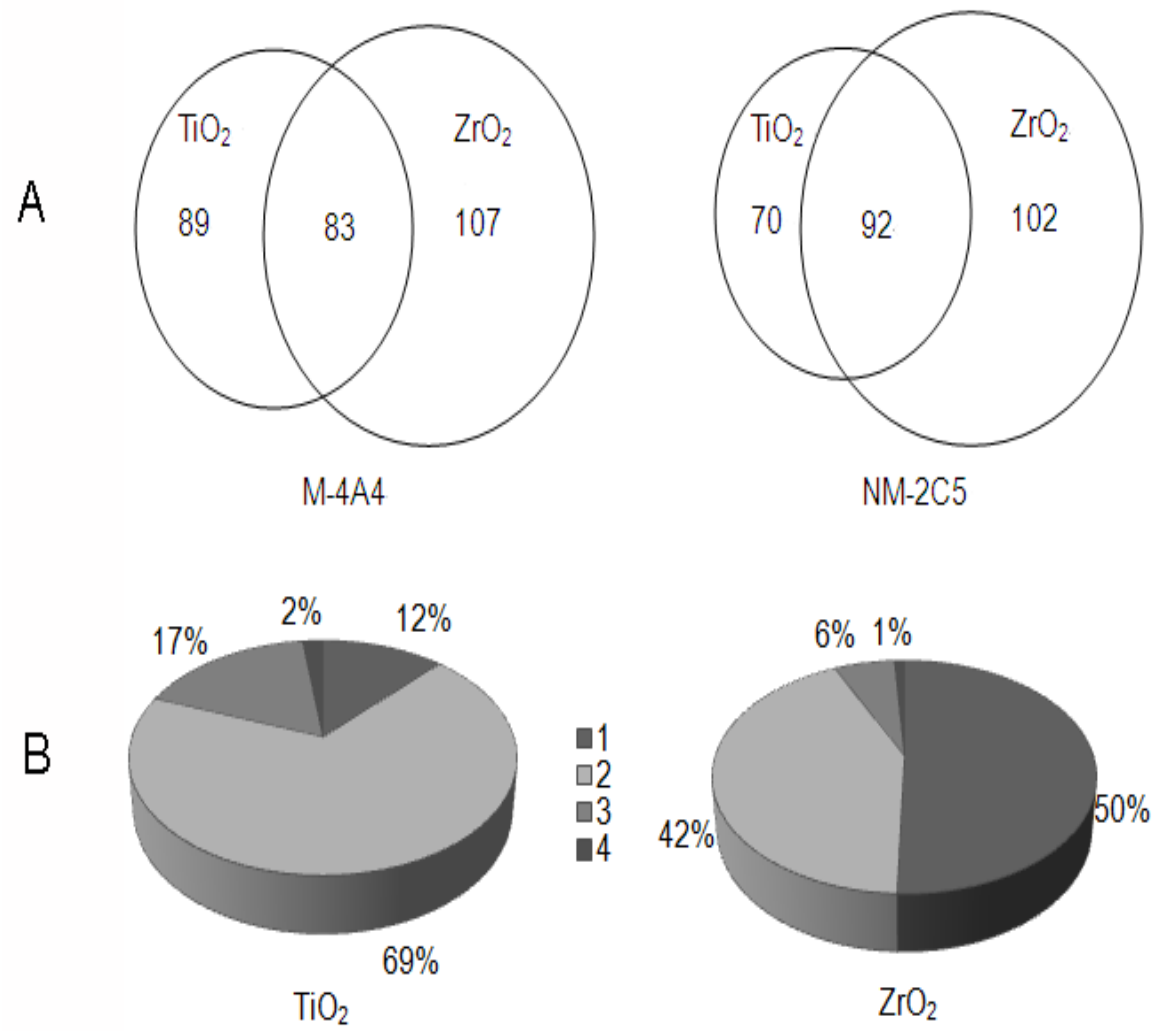


Figure 4-4: A. Overlap between phosphopeptide enrichment methods on the level of identified unique phosphopeptides in the M-4A4 and NM-2C5 cell lines. Around 50% of the unique phosphopeptides identified from the TiO2 method also identified in the ZrO_2 method; B. Distribution of singly (1), doubly (2), triply (3) and quadruply (4) phosphorylated peptides. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons.

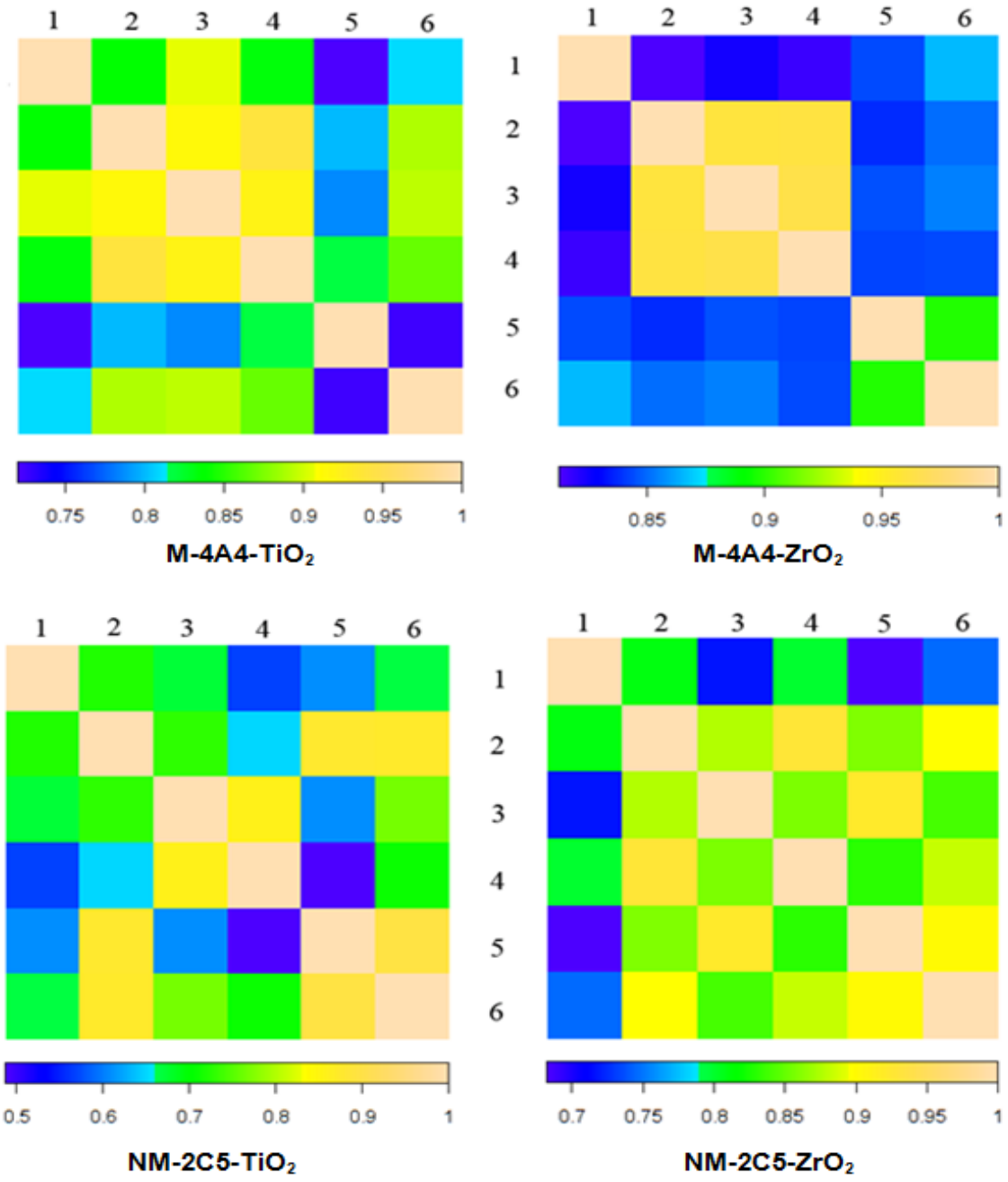


Figure 4-5: Hierarchical clustering of 6 MS2-only scans in the M-4A4 or NM-2C5 cells with TiO2 or ZrO2 particle enrichment. After MS2-only scan analysis, the RAW data file was processed using SEQUEST and validated by TPP. Spectral count data was then extracted from xml files. Correlation factors were calculated with spectral count of proteins. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons.

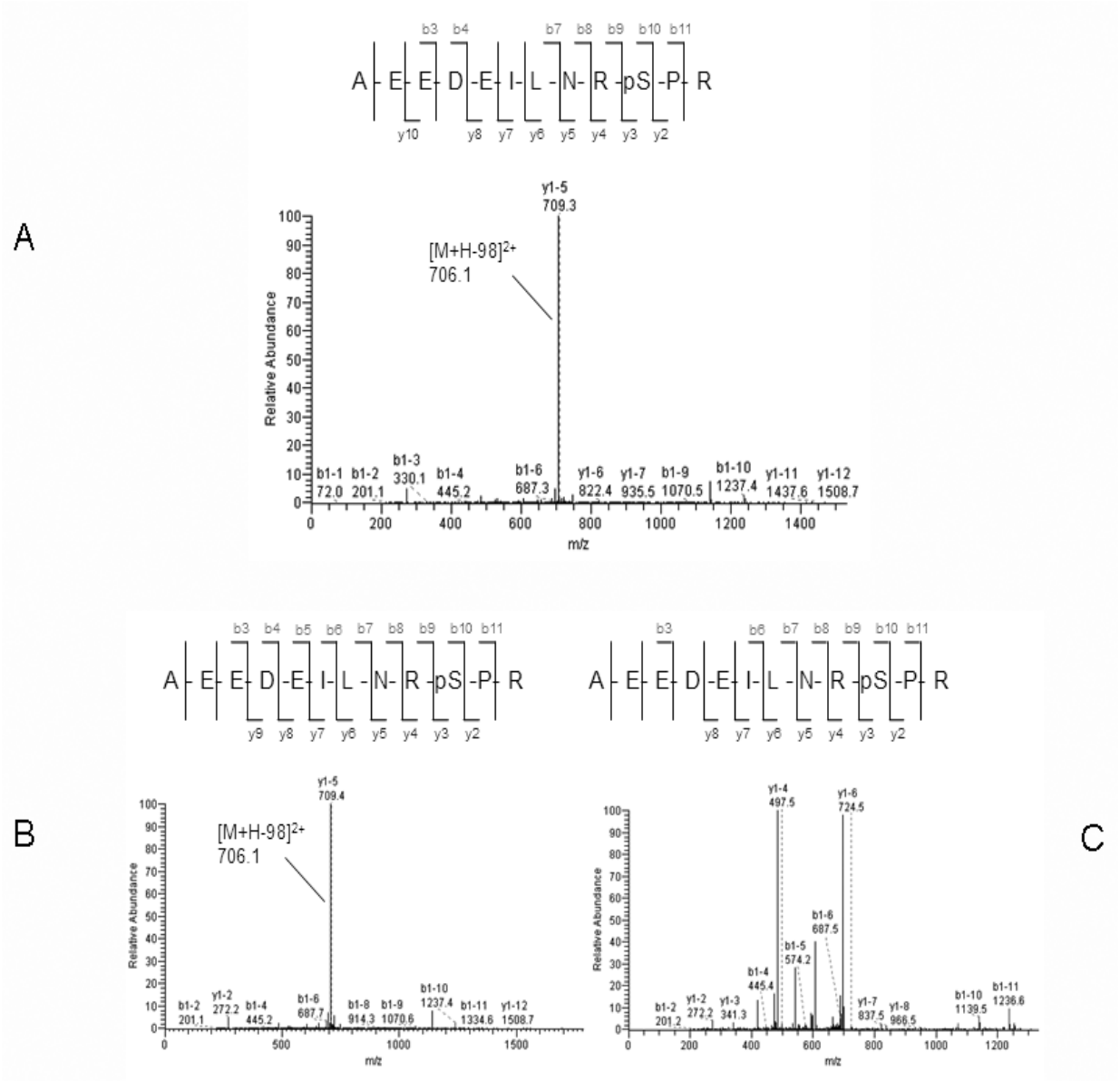


Figure 4-6: Representative spectra of singly phosphorylated peptide AEEDEILNRpSPR assigned to protein calnexin (CANX) identified in MS2-only scan (A) and in MS2 + MS3 scan (B and C). A. In MS2-only scan, MS/MS spectrum shows that the peak at m/z 706.1 represents the doubly charged form of the selected precursor ion at *m/z* 754.4 with loss of one H_3PO_4 group. C. In MS2 + MS3 scan, MS/MS/MS spectrum for the fragment ion at *m/z* 706.1 corresponding to neutral loss of the H_3PO_4 group. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons.

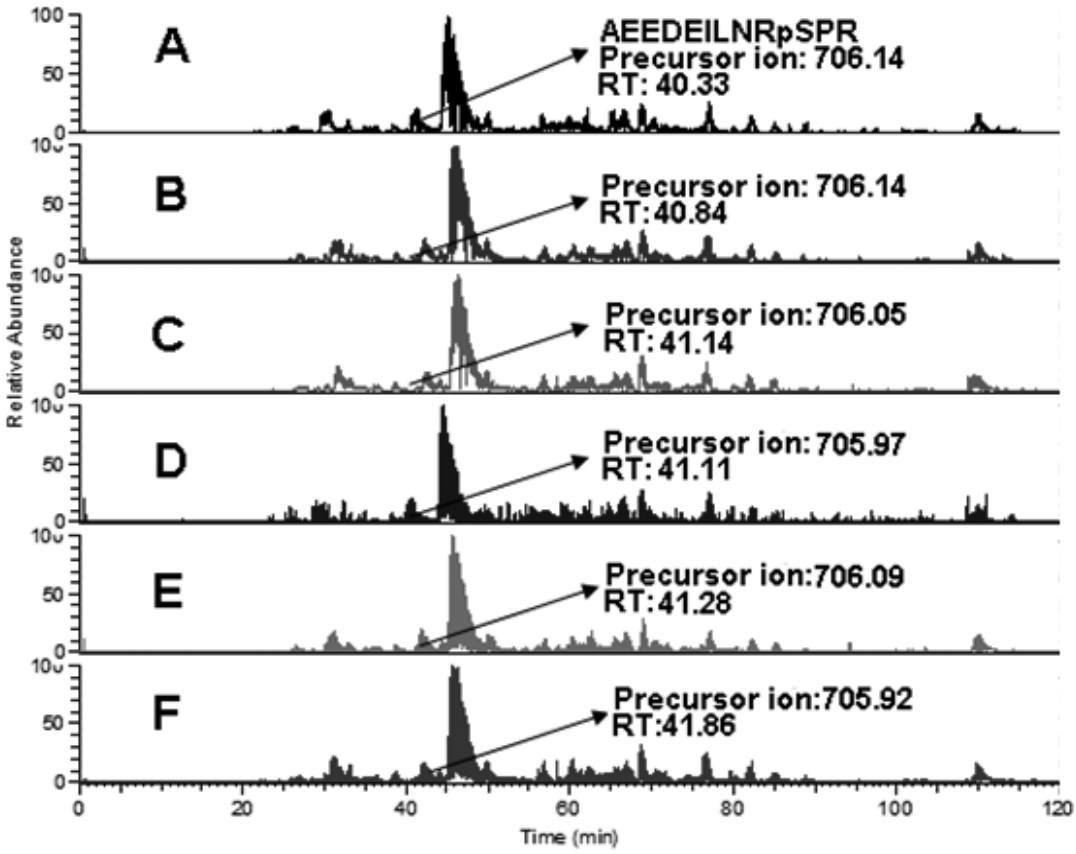
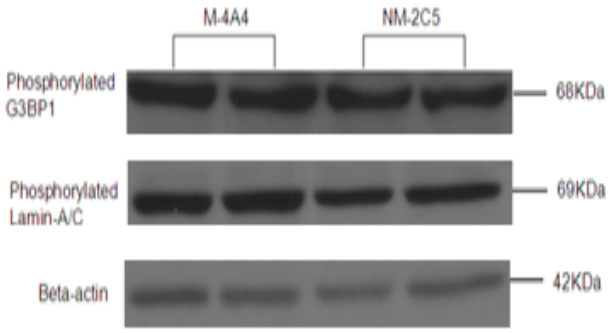


Figure 4-7: Recurrence of the same peptide in different MS2-only and MS2 + MS3 scans in the same sample. Shown is the precursor ion m/z and retention time (RT) of one identified peptide AEEDEILNRpSPR in 3 different MS2-only scans (A, B, C) and 3 different MS2 + MS3 scans (D,E,F) in the NM-2C5 cells enriched with ZrO_2. The very similar retention time strongly supports the utility of integrated quantitative method. The arrow indicates the spectrum where the peptide was identified. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons.



M-4A4						NM-2C5					
5	4	3	4	4	6	2	3	3	3	2	1
14	19	15	22	10	21	9	8	11	12	7	9

Western blotting **Spectral counting**

Figure 4-8: Western blotting analysis of the expression of phosphoproteins LMNA and G3BP1 in M-4A4 and NM-2C5 cell lines. Show right is a comparison of the six replicates of spectral counts of phosphoproteins identified by MS/MS with the immunoblotting. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons

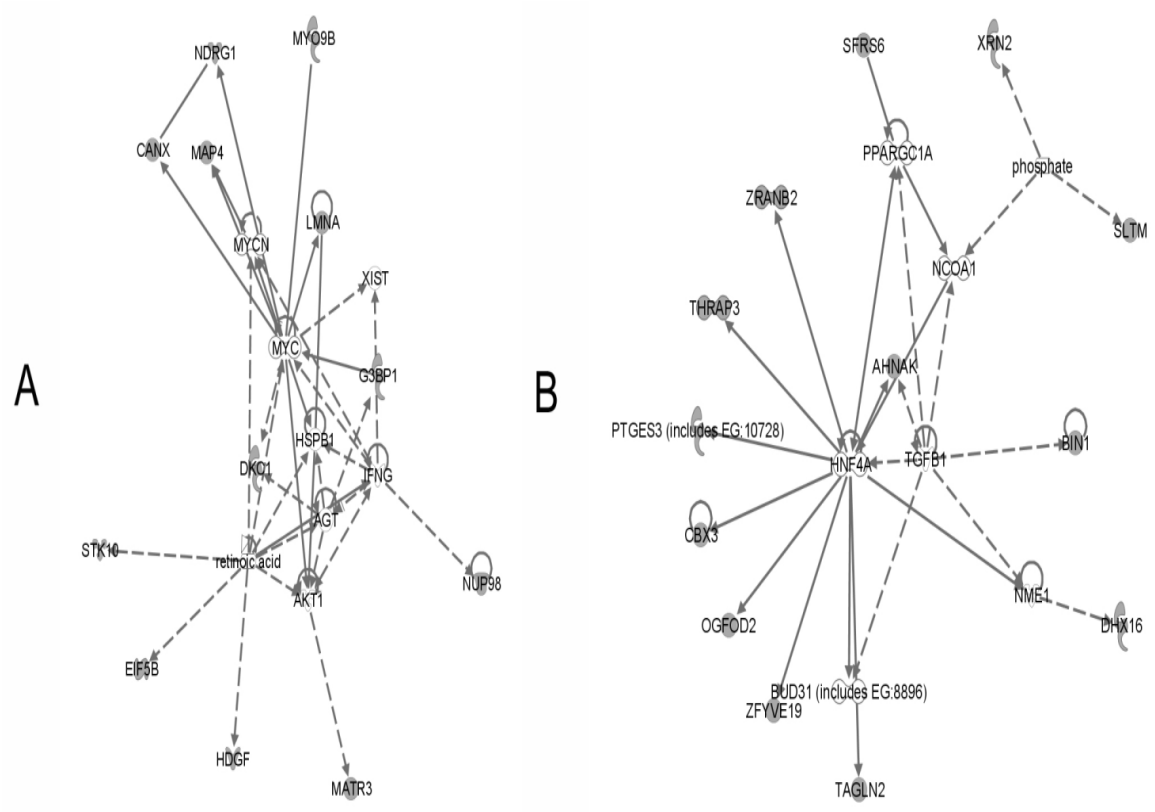


Figure 4-9: Pathway network models 1 (A) and 2 (B). A solid line indicates direct interaction; a dashed line indicates indirect interaction; a line without arrowhead indicates binding; an arrow from protein A to protein B indicates A acts on B. Node shapes are indicative: triangle, kinase; diamond, enzyme; hexagon, translation regulator; trapezoid, transporter; oval (horizontal), transcription regulator; oval (vertical), transmembrane receptor. Proteins identified in this screen are marked in shadow. Reprinted from Xie et al. (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label free quantitative approach. Electrophoresis. 31(11): 1842-1852. Copyright (2010), with permission from John Wiley and Sons.

CHAPTER 5

CONCLUSION AND FUTURE STUDY

5.1 CONCLUSION

This dissertation describes an attempt to integrate proteomics and bioinformatics approaches to mine the human proteome for biomarkers for early detection of cancer. The high complexity of human serum proteome, which is comprised of approximately 100,000 proteins whose concentrations span over 12 orders of magnitude, makes the identification of low abundant biomarkers a major challenge. Several studies have demonstrated that the humoral immune response is the most promising approach to discover potential serum biomarkers for the diagnosis and prognosis of early stage cancers. Moreover, it is part of an ongoing national effort by the National Cancer Institute which has brought together many institutions through its Early Detection Research Network program to develop and discover promising biomarkers and technologies for early detection of cancer.

Two main proteomics approaches are described in the dissertation to identify novel biomarkers in cancers. The first approach involves the use of a natural liquid-based protein microarray platform to study the humoral immune response against tumor antigens in melanoma and pancreatic cancer as described in chapter 2 and 3 respectively. My primary contribution to this dissertation is the statistical analysis of data generated from this protein microarray approach. We presented in Chapter 2 a study comparing

outlier based to traditional mean based approaches in differential expression analysis with applications in protein microarray data in heterogeneous diseases. In this chapter, my work described an attempt to identify differentially expressed proteins as potential biomarkers for cancer early detection. My work demonstrated that an outlier-based approach to differential expression analysis is an alternative approach to the traditional mean-based approach such as the Student t-test. The outlier based approach could be used to analyze the humoral immune response data which is commonly regarded as "heterogeneous" data, i.e. the differential expression is concentrated in the tail regions of the data distribution instead of spreading throughout the distribution. In a simulation based testing of performance between the outlier sum test and Student t-test, we showed that the outlier based analysis has improved power for detection of differentially expressed proteins in the humoral immune response studies that have small sample size, small effect size and large deviation from the normal distribution. While outlier-based approaches offer the potential to extract useful information from studies that yield minimally interesting results from conventional methods, these biomarkers are by definition limited in their predictive power in an unselected population. As illustrated in the melanoma study, even markers that are identified using traditional approaches such as the Student t-method may turn out to have a heterogeneous pattern of differential expression. Thus, we anticipate that while outlier-oriented statistics like the OS statistic may play a useful role, especially in larger studies, another important consequence of these efforts will result from the more widespread adoption of methods to characterize the detailed pattern of differential expression of candidate biomarkers identified through traditional approaches.

Chapter 3 of the dissertation presents a study harnessing the potential of humoral immune response to identify antoantibody-based biomarkers for the early detection of pancreatic cancers. The dataset is obtained from the protein microarray experiment and contains signal intensities of more than 1000 protein fractions extracted and fractionated from a Panc-1 cell line. In this chapter, my work demonstrated the utility of the outlier based approach to differential expression analysis using a large proteomics dataset from a study of humoral immune response in pancreatic cancer versus pancreatitis and normal controls. Here, our group used 2D-liquid separation and protein microarray with my evaluation of two statistical analyses including the Wilcoxon rank sum test and outlier sum test to identify proteins that elicited a differential humoral response pattern between different clinical groups that could be used for further validation. My work combined these two methods together to identify the differentially expressed proteins with improved detection power. I was able to demonstrate how combining the outlier-based approach (e.g. outlier sum test) and the mean-based approach (e.g. Wilcoxon rank sum test) represent a more satisfactory data analysis approach than each approach alone in addressing the difficult data analysis question of how to extract information from heterogeneous response data for biomarker discovery. This idea is not related to meta-analysis in which multiple results of the same question or parameter from several independent (or at least weakly dependent) studies are combined to achieve a higher statistical power to detect an effect than in a single study. However, it can be argued that the strongly dependent results from my combined statistical approach would improve the detection power as well as statistical confidence in the context of biomarker discovery.

In chapter 3, the rationale for using a non-parametric test such as Wilcoxon rank sum test is that we do not know for certain the underlying distribution of the humoral immune response data (e.g. the intensity readouts from the binding between antigens and auto-antibodies). Compared to the parametric tests such as t-test and F-test, non-parametric test only assumes a bare minimum on the underlying data model, namely the observations are independent and the variable under study has underlying continuity. An advantage of using non-parametric tests is that the conclusions are more generalized than those inferred from parametric tests. However, the pitfall of non-parametric tests is in term of statistical power, parametric tests are superior, precisely because of the strength of their assumptions. In chapter 3, the combined statistical analyses of Wilcoxon rank sum test and outlier sum test yielded a total of 9 potential protein biomarkers that can distinguish pancreatic cancer from non-cancer. Based on my analysis, our group chose 5 recombinant proteins for further validation using an independent sample set and confirmed that 3 out of 5 recombinant proteins had significant differential expression in pancreatic cancer versus non-cancer. Further work will require experiments with early stage cancer sera which currently are not available in sufficient numbers for these experiments. Also, further work on the use of truncated or modified forms of these proteins may provide an improved response. In addition, as larger numbers of samples are collected, larger validation sets can be run to test the validity of these potential targets for biomarker response. On the bioinformatics analysis side, it may be interesting to extend the usability of outlier based analysis to other types of mass spectrometry data such as spectral count data. For example, outlier test may be used to

identify outlying peptides with abnormal spectral counts. These peptides are then excluded from protein abundance calculation.

The second approach to identify biomarkers described in the dissertation is based on liquid chromatography mass spectrometry (LC-MS/MS). Compared with the humoral immune response approach, this approach has the advantages of investigating a variety of post-translational modifications including phosphorylation which has been shown to be a major mechanism in tumorigenesis. In the MS-based approach, two different model cell lines are analyzed using a liquid based mass spectrometry (LC-MS/MS) coupled with two dimensional separation of protein fractions based on their pI and hydrophobicity. Then, a label-free, spectral count method was applied for the quantitation of specific proteins to identify differentially expressed proteins with respect to the cellular phenotype. In chapter 4, we used a comparative MS-based approach to survey phosphorylation profiles for an isogenic pair of human breast cancer cell lines. We also described a general integrated framework for quantifying enriched phosphoproteins in the two cell lines by combining automatic validation of the MS2+MS3 scan for phosphopeptide identification, with a subsequent MS2-only scan for spectral counting. My specific contribution to this project is the spectral counting analysis on MS-MS (MS2) data. Label-free quantitation methods using spectral count have received increased attention as promising alternatives that avoid some of the disadvantages of using stable isotope labeling methods. A difficult data analysis question related with phosphoproteins quantitation involved identification of significant changes in abundance level as many phosphoproteins are low abundance proteins, often have less than 10 spectral counts. For example, a two-fold change from 1 to 2 spectra will be less convincing to be a significant change than a two-

fold change from 10 to 20 spectra. To the best of our knowledge, there is no satisfactory statistical approach for this challenge yet. In this chapter, my work evaluated a label free quantitative analysis to determine the expression level changes of phosphoproteins involved in the metastasis of breast cancer. Specifically, my work described the enhancement of spectral counting as a quantitative proteomics tools in low abundance phosphoproteins as cancer biomarkers using standard approaches such as NSAF (normalized spectral abundance factor)[139]. The spectral count quantitative approach successfully identified 27 phosphoproteins as being differentially expressed with respect to tumor cell phenotype. In addition, our group showed that there is good agreement between the spectral counting quantification and Western blotting in two commercially available phospho-specific antibodies for 2 differentially expressed phosphoproteins (LMNA and G3BP1). Taken together, my work demonstrated the effectiveness of NSAF as a quantitative method for low abundance phosphoproteins as cancer biomarkers.

Collectively, our results represent an exploration of bioinformatics and proteomics approaches in biomarker discovery in early detection of cancers. Along with a list of potential biomarkers discovered in melanoma, pancreatic and breast cancer, our results provide new insight into disease pathogenesis and biomarker identification. Specifically, our results demonstrate great potential of the humoral immune response in tackling the early cancer detection challenges.

5.2 FUTURE STUDY

The ultimate goals of cancer early detection require methods with high sensitivity, dynamic range, throughput and multiplexing capability to mine the complex human

plasma/serum proteome for potential biomarkers. To accelerate the process from biomarker discovery to pre-validation, an integrated approach combining the power of proteomics and bioinformatics is needed. My research foci are biomarker discovery using approaches in bioinformatics and proteomics to harness the potential of humoral immune response. Therefore, the future study will focus on biomarker validation using a larger independent patient cohort. However, a challenge in biomarker discovery in proteomics study is the dependence on the availability of clinical samples as most cancer samples are rare and difficult to obtain. Therefore, we will use the resource and support from a large collaborative network such as the Early Detection Research Network which is critical for this endeavor.

As specified per the Early Detection Research Network (EDRN), acceptable cancer biomarkers are judged based on criteria such as high sensitivity and specificity, ability to obtain from non-invasive methods, reliability and repeatability in testing, and high predictive value for clinical disease. In the next step, we will develop statistical approaches to measure the predictive value such as ROC (receiver operating curve) analyses. Following the use of ROC analyses for measuring predictive power of biomarkers in distinguishing cancer samples to non-cancer samples compared with existing predictors, we will begin collecting clinical data to validate predictive biomarkers in early detection cases. These steps are corresponding to phase 2 and 3 of the EDRN's 5 phase approach to biomarker development for early detection.

As demonstrated in the recombinant protein validation experiment in chapter 3, one area of difficulty in humoral response experiments is the low signal intensity that is often present in the arrays. While differential responses are observed for certain potential

cancer protein markers the response overall is not remarkably high. This weak response could be a result of protein immobilization on the slide which renders the protein unable to move about such that binding sites are blocked from reagent molecules. It may be interesting to investigate whether reducing the protein size by chemical means may facilitate exposure of these binding sites thereby enhancing the overall sensitivity of the humoral response experiments.

Our understanding of the underlying mechasisms of humoral immuner response and its role in cancer are still limited yet it is critical to the development of mechanism specific therapeutic immunotherapy strategies. The level of complexity of the potentially relevant biological pathways in the immune response to cancer is too high to be easily understood without intelligent integration of multiple levels of biological information. Although the central dogma of molecular biology states that proteins are closer to actual biological functions of cells than mRNAs or DNAs[140], proteomics only represents one aspect of cancer biomarkers in addition to genetic biomarkers, and changes in the transcriptome do not always correlate with protein expression or activity[141-142]. Therefore, integrating gene expression data (transcriptome) with protein (proteome) and metabolites (metabolome) will provide a more complete picture of humoral immune response in cancer.

BIBLIOGRAPHY

1. Jemal A, Siegel R, Xu J, Ward E: **Cancer statistics, 2010**. *CA: a cancer journal for clinicians* 2010:caac. 20073v20071.
2. Jacobs IJ, Skates SJ, MacDonald N, Menon U, Rosenthal AN, Davies AP, Woolas R, Jeyarajah AR, Sibley K, Lowe DG: **Screening for ovarian cancer: a pilot randomised controlled trial**. *The Lancet* 1999, **353**(9160):1207-1210.
3. Srinivas PR, Kramer BS, Srivastava S: **Trends in biomarker research for cancer detection**. *The lancet oncology* 2001, **2**(11):698-704.
4. DeLeo AB, Jay G, Appella E, Dubois GC, Law LW, Old LJ: **Detection of a transformation-related antigen in chemically induced sarcomas and other transformed cells of the mouse**. *Proceedings of the National Academy of Sciences of the United States of America* 1979, **76**(5):2420.
5. Castelli C, Rodolfo M, Luksch R, Stockert E, Chen YT, Collini P, Ranzani T, Lombardo C, Dalerba P, Rivoltini L: **Antigen specific immunity in neuroblastoma patients: Antibody and T cell recognition of NY-ESO-1 tumor antigen**. *Proceedings of the American Association for Cancer Research* 2004, **2004**(1):1261.
6. Disis ML, Knutson KL, Schiffman K, Rinn K, McNeel DG: **Pre-existent immunity to the HER-2/neu oncogenic protein in patients with HER-2/neu overexpressing breast and ovarian cancer**. *Breast cancer research and treatment* 2000, **62**(3):245-252.
7. Litvak DA, Gupta RK, Yee R, Wanek LA, Ye W, Morton DL: **Endogenous immune response to early-and intermediate-stage melanoma is correlated with outcomes and is independent of locoregional relapse and standard prognostic factors* 1**. *Journal of the American College of Surgeons* 2004, **198**(1):27-35.
8. Stockert E, Jäger E, Chen Y-T, Scanlan MJ, Gout I, Karbach J, Arand M, Knuth A, Old LJ: **A Survey of the Humoral Immune Response of Cancer Patients to a Panel of Human Tumor Antigens**. *The Journal of Experimental Medicine* 1998, **187**(8):1349-1354.
9. Anderson KS, Sibani S, Wallstrom G, Qiu J, Mendoza EA, Raphael J, Hainsworth E, Montor WR, Wong J, Park JG *et al*: **Protein Microarray Signature of Autoantibody Biomarkers for the Early Detection of Breast Cancer**. *Journal of Proteome Research* 2010, **10**(1):85-96.
10. **Frequently Asked Questions** [http://edrn.nci.nih.gov/advocates/frequently-asked-questions]
11. Hanash SM, Pitteri SJ, Faca VM: **Mining the plasma proteome for cancer biomarkers**. *Nature* 2008, **452**(7187):571-579.
12. Anderson L, Anderson NG: **High resolution two-dimensional electrophoresis of human plasma proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5421.
13. Hanash S, Taguchi A: **The grand challenge to decipher the cancer proteome**. *Nat Rev Cancer* 2010, **10**(9):652-660.
14. Beck G, Habicht GS: **Immunity and the invertebrates**. *Scientific American* 1996, **275**(5):60.

15. Lutzky J, Gonzalez-Angulo AM, Orzano JA: **Antibody-based vaccines for the treatment of melanoma**. *Seminars in oncology* 2002, **29**(5):462-470.
16. Montefiori D, Sattentau Q, Flores J, Esparza J, Mascola J: **Antibody-based HIV-1 vaccines: recent developments and future directions**. *PLoS medicine* 2007, **4**(12):e348.
17. Caron M, Choquet-Kastylevsky G, Joubert-Caron R: **Cancer immunomics using autoantibody signatures for biomarker discovery**. *Molecular & Cellular Proteomics* 2007, **6**(7):1115.
18. Shankaran V, Ikeda H, Bruce AT, White JM, Swanson PE, Old LJ, Schreiber RD: **IFN[gamma] and lymphocytes prevent primary tumour development and shape tumour immunogenicity**. *Nature* 2001, **410**(6832):1107-1111.
19. Hanash S: **Harnessing the Immune Response for Cancer Detection**. *Cancer Epidemiology Biomarkers & Prevention* 2011, **20**(4):569-570.
20. Hirohashi Y, Torigoe T, Inoda S, Takahashi A, Morita R, Nishizawa S, Tamura Y, Suzuki H, Toyota M, Sato N: **Immune response against tumor antigens expressed on human cancer stem-like cells/tumor-initiating cells**. *Immunotherapy* 2010, **2**(2):201-211.
21. Sundaresh S, Doolan DL, Hirst S, Mu Y, Unal B, Davies DH, Felgner PL, Baldi P: **Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques**. *Bioinformatics* 2006, **22**(14):1760-1766.
22. Stafford P, Johnston S: **Microarray technology displays the complexities of the humoral immune response**. *Expert Review of Molecular Diagnostics* 2011, **11**(1):5-8.
23. Anderson KS, LaBaer J: **The Sentinel Within: Exploiting the Immune System for Cancer Biomarkers†**. *Journal of Proteome Research* 2005, **4**(4):1123-1133.
24. Desmetz C, Cortijo C, Mangé A, Solassol J: **Humoral response to cancer as a tool for biomarker discovery**. *Journal of Proteomics* 2009, **72**(6):982-988.
25. Chen Y-T, Scanlan MJ, Sahin U, Türeci Ö, Gure AO, Tsang S, Williamson B, Stockert E, Pfreundschuh M, Old LJ: **A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening**. *Proceedings of the National Academy of Sciences* 1997, **94**(5):1914-1918.
26. Park S, Lim Y, Lee D, Cho B, Bang Y-J, Sung S, Kim H-Y, Kim D-K, Lee Y-S, Song Y *et al*: **Identification and characterization of a novel cancer/testis antigen gene CAGE-1**. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 2003, **1625**(2):173-182.
27. Pallasch CP, Struss A-K, Munnia A, König J, Steudel W-I, Fischer U, Meese E: **Autoantibodies against GLEA2 and PHF3 in glioblastoma: Tumor-associated autoantibodies correlated with prolonged survival**. *International Journal of Cancer* 2005, **117**(3):456-459.
28. Jongeneel V: **Towards a cancer immunome database**. *Cancer Immun* 2001, **1**(3).
29. Fosså A, Alsøe L, Crameri R, Funderud S, Gaudernack G, Smeland EB: **Serological cloning of cancer/testis antigens expressed in prostate cancer using cDNA phage surface display**. *Cancer Immunology, Immunotherapy* 2004, **53**(5):431-438.

30. Sioud M, Hansen MH: **Profiling the immune response in patients with breast cancer by phage displayed cDNA libraries**. *European Journal of Immunology* 2001, **31**(3):716-725.
31. Yang L, Guo S, Li Y, Zhou S, Tao S: **Protein microarrays for systems biology**. *Acta Biochimica et Biophysica Sinica* 2011, **43**(3):161-171.
32. Haab BB: **Advances in protein microarray technology for protein expression and interaction profiling**. *Current opinion in drug discovery & development* 2001, **4**(1):116.
33. Zhu H, Snyder M: **Protein chip technology**. *Current opinion in chemical biology* 2003, **7**(1):55-63.
34. Bertone P, Snyder M: **Advances in functional protein microarray technology**. *Febs Journal* 2005, **272**(21):5400-5411.
35. Patwa T, Li C, Simeone DM, Lubman DM: **Glycoprotein analysis using protein microarrays and mass spectrometry**. *Mass spectrometry reviews* 2010, **29**(5):830-844.
36. Patwa TH, Zhao J, Anderson MA, Simeone DM, Lubman DM: **Screening of glycosylation patterns in serum using natural glycoprotein microarrays and multi-lectin fluorescence detection**. *Analytical chemistry* 2006, **78**(18):6411-6421.
37. Lueking A, Possling A, Huber O, Beveridge A, Horn M, Eickhoff H, Schuchardt J, Lehrach H, Cahill DJ: **A nonredundant human protein chip for antibody screening and serum profiling**. *Molecular & Cellular Proteomics* 2003, **2**(12):1342.
38. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R: **Global analysis of protein phosphorylation in yeast**. *Nature* 2005, **438**(7068):679-684.
39. Chen CS, Korobkova E, Chen H, Zhu J, Jian X, Tao SC, He C, Zhu H: **A proteome chip approach reveals new DNA damage recognition activities in Escherichia coli**. *Nature methods* 2007, **5**(1):69-74.
40. Ramachandran N, Anderson KS, Raphael JV, Hainsworth E, Sibani S, Montor WR, Pacek M, Wong J, Eljanne M, Sanda MG: **Tracking humoral responses using self assembling protein microarrays**. *PROTEOMICS–Clinical Applications* 2008, **2**(10 11):1518-1527.
41. Sreekumar A, Nyati MK, Varambally S, Barrette TR, Ghosh D, Lawrence TS, Chinnaiyan AM: **Profiling of Cancer Cells Using Protein Microarrays**. *Cancer research* 2001, **61**(20):7585.
42. Haab BB: **Antibody arrays in cancer research**. *Molecular & Cellular Proteomics* 2005, **4**(4):377.
43. Paweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW, Emmert-Buck MR, Roth MJ, III EFP, Liotta LA: **Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front**. *gene* 2001, **20**:1981-1989.
44. Royce TE, Rozowsky JS, Luscombe NM, Emanuelsson O, Yu H, Zhu X, Snyder M, Gerstein MB: **Extrapolating traditional DNA microarray statistics to tiling and protein microarray technologies**. *Methods in enzymology* 2006, **411**:282.

45. Bolstad BM, Irizarry RA, Åstrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185.
46. Pollard HB, Srivastava M, Eidelman O, Jozwik C, Rothwell SW, Mueller GP, Jacobowitz DM, Darling T, Guggino WB, Wright J: **Protein microarray platforms for clinical proteomics**. *PROTEOMICS–Clinical Applications* 2007, **1**(9):934-952.
47. Huber W, Von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* 2002, **18**(suppl 1):S96.
48. Kreil DP, Karp NA, Lilley KS: **DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results**. *Bioinformatics* 2004, **20**(13):2026.
49. Luevano M, Bernard HU, Barrera-Saldaña HA, Trevino V, Garcia-Carranca A, Villa LL, Monk BJ, Tan X, Davies DH, Felgner PL: **High-throughput profiling of the humoral immune responses against thirteen human papillomavirus types by proteome microarrays**. *Virology* 2010, **405**(1):31-40.
50. Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments**. *Genome Biol* 2003, **4**(4):210.
51. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, **278**(5338):680.
52. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays**. *Nature genetics* 1999, **21**(1 Suppl):10-14.
53. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes**. *Bioinformatics* 2001, **17**(6):509-519.
54. Zhong L, Hidalgo GE, Stromberg AJ, Khattar NH, Jett JR, Hirschowitz EA: **Using protein microarray as a diagnostic assay for non-small cell lung cancer**. *American journal of respiratory and critical care medicine* 2005, **172**(10):1308.
55. Tibshirani R, Hastie T: **Outlier sums for differential gene expression analysis**. *Biostatistics* 2007, **8**(1):2-8.
56. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, Varambally S, Cao X, Tchinda J, Kuefer R *et al*: **Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer**. *Science* 2005, **310**(5748):644-648.
57. Wu B: **Cancer outlier differential gene expression detection**. *Biostatistics* 2007, **8**(3):566-575.
58. Lian H: **MOST: detecting cancer differential gene expression**. *Biostatistics* 2008, **9**(3):411-418.
59. Wang Y, Rekaya R: **LSOSS: Detection of Cancer Outlier Differential Gene Expression**. *Biomarker Insights* 2010, **5**:69.
60. Chen L-A, Chen D-T, Chan W: **The distribution-based p-value for the outlier sum in differential gene expression analysis**. *Biometrika* 2010, **97**(1):246-253.
61. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL: **Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene**. *Science* 1987, **235**(4785):177.

62. Tockman MS, Mulshine JL, Piantadosi S, Erozan YS, Gupta PK, Ruckdeschel JC, Taylor PR, Zhukov T, Zhou WH, Qiao YL: **Prospective detection of preclinical lung cancer: results from two studies of heterogeneous nuclear ribonucleoprotein A2/B1 overexpression**. *Clinical cancer research* 1997, **3**(12):2237.
63. Shah RB, Mehra R, Chinnaiyan AM, Shen R, Ghosh D, Zhou M, MacVicar GR, Varambally S, Harwood J, Bismar TA: **Androgen-independent prostate cancer is a heterogeneous group of diseases**. *Cancer research* 2004, **64**(24):9209.
64. Balch CM, Soong SJ, Gershenwald JE, Thompson JF, Reintgen DS, Cascinelli N, Urist M, McMasters KM, Ross MI, Kirkwood JM: **Prognostic factors analysis of 17,600 melanoma patients: validation of the American Joint Committee on Cancer melanoma staging system**. *Journal of Clinical Oncology* 2001, **19**(16):3622.
65. Liu Y, He J, Xie X, Su G, Teitz-Tennenbaum S, Sabel MS, Lubman DM: **Serum Autoantibody Profiling Using a Natural Glycoprotein Microarray for the Prognosis of Early Melanoma**. *Journal of Proteome Research* 2010, **9**(11):6044-6051.
66. Hu J: **Cancer outlier detection based on likelihood ratio test**. *Bioinformatics* 2008, **24**(19):2193-2199.
67. Cohen J: **A power primer**. *Psychological Bulletin* 1992, **112**(1):155-159.
68. Hedges LV, Olkin I: **Statistical Method for Meta-Analysis**: Orlando: Academic Press; 1985.
69. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control**. *Nature medicine* 2004, **10**(8):789-799.
70. Hanahan D: **The hallmarks of cancer**. *Cell* 2000, **100**(1):57-70.
71. Shackleton M, Quintana E, Fearon ER, Morrison SJ: **Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution**. *Cell* 2009, **138**(5):822-829.
72. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ: **Cancer statistics, 2009**. *CA: a cancer journal for clinicians* 2009, **59**(4):225-249.
73. Rosty C, Goggins M: **Early detection of pancreatic carcinoma**. *Hematology/oncology clinics of North America* 2002, **16**(1):37.
74. Garcea G, Neal C, Pattenden C, Steward W, Berry D: **Molecular prognostic markers in pancreatic cancer: a systematic review**. *European Journal of Cancer* 2005, **41**(15):2213-2236.
75. Tan EM: **Autoantibodies as reporters identifying aberrant cellular mechanisms in tumorigenesis**. *Journal of Clinical Investigation* 2001, **108**(10):1411-1416.
76. Casiano CA, Mediavilla-Varela M, Tan EM: **Tumor-associated antigen arrays for the serological diagnosis of cancer**. *Molecular & Cellular Proteomics* 2006, **5**(10):1745.
77. Desmetz C, Maudelonde T, Mangé A, Solassol J: **Identifying autoantibody signatures in cancer: a promising challenge**. *Expert Review of Proteomics* 2009, **6**(4):377-386.
78. Hall JC, Casciola-Rosen L, Rosen A: **Altered structure of autoantigens during apoptosis**. *Rheumatic diseases clinics of North America* 2004, **30**(3):455.

79. Pollard KM, Lee DK, Casiano CA, Bluthner M, Johnston MM, Tan EM: **The autoimmunity-inducing xenobiotic mercury interacts with the autoantigen fibrillarin and modifies its molecular and antigenic properties**. *The Journal of Immunology* 1997, **158**(7):3521.
80. Utz PJ, Anderson P: **Posttranslational protein modifications, apoptosis, and the bypass of tolerance to autoantigens**. *apoptosis*, **14**:19.
81. Rosen A, Casciola-Rosen L, Wigley F: **Role of metal-catalyzed oxidation reactions in the early pathogenesis of scleroderma**. *Current opinion in rheumatology* 1997, **9**(6):538.
82. Ben-Mahrez K, Thierry D, Sorokine I, Danna-Muller A, Kohiyama M: **Detection of circulating antibodies against c-myc protein in cancer patient sera**. *British Journal of Cancer* 1988, **57**(6):529.
83. Disis ML, Calenoff E, McLaughlin G, Murphy AE, Chen W, Groner B, Jeschke M, Lydon N, McGlynn E, Livingston RB: **Existent T-cell and antibody immunity to HER-2/neu protein in patients with breast cancer**. *Cancer research* 1994, **54**(1):16.
84. Disis ML, Cheever MA: **Oncogenic proteins as tumor antigens**. *Current opinion in immunology* 1996, **8**(5):637-642.
85. Soussi T: **p53 Antibodies in the sera of patients with various types of cancer: a review**. *Cancer research* 2000, **60**(7):1777.
86. Brichory FM, Misek DE, Yim AM, Krause MC, Giordano TJ, Beer DG, Hanash SM: **An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(17):9824.
87. Gnjatic S, Wheeler C, Ebner M, Ritter E, Murray A, Altorki NK, Ferrara CA, Hepburne-Scott H, Joyce S, Koopman J: **Seromic analysis of antibody responses in non-small cell lung cancer patients and healthy donors using conformational protein arrays**. *Journal of immunological methods* 2009, **341**(1-2):50.
88. Desmetz C, Bascoul-Mollevi C, Rochaix P, Lamy PJ, Kramar A, Rouanet P, Maudelonde T, Mangé A, Solassol J: **Identification of a new panel of serum autoantibodies associated with the presence of in situ carcinoma of the breast in younger women**. *Clinical cancer research* 2009, **15**(14):4733.
89. Looi KS, Nakayasu ES, Diaz RA, Tan EM, Almeida IC, Zhang JY: **Using proteomic approach to identify tumor-associated antigens as markers in hepatocellular carcinoma**. *Journal of Proteome Research* 2008, **7**(9):4004-4012.
90. Xia Q, Kong XT, Zhang GA, Hou XJ, Qiang H, Zhong RQ: **Proteomics-based identification of DEAD-box protein 48 as a novel autoantigen, a prospective serum marker for pancreatic cancer**. *Biochemical and biophysical research communications* 2005, **330**(2):526-532.
91. Hamanaka Y, Suehiro Y, Fukui M, Shikichi K, Imai K, Hinoda Y: **Circulating anti MUC1 IgG antibodies as a favorable prognostic factor for pancreatic cancer**. *International Journal of Cancer* 2003, **103**(1):97-100.

92. Kotera Y, Fontenot JD, Pecher G, Metzgar RS, Finn OJ: **Humoral immunity against a tandem repeat epitope of human mucin MUC-1 in sera from breast, pancreatic, and colon cancer patients**. *Cancer research* 1994, **54**(11):2856.
93. Raedle J, Oremek G, Welker M, Roth WK, Caspary WF, Zeuzem S: **p53 autoantibodies in patients with pancreatitis and pancreatic carcinoma**. *Pancreas* 1996, **13**(3):241.
94. Sanz-Ortega J, Steinberg S, Moro E, Saez M, Lopez J, Sierra E, Sanz-Esponera J, Merino M: **Comparative study of tumor angiogenesis and immunohistochemistry for p53, c-ErbB2, c-myc and EGFr as prognostic factors in gastric cancer**. *Histology and histopathology* 2000, **15**(2):455.
95. Yan F, Sreekumar A, Laxman B, Chinnaiyan AM, Lubman DM, Barder TJ: **Protein microarrays using liquid phase fractionation of cell lysates**. *Proteomics* 2003, **3**(7):1228-1235.
96. Patwa TH, Li C, Poisson LM, Kim HY, Pal M, Ghosh D, Simeone DM, Lubman DM: **The identification of phosphoglycerate kinase 1 and histone H4 autoantibodies in pancreatic cancer patient serum using a natural protein microarray**. *Electrophoresis* 2009, **30**(12):2215-2226.
97. Canelle L, Bousquet J, Pionneau C, Deneux L, Imam-Sghiouar N, Caron M, Joubert-Caron R: **An efficient proteomics-based approach for the screening of autoantibodies**. *Journal of immunological methods* 2005, **299**(1-2):77-89.
98. Tibshirani R, Hastie T: **Outlier sums for differential gene expression analysis**. *Biostatistics* 2007, **8**(1):2.
99. Lieber M, Mazzetta J, Nelson Rees W, Kaplan M, Todaro G: **Establishment of a continuous tumor cell line (PANC 1) from a human carcinoma of the exocrine pancreas**. *International Journal of Cancer* 1975, **15**(5):741-747.
100. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R: **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer**. *Science* 2005, **310**(5748):644.
101. Brahimi-Horn C, Pouysségur J: **The role of the hypoxia-inducible factor in tumor metabolism growth and invasion**. *Bull Cancer* 2006, **93**(8):E73-80.
102. Wang J, Dai J, Jung Y, Wei CL, Wang Y, Havens AM, Hogg PJ, Keller ET, Pienta KJ: **A glycolytic mechanism regulating an angiogenic switch in prostate cancer**. *Cancer research* 2007, **67**(1):149.
103. Kurayoshi M, Oue N, Yamamoto H, Kishida M, Inoue A, Asahara T, Yasui W, Kikuchi A: **Expression of Wnt-5a is correlated with aggressiveness of gastric cancer by stimulating cell migration and invasion**. *Cancer research* 2006, **66**(21):10439.
104. Zieker D, Königsrainer I, Traub F, Nieselt K, Knapp B, Schillinger C, Stirnkorb C, Fend F, Northoff H, Kupka S: **PGK1 a potential marker for peritoneal dissemination in gastric cancer**. *Cellular Physiology and Biochemistry* 2008, **21**(5-6):429-436.
105. Shichijo S, Azuma K, Komatsu N, Ito M, Maeda Y, Ishihara Y, Itoh K: **Two Proliferation-Related Proteins, TYMS and PGK1, Could Be New Cytotoxic T Lymphocyte-Directed Tumor-Associated Antigens of HLA-A2+ Colon Cancer**. *Clinical cancer research* 2004, **10**(17):5828-5836.

106. Kreunin P, Urquidi V, Lubman DM, Goodison S: **Identification of metastasis-associated proteins in a human tumor metastasis model using the mass-mapping technique**. *Proteomics* 2004, **4**(9):2754-2765.
107. Urquidi V, Sloan D, Kawai K, Agarwal D, Woodman AC, Tarin D, Goodison S: **Contrasting expression of thrombospondin-1 and osteopontin correlates with absence or presence of metastatic phenotype in an isogenic model of spontaneous human breast cancer metastasis**. *Clin Cancer Res* 2002, **8**(1):61-74.
108. Goodison S, Yuan J, Sloan D, Kim R, Li C, Popescu NC, Urquidi V: **The RhoGAP protein DLC-1 functions as a metastasis suppressor in breast cancer cells**. *Cancer Res* 2005, **65**(14):6042-6053.
109. Goodison S, Kawai K, Hihara J, Jiang P, Yang M, Urquidi V, Hoffman RM, Tarin D: **Prolonged dormancy and site-specific growth potential of cancer cells spontaneously disseminated from nonmetastatic breast tumors as revealed by labeling with green fluorescent protein**. *Clinical Cancer Research* 2003, **9**(10):3808-3814.
110. Leth-Larsen R, Lund R, Hansen HV, Laenkholm AV, Tarin D, Jensen ON, Ditzel HJ: **Metastasis-related Plasma Membrane Proteins of Human Breast Cancer Cells Identified by Comparative Quantitative Mass Spectrometry**. *Molecular & Cellular Proteomics* 2009, **8**(6):1436-1449.
111. Montel V, Huang TY, Mose E, Pestonjamasp K, Tarin D: **Expression profiling of primary tumors and matched lymphatic and lung metastases in a xenogeneic breast cancer model**. *American Journal of Pathology* 2005, **166**(5):1565-1579.
112. Krueger KE, Srivastava S: **Posttranslational protein modifications: current implications for cancer detection, prevention, and therapeutics**. *Mol Cell Proteomics* 2006, **5**(10):1799-1810.
113. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S *et al*: **Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents**. *Mol Cell Proteomics* 2004, **3**(12):1154-1169.
114. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Mol Cell Proteomics* 2002, **1**(5):376-386.
115. Mueller LN, Brusniak MY, Mani DR, Aebersold R: **An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data**. *J Proteome Res* 2008, **7**(1):51-61.
116. Rikova K, Guo A, Zeng Q, Possemato A, Yu J, Haack H, Nardone J, Lee K, Reeves C, Li Y *et al*: **Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer**. *Cell* 2007, **131**(6):1190-1203.
117. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M: **Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein**. *Mol Cell Proteomics* 2005, **4**(9):1265-1272.

118. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG: **Comparison of label-free methods for quantifying human proteins by shotgun proteomics**. *Mol Cell Proteomics* 2005, **4**(10):1487-1502.
119. Ulintz PJ, Bodenmiller B, Andrews PC, Aebersold R, Nesvizhskii AI: **Investigating MS2/MS3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence**. *Mol Cell Proteomics* 2008, **7**(1):71-87.
120. Yu LR, Zhu Z, Chan KC, Issaq HJ, Dimitrov DS, Veenstra TD: **Improved titanium dioxide enrichment of phosphopeptides from HeLa cells and high confident phosphopeptide identification by cross-validation of MS/MS and MS/MS/MS spectra**. *J Proteome Res* 2007, **6**(11):4150-4162.
121. Stulemeijer IJE, Joosten M, Jensen ON: **Quantitative Phosphoproteomics of Tomato Mounting a Hypersensitive Response Reveals a Swift Suppression of Photosynthetic Activity and a Differential Role for Hsp90 Isoforms**. *Journal of Proteome Research* 2009, **8**(3):1168-1182.
122. Jiang X, Han G, Feng S, Jiang X, Ye M, Yao X, Zou H: **Automatic validation of phosphopeptide identifications by the MS2/MS3 target-decoy search strategy**. *J Proteome Res* 2008, **7**(4):1640-1649.
123. Liu H, Sadygov RG, Yates JR, 3rd: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics**. *Anal Chem* 2004, **76**(14):4193-4201.
124. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP: **Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae**. *J Proteome Res* 2006, **5**(9):2339-2347.
125. Xie X, Li S, Liu S, Lu Y, Shen P, Ji J: **Proteomic analysis of mouse islets after multiple low-dose streptozotocin injection**. *Biochim Biophys Acta* 2008, **1784**(2):276-284.
126. Hunter T, Sefton BM: **Transforming gene product of Rous sarcoma virus phosphorylates tyrosine**. *Proc Natl Acad Sci U S A* 1980, **77**(3):1311-1315.
127. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M: **Global, in vivo, and site-specific phosphorylation dynamics in signaling networks**. *Cell* 2006, **127**(3):635-648.
128. Lu B, Ruse C, Xu T, Park SK, Yates J, 3rd: **Automatic validation of phosphopeptide identifications from tandem mass spectra**. *Anal Chem* 2007, **79**(4):1301-1310.
129. Chen Y, Olopade OI: **MYC in breast tumor progression**. *Expert Rev Anticancer Ther* 2008, **8**(10):1689-1698.
130. Calogero RA, Cordero F, Forni G, Cavallo F: **Inflammation and breast cancer. Inflammatory component of mammary carcinogenesis in ErbB2 transgenic mice**. *Breast Cancer Res* 2007, **9**(4):211.
131. Simeone AM, Tari AM: **How retinoids regulate breast cancer cell proliferation and apoptosis**. *Cell Mol Life Sci* 2004, **61**(12):1475-1484.
132. Haas M, Jost E: **Functional analysis of phosphorylation sites in human lamin A controlling lamin disassembly, nuclear transport and assembly**. *Eur J Cell Biol* 1993, **62**(2):237-247.

133. Cenni V, Sabatelli P, Mattioli E, Marmiroli S, Capanni C, Ognibene A, Squarzoni S, Maraldi NM, Bonne G, Columbaro M *et al*: **Lamin A N-terminal phosphorylation is associated with myoblast activation: impairment in Emery-Dreifuss muscular dystrophy**. *J Med Genet* 2005, **42**(3):214-220.
134. Cenni V, Bertacchini J, Beretti F, Lattanzi G, Bavelloni A, Riccio M, Ruzzene M, Marin O, Arrigoni G, Parnaik V *et al*: **Lamin A Ser404 is a nuclear target of Akt phosphorylation in C2C12 cells**. *J Proteome Res* 2008, **7**(11):4727-4735.
135. Bussolati G, Marchio C, Gaetano L, Lupo R, Sapino A: **Pleomorphism of the nuclear envelope in breast cancer: a new approach to an old problem**. *J Cell Mol Med* 2008, **12**(1):209-218.
136. Tourriere H, Gallouzi IE, Chebli K, Capony JP, Mouaikel J, van der Geer P, Tazi J: **RasGAP-associated endoribonuclease G3Bp: selective RNA degradation and phosphorylation-dependent localization**. *Mol Cell Biol* 2001, **21**(22):7747-7760.
137. Planas-Silva MD, Bruggeman RD, Grenko RT, Smith JS: **Overexpression of c-Myc and Bcl-2 during progression and distant metastasis of hormone-treated breast cancer**. *Exp Mol Pathol* 2007, **82**(1):85-90.
138. Barnes CJ, Li F, Mandal M, Yang Z, Sahin AA, Kumar R: **Heregulin induces expression, ATPase activity, and nuclear localization of G3BP, a Ras signaling component, in human breast tumors**. *Cancer Res* 2002, **62**(5):1251-1255.
139. Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL, Washburn MP: **Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors**. *Methods* 2006, **40**(4):303-311.
140. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J: **Molecular cell biology**. *New York* 1995.
141. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast**. *Molecular and cellular biology* 1999, **19**(3):1720.
142. Cravatt BF, Simon GM, Yates III JR: **The biological impact of mass-spectrometry-based proteomics**. *Nature* 2007, **450**(7172):991-1000.