5-2018

Proteomics Strategies to Develop Proteins of Post-translational Modifications in Plasma-Derived Extracellular Vesicles as Disease Markers

I-Hsuan (Blair) Chen
Purdue University

PROTEOMICS STRATEGIES TO DEVELOP PROTEINS OF POST-TRANSLATIONAL MODIFICATIONS IN PLASMA-DERIVED EXTRACELLULAR VESICLES AS DISEASE MARKERS

by

I-Hsuan (Blair) Chen

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biochemistry

West Lafayette, Indiana

May 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. W Andy Tao, Chair
Department of Biochemistry
Dr. Ourania Andrisani
Department of Basic Medical Sciences
Dr. Humaira Gowher
Department of Biochemistry
Dr. Jer-Yen Yang
Department of Basic Medical Sciences

Approved by:
Dr. Jason Cannon
Head of the Graduate Program

To my father, who taught me to be strong; my mother, who taught me to be independent; my sister, who helped me on taking care of my family.

With all my love

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. W. Andy Tao, for his guidance and support in letting me attempt the new project that I am passionate about. I have learned so much during my time in his lab, not just science, but also collaborating with other scientist and presenting my work. I also need to acknowledge my thesis committees, Dr. Humaira Gowher, Dr. Ourania Andrasani and Dr. Jer-Yen Yang for the suggestions and support of my projects. Of course I need to thank my collaborators for teaching me a lot of new things, Dr. Liqin Zhu for the HCC mice project, and IU biobank providing me so many samples so I can proceed my projects.

I also need to thank the wonderful lab mates I ‘ve ever had over four years. I appreciate the mentorship of Dr. Anton Iliuk; the advice and friendship of Chuan-Chih, Li Pan and Justine Arrington; the support and friendship of Hillary Andaluz, Peipei Zhu and Sebastian Paze. It’s all you guys made my time here enjoyable and have so many good memories.

In the end I must thank my loving family, my Parents Ming-Yang Chen, Wen-Lin Chen and my sister Yi-Chun Chen, my grandmother Hung-Ming Chang, who inspired me for pursuing the PhD, Grandma, I have kept my promise to you! Thank you all for encouraging me and being my biggest fan all the time. Thank you as well my new and old friends, who supported me; and my roommate Amy Hung, who was on the same journey, always being a good listener. Last but not the least, thank you to my best partner Yu-Chen Chang, for your understanding, helping me with my writing, and always being supportive.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAA	2-chloroacetamide
EVs	Extracellular Vesicles
IMAC	Immobilized metal ion affinity chromatography
LC-MS	Liquid chromatography-mass spectrometry
MVs	Microvesicles
PolyMAC	Polymer-based metal-ion affinity capture
PTMs	Post translational modifications
TCEP	Tris(2-carboxyethyl)phosphine hydrochloride

ABSTRACT

Author: Chen, I-Hsuan. Ph.D.
Institution: Purdue University
Degree Received: May 2018
Title: Proteomics Strategies to Develop Proteins of Post-Translational Modifications in Plasma-Derived Extracellular Vesicles as Disease Markers
Major Professor: Weiguo Andy Tao

Blood tests, which are the most wide spread diagnosis procedure in clinical analysis, apply blood biomarkers to categorize patients and support treatment decisions. However, existing biomarkers often lack specificity and are far from comprehensive. Mass spectrometry-based proteomics allow users to characterize plasma protein in great depth and has become a powerful tool in the biomarker discovery area. However, because of the extremely high dynamic range of plasma, being able identify thousands of plasma proteins using methods such as Liquid chromatography-tandem mass spectrometry (LC-MS/MS) remains a challenge. Furthermore, recent discoveries of extracellular vesicles (EVs) have proven that EVs have a high possibility for becoming the source for biomarker discovery and disease diagnosis. In addition to the protein in EVs, post-translation modification proteins (PTMs proteins) are also interesting targets because the PTMs proteins are involved with many cancer-related signaling transductions. This dissertation proposes proteomics strategies of using PTMs proteins in plasma-derived extracellular vesicles as breast cancer markers. Initially, Chapter One highlights the potential of using phosphoproteins in extracellular vesicles as markers for breast cancer. Chapter Two delves into the development of a pipeline proteomics strategy that utilizes glycoproteins in EVs as breast cancer markers. Finally, Chapter Three explores the details of different subtypes, which presents the possibility of leveraging three PTMs including phosphorylation, acetylation and glycosylation to distinguish three major breast cancer subtypes.

CHAPTER 1. PHOSPHOPROTEINS IN EXTRACELLULAR VESICLES AS CANDIDATE MARKERS FOR BREAST CANCER

1.1. Summary

The state of protein phosphorylation can be a key determinant of cellular physiology such as early stage cancer, but the development of phosphoproteins in biofluids for disease diagnosis remains elusive. Here we demonstrate a strategy to isolate and identify phosphoproteins in extracellular vesicles (EVs) from human plasma as potential markers to differentiate disease from healthy states. We identified close to 10,000 unique phosphopeptides in EVs isolated from small volumes of plasma samples. Using label-free quantitative phosphoproteomics, we identified 144 phosphoproteins in plasma EVs that are significantly higher in patients diagnosed with breast cancer compared with healthy controls. Several biomarkers were validated in individual patients using paralleled reaction monitoring for targeted quantitation. This study demonstrates that the development of phosphoproteins in plasma EV as disease biomarkers is highly feasible and may transform cancer screening and monitoring.

1.2. Introduction

Early diagnosis and monitoring of diseases such as cancers through blood tests has been a decades long aim of medical diagnostics. Since protein phosphorylation is one of the most important and widespread molecular regulatory mechanisms that controls almost all aspects of cellular functions(1, 2), the status of phosphorylation events conceivably provides clues regarding disease status (3). However, few phosphoproteins have been developed as disease markers. Assays of phosphoproteins from tissues face tremendous challenges due to the invasive nature of tissue biopsy and the highly dynamic nature of protein phosphorylation during the typically long and complex procedure of tissue biopsy. Furthermore, biopsy tissue from tumors is not available for monitoring patient response over the course of treatment. Development of phosphoproteins as disease biomarkers from biofluids is even more challenging due to the presence of active phosphatases in high concentration in blood. With a few high abundant proteins representing over

95% of the mass in blood, few phosphorylated proteins in plasma/serum can be identified with stable and detectable concentration.

The recent discovery of extracellular vesicles (EVs), including microvesicles and exosomes, and their potentially important cellular functions in tumor biology and metastasis has presented them as intriguing sources for biomarker discovery and disease diagnosis (4-6). Critical for immune regulation and intercellular communication, EVs have many differentiating characteristics of cancer cell-derived cargo, including mutations, active miRNAs, and signaling molecules with metastatic features (7, 8). The growing body of functional studies have provided strong evidence that these EV-based disease markers can be identified well before the onset of symptoms or physiological detection of a tumor, making them a promising candidate for early-stage cancer and other diseases (6, 9). Interestingly, EVs are membrane-encapsulated nano- or microparticles, which protects their inside contents from external proteases and other enzymes (10-12). These features make them highly stable in a biofluid for extended periods of time, and also allow us to potentially develop phosphoproteins in EVs for medical diagnoses. The ability to detect the genome output – active proteins, in particular phosphoproteins – can provide more direct real time information about the organism's physiological functions and disease progression, particularly in cancers.

We aimed to develop EV phosphoproteins as potential disease markers by focusing on breast cancer in this study. To this end, we isolated and identified the largest group of EV phosphoproteins to date from both microvesicles and exosomes, and measured phosphorylation changes across breast cancer patients and healthy individuals. We subsequently identified multiple potential candidates and verified several among patients and healthy controls. The EV phosphoproteomics approach demonstrated here can be applied to other systems and thus establish a new strategy for biomarker discovery.

1.3. Experimental procedure

1.3.1. Plasma Samples

The Indiana University Institutional Review Board approved the use of human plasma samples. Blood samples were collected from 6 healthy females and from 18 breast cancer patients that

obtained through the IU Simon Cancer Center. Plasma samples were collected by standard protocol, in brief, plasma sample processing was initiated within 30 min of blood draw to an ethylenediaminetetraacetic acid (EDTA) containing tube. Samples were spun for 30 min at 3500 rpm to remove all cell debris and platelet.

1.3.1. Extracellular vesicles isolation

A total 5.5ml pool plasma samples were collected from both healthy control and breast cancer patient group for technical replicates phosphoproteomics. Plasma samples were centrifuged at 20,000 xg at 4 ^{o}C for 1hr. Pellets were washed with cold PBS and centrifuged again at 20,000 xg at 4 ^{o}C for 1 hr, the pellets were microvesicles. Supernatant of first centrifugation were further centrifuged at ultra-high speed 100,000 xg at 4 ^{o}C for 1hr. Pellets were wash with cold PBS and centrifuged at 100,000 xg for 1hr again. The pellets from ultra-high speed centrifugation were exosome.

.

1.3.2. Protein digestion

The digestion was performed with phase transfer surfactant aids (PTS) digestion(13). Extracellular vesicles were solubilized in lysis buffer containing 12mM sodium deoxycholate (SDC), 12mM sodium lauroyl sarcosinate (SLS) and phosphatase inhibitor cocktail (Sigma-Aldrich, St. Louis, MO) in 100mM Tris-HCl, pH8.5. Proteins were reduced and alkylated with 10 mM tris-(2-carboxyethyl)phosphine (TECP) and 40 mM chloroacetamide (CAA) at 95 °C for 5 min. Alkylated proteins were diluted to 5 fold by 50mM triethylammonium bicarbonate (TEAB) and digested with Lys-C (Wako, Japan) in a 1:100 (w/w) enzyme to protein ratio for 3 hr at 37 °C. Trypsin was added to a final 1:50 (w/w) enzyme-to-protein ratio for overnight digestion. The digested peptides were acidified with trifluoroacetic acid (TFA) to final concentration of 0.5% TFA, and 250ul of Ethyl acetate was added to 250ul digested solution. The mixture was shaken for 2 min, then centrifuged at 13,200 rpm for 2 min to obtain aqueous and organic phases. The aqueous phase was collected and desalted using a 100 mg of Seppak C18 column (Waters, Milford, MA).

1.3.3. Phosphoproteomics enrichment

The phosphopeptide enrichment was performed according to the reported protocol with some modifications(14). The in-house constructed IMAC tip was made by capping the end with a 20µm polypropylene frits disk (Agilent, Wilmington, DE, USA). The tip was packed with 5 mg of Ni-NTA silica resin by centrifugation. Prior to sample loading, Ni^{2+} ions were removed by 100 mM EDTA solution. Furthermore, the beads were chelating with Fe^{3+} and equilibrated with loading buffer (6% (v/v) acetic acid (AA) at pH 2.7). Tryptic peptides were reconstituted in loading buffer and loaded onto the IMAC tip. After successive washes with 4% (v/v) AA, 25% ACN, and 6% (v/v) AA, the bound phosphopeptides were eluted with 200 mM $NH_4H_2PO_4$. The eluted phosphopeptides were desalted using C-18 StageTips (15).

1.3.4. LC-MS/MS analysis

The phosphopeptides were dissolved in 4 µL of 0.3% formic acid (FA) with 3% ACN and injected into an Easy-nLC 1000 (Thermo Fisher Scientific). Peptides were separated on a 45 cm in-house packed column (360 µm OD × 75 µm ID) containing C18 resin (2.2 µm, 100Å, Michrom Bioresources) with a 30 cm column heater (Analytical Sales and Services) and the temperature was set at 50 °C. The mobile phase buffer consisted of 0.1% FA in ultra pure water (buffer A) with an eluting buffer of 0.1% FA in 80% ACN (buffer B) run over either with a linear 45 min or 60 min gradient of 6%-30% buffer B at flow rate of 250 nL/min. The Easy-nLC 1000 was coupled online with a Velos Pro LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific). The mass spectrometer was operated in the data dependent mode in which a full scan MS (from m/z 350-1500 with the resolution of 30,000 at m/z 400). The 10 most intense ions were subjected to collision induced dissociation (CID) fragmentation (normalized collision energy (NCE) 30%, AGC 3e4, max injection time 100 ms).

1.3.5. PRM analysis

Peptide samples were dissolved in 8 µl of 0.1% formic acid and injected 6ul into easy nLC 1200 (Thermo) HPLC system. Eluent was introduced into the mass spectrometer using 10cm

PicoChip® columns filled with 3uM ReprosilPUR C18 (New Objective, Woburn, MA) operated at 2.6 kV. The mobile phase buffer consists of 0.1% formic acid in water with an eluting buffer of 0.1% formic acid (Buffer A) in 90% CH3CN (Buffer B). The LC flow rate was 300nl/min. The gradient was set as 0–30% Buffer B for 30 mins and 30-80% for 10mins. The sample was acquired on Q Exactive HF (Thermo, Germany). Each sample was analyzed under parallel reaction monitoring (PRM) with an isolation width of ±0.7 Th. In all experiments, a full mass spectrum at 60,000 resolution relative to m/z 200 (AGC target 3 × 106, 100 ms maximum injection time, m/z 400–1600) was followed by up to 20 PRM scans at 15000 resolution (AGC target 1e5, 50 ms maximum injection time) as triggered by a unscheduled inclusion list. Higher energy collisional dissociation (HCD) was used with 30eV normalized collision energy.

1.3.6. Data processing

The raw files were searched directly UniprotKB database version Jan2015 with no redundant entries using MaxQuant software (version 1.5.4.1)(16) with Andromeda search engine. Initial precursor mass tolerance was set at 20 p.p.m. and the final tolerance was set at 6 p.p.m., and ITMS MS/MS tolerance was set at 0.6 Da. Search criteria included a static carbamidomethylation of cysteines (+57.0214 Da) and variable modifications of (1) oxidation (+15.9949 Da) on methionine residues, (2) acetylation (+42.011 Da) at N-terminus of protein, and (3) phosphorylation (+79.996 Da) on serine, threonine or tyrosine residues were searched. Search was performed with Trypsin/P digestion and allowed a maximum of two missed cleavages on the peptides analyzed from the sequence database. The false discovery rates of proteins, peptides and phosphosites were set at 0.01. The minimum peptide length was six amino acids, and a minimum Andromeda score was set at 40 for modified peptides. A site localization probability of 0.75 was used as the cut off for localization of phosphorylation sites. All the peptide spectral matches and MS/MS spectra can be viewed through MaxQuant viewer. All the localized phosphorylation sites and corresponding phosphoproteins were submitted to pLogo software (17) and Panther (18) to determine the phosphorylation motifs and gene ontology, respectively. PRM data were manually curated within Skyline (version 3.5.0.9319)(19)

1.3.7. Quantitation and Statistical Rationale

All data were analyzed by using the Perseus software (version 1.5.4.1) (20). For quantification of both proteomic and phosphoproteomic, the intensities of peptides and phosphopeptides were extracted through MaxQuant, and the missing values of intensities were replaced by normal distribution with a downshift of 1.8 standard deviations and a width of 0.3 standard deviations. The significantly increased phosphosites or proteins in patient samples were identified by the p-value is significant from a two sample t-test with a permutation-based FDR cut off 0.05 with S0 set on 0.2 for all of data sets. The up-regulated candidate networks were predicted in STRING version 10.0(21) with the interaction score ≥ 0.4, and the signal networks were visualized using Cytoscape version 3.4.0(22) with MCODE plugin version 1.4.2(23)

1.4. Result

1.4.1. Identification of 9,643 Unique Phosphopeptides from Plasma Microvesicles and Exosomes.

The workflow for the isolation of EVs, enrichment of phosphopeptides, and EV phosphoproteome analyses is illustrated in Fig. 1.1. Microvesicles and exosomes were isolated from human plasma samples through high-speed and ultra-highspeed centrifugations, respectively, an approach that has been used in previous studies (13–15). For the initial screening, the plasma samples were collected and pooled from healthy individuals ($n = 6$) and from patients diagnosed with breast cancer ($n = 18$). After lysis of EVs, proteins were extracted and peptides generated using trypsin with the aid of phase transfer surfactants for better digestion efficiency and fewer missed tryptic sites (24). Phosphopeptides were enriched and analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) on a high-speed, high-resolution mass spectrometer. For each phosphopeptide sample, three technical replicates were performed. Label-free quantification was performed to determine differential phosphorylation of EV proteins in the plasma of control and breast cancer patient samples.

The strategy allowed us to identify 9,643 unique phosphopeptides, including 9,225 from microvesicles and 1,014 from exosomes, representing 1,934 and 479 phosphoproteins in

microvesicles and exosomes, respectively. On average, close to 7,000 unique EV phosphopeptides were identified from 1 mL human plasma. As shown in Fig. 1. 2A and Fig. 1.3 A, more than 50% of exosome phosphopeptides were also identified in microvesicles. Gene ontology analysis of the phosphoproteins indicated overall similar cellular components and biological functions between microvesicles and exosomes (Fig. 1.2B and Fig. 1.3B). Although previous large-scale phosphoproteomics studies revealed that phosphorylation preferentially targets nuclear proteins (25, 26), a significant portion of the EV phosphoproteomes are distinctively from membranes and organelles. As expected, proteins annotated as extracellular were significantly overrepresented in the EV phosphoproteomes. We also found that many EV phosphoproteins are involved in cell–cell communication, stimulus response, and biogenesis.

The EV phosphoproteome analyses revealed that the distribution of tyrosine, threonine, and serine phosphorylation (pY, pT and pS) sites is 2.0%, 14.1%, and 83.9%, respectively, for microvesicle phosphoproteins, which is similar to previously reported site distribution in in vivo human phosphoproteomes(27) . Interestingly, the distribution of pY in exosomes is an order of magnitude higher, at 13.7%, which is quite close to the distribution of pT, at 16.1% (Fig. 1.2C). This apparent discrepancy may reflect the different origins of microvesicles and exosomes. Microvesicles bud directly from the plasma membrane, whereas exosomes are represented by endosome-associated proteins, in which proteins such as integrins, hormone receptors, growth factor receptors, receptor tyrosine kinases, and nonreceptor tyrosine kinases such as Src kinases are involved. A further motif analysis of pS/T phosphorylation sites revealed overall similar distribution of general motif to cellular phosphoproteome; for example, the most abundant class of sites is acidophilic, followed by proline-directed and basophilic (Fig. 1.4 A). However, in the exosome phosphoproteome, proline-directed phosphorylation constitutes only half of that in microvesicles, and therefore the motif assay does not show dominant –SP- motif in the exosome phosphoproteome (Fig. 1.4B).

1.4.2. Cancer-Specific Phosphoproteins in EV.

Label-free quantitation of phosphopeptides with the probability score of phosphorylation site location over 0.75 was used to identify differential phosphorylation events in patients with breast cancer from those in healthy individuals. We quantified 3,607 and 461 unique phosphosites and

identified 156 and 271 phosphosites with significant changes [false discovery rate (FDR) < 0.05 and S0 = 0.2] in microvesicles and exosomes, respectively (Fig. 1.5 A and B). Differential phosphorylation may be a result of changes in protein expression or changes of a particular site's phosphorylation. To distinguish these factors, we also performed label-free quantitation of total proteomes for both microvesicles and exosomes. We identified 1,996 proteins, 34.4% of which were also identified with phosphopeptide enrichment. In comparison, 862 proteins were detected in the phosphorylation data alone, indicating that phosphoproteins are typically of low abundance, escaping detection via the shotgun proteomics approach. Quantitative analyses of EV proteomes revealed strikingly similar expression of most proteins in healthy individuals and patients with cancer (Fig. 1.5A). In comparison, there are a larger number of phosphorylation sites with significant changes in patient samples, indicating that these phosphorylation differences between patients with cancer and healthy individuals are not a result of changes in protein expression, and thus reflect phosphorylation truly specific to patients with cancer. The result also justifies our approach to developing protein phosphorylation changes, instead of protein expression changes, as the measurement of disease progression. EV proteomic analyses also revealed that several protein markers were only identified in microvesicles or exosomes specifically, but at the same time, there are some protein markers identified in both particles (Fig. 1.6). Western blotting was carried out with the antibody against CD 31, which is considered an endothelial-derived microvesicles marker. Although CD 31 was mainly identified in microvesicles, the Western blotting (WB) experiment and MS data indicated that the current isolation method based on ultracentrifugation is not entirely specific.

We compared these phosphosites representing 197 unique phosphopeptides that showed significant increase in patients with breast cancer with all identified unique phosphopeptides in EV phosphoproteomes (Figs. 1.7 and 1.8). Again, the disparity of relative abundance of pY/pT/pS and sequence motif in microvesicle and exosomes may be a result of their different origins. Although phosphopeptides that showed a significant decrease in patients with breast cancer might be interesting, it is conceivable that these phosphopeptides were not necessarily down-regulated in EV pools, as EVs from other cell sources could compensate them. Therefore, we focused our attention on these 197 unique phosphopeptides. Motif analyses of the corresponding phosphosites found that proline-directed motif (s/tP) decreased significantly, whereas the AB motif increased. In terms of cellular components, the up-regulated phosphoproteins showed a slightly increased

share of membrane proteins in MV, whereas there is increase in extracellular proteins in exosome. We further compared the 197 unique phosphopeptides with a recent comprehensive proteogenomic study in which breast phosphoproteomics studies were carried out in tissues from 105 patients with breast cancer (28). We found that a significant portion of these 197 phosphopeptides (>60%) were also identified by the proteogenomic study (Fig. 1.9A), indicating that EV phosphoproteome is sensitive and that quantitative analyses of EV phosphoproteomics can identify phosphorylation events that are disease specific. However, because EVs can be released from diverse types of cells, the difference could be the result of distinctive immune response or other factors in healthy individuals and patients with cancer. Nevertheless, the results highlight the advantage of analyzing EV phosphoproteome through liquid biopsy over tissue biopsy, which is invasive and subject to variation because of the long procedure.

To better understand the biological roles of differential phosphorylation events, we examined phosphoproteins specific to patients with cancer, using STRING to identify enriched gene ontology categories and signaling networks (21). We found that several crucial functions related to cancer metastasis, membrane reorganization, and intercellular communication were enriched in cancer-specific EV phosphoproteins (Fig. 1.9B). It is interesting to reveal the central role of SRC tyrosine kinase with multiple phosphoproteins identified in the study, which is consistent with previous studies linking an elevated level of activity of SRC to cancer progression by promoting other signals. Please note that although 16% of phosphoproteins that were up-regulated in patients with cancer are membrane proteins, and because of relative lack of protein–protein interaction data with membrane proteins, these membrane proteins were not implicated in the STRING analysis.

1.4.3. Verification of Phosphorylation Specific to Patients with Cancer, Using Parallel Reaction Monitoring

Because breast cancer is extremely heterogeneous, the chance to identify a single diagnostic biomarker is likely rare. Instead, the identification of a panel of candidate markers that reflect the onset and progression of key disease-related signaling events would be feasible to offer better prognostic value. In an effort to validate the differential phosphorylation of potential markers in patients with cancer, we applied parallel reaction monitoring (PRM) (29) to quantify individual EV phosphopeptides in plasma from patients with breast cancer and healthy individuals. Because

phosphospecific antibodies suitable for construction of ELISA are rarely available, targeted, quantitative MS approaches such as PRM and MRM (multireaction monitoring) are essential for initial validation. As a demonstration that PRM can be used to initially verify candidate phosphoproteins, we selected four phosphoproteins: Ral GTPase activating protein subunit alpha-2 (RALGAPA2), cGMP dependent protein kinase1 (PKG1), tight junction protein 2 (TJP2), and nuclear transcription factor, X box binding protein 1 (NFX1). These four proteins showed significant phosphorylation up-regulation in patients with cancer, were previously reported as phosphoproteins, and have been implicated in multiple breast cancer studies (30-33)

Quantitative assays based on PRM were performed with plasma EV samples from 13 patients with cancer (eight additional patient samples) and seven healthy controls (one additional control). The relative abundance data of phosphopeptides from four individual proteins are presented as a linear box-and-whiskers plot (Fig. 1.10). With reference from the figure, RALGAPA2, PKG1, and TJP2 were observed to be significantly elevated in patients with breast cancer compared with in control patients. However, the fold difference is noticeably smaller in PRM than label-free quantification. In particular, NFX1 phosphorylation was only identified in breast cancer samples, and not in healthy controls, but because of large variation among individual samples, the difference of NFX1 phosphorylation on the specific site is statistically inconclusive. The data may be the reflection of dynamic suppression of targeted proteomics such as MRM and PRM. Nevertheless, large variation among clinical samples underscores current challenges facing biomarker validation.

1.5. Discussion

MS-based proteomic profiling and quantitation holds enormous promise for uncovering biomarkers. However, successful applications to human diseases remain limited. This is, in large part, a result of the complexity of biofluids that have an extremely wide dynamic range and are typically dominated by a few highly abundant proteins. This prevents the development of a coherent, practical pipeline for systemic screening and validation. Here, we reported in-depth analyses of phosphoproteomes in plasma EVs and demonstrated the feasibility of developing phosphoproteins as potential disease biomarkers. Previous studies typically could only identify a small number of phosphoproteins in plasma, likely as a result of the presence of phosphatases in the bloodstream, and the level of phosphorylation does not have any clear meaningful connection

to biological status (34, 35). We presented an MS-based strategy that includes the isolation of EV particles from human blood, enrichment of EV phosphopeptides, LC-MS/MS analyses, and PRM quantification for biomarker discovery and quantitative verification. We analyzed samples from patients with breast cancer, in comparison with healthy controls, to identify candidate breast cancer biomarkers. These candidates will need to be further evaluated in larger, heterogeneous patient cohorts of defined breast cancer subtypes in the future. The study highlights our ability to isolate and identify thousands of phosphopeptides from limited volumes of biobanked human plasma samples. These findings provide a proof of principle for this strategy to be used to explore existing resources for a wide range of diseases.

Recently, liquid biopsies (analysis of biofluids such as plasma and urine) have gained much attention for cancer research and clinical care, as they offer multiple advantages in clinical settings, including their noninvasive nature, a suitable sample source for longitudinal disease monitoring, better screenshot of tumor heterogeneity, and so on. Current liquid biopsies primarily focus on the detection and downstream analysis of circulating tumor cells and circulating tumor DNA. A major obstacle with the cur-ent methods is the heterogeneity and extreme rarity of the circulating tumor cells and circulating DNA. EVs offer all the same attractive advantages of a liquid biopsy, but without the sampling limitation of circulating tumor cells and circulating tumor DNA. At present, most of the studies on EVs focus on microRNAs and a small portion on EV proteins. The ability to detect the genome output, and in particular functional proteins such as phosphoproteins, can arguably provide more useful real-time information about the organism's physiological functions and disease progression, such as in the early detection and monitoring of cancers.

Our study clearly indicates that EV phosphoproteomes can be readily captured and analyzed. It is interesting to know that EV phosphoproteins are stable over a long period of storage time (the plasma samples from Indiana Biobank were collected more than 5 y ago), which is critical for applications in clinical tests. However, a thorough investigation on EV phophoproteome stability might be necessary, as cellular phosphorylation events are extremely dynamic and EVs are circulating in the blood for long periods of time. EV phosphoproteomes may mainly represent phosphorylation events that are constitutively active, and therefore insensitive to capturing acute events. All these questions can be addressed with further studies on well-defined EV samples, possibly using animal models.

Last, although we present here a feasible strategy to develop phosphoproteins as potential

disease markers, it relies on the isolation of a good quantity of EVs with high reproducibility. At this stage, the isolation of microvesicles and exosomes is primarily based on differential high-speed centrifugation, which is not highly specific and is unlikely suitable for clinical settings. Immunoprecipitation of microvesicles and exosomes may introduce bias and contaminations from plasma proteins. The development of phosphoproteins as biomarkers is also severely limited by the availability of phosphospecific antibodies. The inability to develop ELISA or similar immunobased assays will inevitably depend on alternative validation methods such as MS-based targeted quantitation and nonantibody-based methods (36, 37). The complexity of biofluids and the necessity of including EV isolation and phosphopeptide isolation in a sample preparation will no doubt add extra challenges to the accuracy of MS-based targeted quantitation of heterogeneous clinical samples.

1.6. Data Access

The raw data, MaxQuant output text files, and supplementary MS quantitation tables for all proteomic analyses have been deposited to the ProteomeXchange Consortium(38) (http://proteomecentral.proteomexchange.org) with the data set identifier PXD005214.

1.7. Reference

1. Hunter, T. (2000) Signaling--2000 and beyond. *Cell* 100, 113-127
2. Kabuyama, Y., Resing, K. A., and Ahn, N. G. (2004) Applying proteomics to signaling networks. *Curr Opin Genet Dev* 14, 492-498
3. Iliuk, A. B., Arrington, J. V., and Tao, W. A. (2014) Analytical challenges translating mass spectrometry-based phosphoproteomics from discovery to clinical applications. *Electrophoresis*
4. Melo, S. A., Luecke, L. B., Kahlert, C., Fernandez, A. F., Gammon, S. T., Kaye, J., LeBleu, V. S., Mittendorf, E. A., Weitz, J., Rahbari, N., Reissfelder, C., Pilarsky, C., Fraga, M. F., Piwnica-Worms, D., and Kalluri, R. (2015) Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature* 523, 177-182

5. Gonzales, P. A., Pisitkun, T., Hoffert, J. D., Tchapyjnikov, D., Star, R. A., Kleta, R., Wang, N. S., and Knepper, M. A. (2009) Large-scale proteomics and phosphoproteomics of urinary exosomes. *Journal of the American Society of Nephrology : JASN* 20, 363-379
6. Boukouris, S., and Mathivanan, S. (2015) Exosomes in bodily fluids are a highly stable resource of disease biomarkers. *Proteomics Clin Appl* 9, 358-367
7. Zhang, Y., and Wang, X. F. (2015) A niche role for cancer exosomes in metastasis. *Nature cell biology* 17, 709-711
8. Costa-Silva, B., Aiello, N. M., Ocean, A. J., Singh, S., Zhang, H., Thakur, B. K., Becker, A., Hoshino, A., Mark, M. T., Molina, H., Xiang, J., Zhang, T., Theilen, T. M., Garcia-Santos, G., Williams, C., Ararso, Y., Huang, Y., Rodrigues, G., Shen, T. L., Labori, K. J., Lothe, I. M., Kure, E. H., Hernandez, J., Doussot, A., Ebbesen, S. H., Grandgenett, P. M., Hollingsworth, M. A., Jain, M., Mallya, K., Batra, S. K., Jarnagin, W. R., Schwartz, R. E., Matei, I., Peinado, H., Stanger, B. Z., Bromberg, J., and Lyden, D. (2015) Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nature cell biology* 17, 816-826
9. Saraswat, M., Joenvaara, S., Musante, L., Peltoniemi, H., Holthofer, H., and Renkonen, R. (2015) N-linked (N-) glycoproteomics of urinary exosomes. [Corrected]. *Mol Cell Proteomics* 14, 263-276
10. Sokolova, V., Ludwig, A. K., Hornung, S., Rotan, O., Horn, P. A., Epple, M., and Giebel, B. (2011) Characterisation of exosomes derived from human cells by nanoparticle tracking analysis and scanning electron microscopy. *Colloids and surfaces. B, Biointerfaces* 87, 146-150
11. Palmisano, G., Jensen, S. S., Le Bihan, M. C., Laine, J., McGuire, J. N., Pociot, F., and Larsen, M. R. (2012) Characterization of membrane-shed microvesicles from cytokine-stimulated beta-cells using proteomics strategies. *Mol Cell Proteomics* 11, 230-243
12. Cocucci, E., and Meldolesi, J. (2015) Ectosomes and exosomes: shedding the confusion between extracellular vesicles. *Trends in cell biology* 25, 364-372
13. Masuda, T., Sugiyama, N., Tomita, M., and Ishihama, Y. (2011) Microscale phosphoproteome analysis of 10,000 cells from human cancer cell lines. *Anal Chem* 83, 7698-7703

14. Tsai, C. F., Hsu, C. C., Hung, J. N., Wang, Y. T., Choong, W. K., Zeng, M. Y., Lin, P. Y., Hong, R. W., Sung, T. Y., and Chen, Y. J. (2014) Sequential phosphoproteomic enrichment through complementary metal-directed immobilized metal ion affinity chromatography. *Anal Chem* 86, 685-693
15. Rappsilber, J., Mann, M., and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* 2, 1896-1906
16. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367-1372
17. O'Shea, J. P., Chou, M. F., Quader, S. A., Ryan, J. K., Church, G. M., and Schwartz, D. (2013) pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 10, 1211-1212
18. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 44, D336-342
19. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26, 966-968
20. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13, 731-740
21. Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 28, 3442-3444
22. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498-2504
23. Bader, G. D., and Hogue, C. W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4, 2

24. Masuda, T., Saito, N., Tomita, M., and Ishihama, Y. (2009) Unbiased quantitation of Escherichia coli membrane proteome using phase transfer surfactants. *Mol Cell Proteomics* 8, 2770-2777
25. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127, 635-648
26. Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D., Lam, H., Schmidt, A., Rinner, O., Mueller, L. N., Shannon, P. T., Pedrioli, P. G., Panse, C., Lee, H. K., Schlapbach, R., and Aebersold, R. (2007) PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol Syst Biol* 3, 139
27. Sharma, K., D'Souza, R. C., Tyanova, S., Schaab, C., Wisniewski, J. R., Cox, J., and Mann, M. (2014) Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep* 8, 1583-1594
28. Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., Huang, K. L., Lin, C., McLellan, M. D., Yan, P., Davies, S. R., Townsend, R. R., Skates, S. J., Wang, J., Zhang, B., Kinsinger, C. R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A. G., Fenyo, D., Ellis, M. J., Carr, S. A., and Nci, C. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 55-62
29. Bourmaud, A., Gallien, S., and Domon, B. (2016) Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications. *Proteomics* 16, 2146-2159
30. Peri, S., de Cicco, R. L., Santucci-Pereira, J., Slifker, M., Ross, E. A., Russo, I. H., Russo, P. A., Arslan, A. A., Belitskaya-Levy, I., Zeleniuch-Jacquotte, A., Bordas, P., Lenner, P., Ahman, J., Afanasyeva, Y., Johansson, R., Sheriff, F., Hallmans, G., Toniolo, P., and Russo, J. (2012) Defining the genomic signature of the parous breast. *BMC Med Genomics* 5, 46
31. Gong, Y., Xu, C. Y., Wang, J. R., Hu, X. H., Hong, D., Ji, X., Shi, W., Chen, H. X., Wang, H. B., and Wu, X. M. (2014) Inhibition of phosphodiesterase 5 reduces bone mass by suppression of canonical Wnt signaling. *Cell Death Dis* 5, e1544

32. Nam, S., Chang, H. R., Jung, H. R., Gim, Y., Kim, N. Y., Grailhe, R., Seo, H. R., Park, H. S., Balch, C., Lee, J., Park, I., Jung, S. Y., Jeong, K. C., Powis, G., Liang, H., Lee, E. S., Ro, J., and Kim, Y. H. (2015) A pathway-based approach for identifying biomarkers of tumor progression to trastuzumab-resistant breast cancer. *Cancer Lett* 356, 880-890
33. Yi, T., Zhai, B., Yu, Y., Kiyotsugu, Y., Raschle, T., Etzkorn, M., Seo, H. C., Nagiec, M., Luna, R. E., Reinherz, E. L., Blenis, J., Gygi, S. P., and Wagner, G. (2014) Quantitative phosphoproteomic analysis reveals system-wide signaling pathways downstream of SDF-1/CXCR4 in breast cancer stem cells. *Proc Natl Acad Sci U S A* 111, E2182-2190
34. Jaros, J. A., Guest, P. C., Ramoune, H., Rothermundt, M., Leweke, F. M., Martins-de-Souza, D., and Bahn, S. (2012) Clinical use of phosphorylated proteins in blood serum analysed by immobilised metal ion affinity chromatography and mass spectrometry. *J Proteomics* 76 Spec No., 36-42
35. Hu, L., Zhou, H., Li, Y., Sun, S., Guo, L., Ye, M., Tian, X., Gu, J., Yang, S., and Zou, H. (2009) Profiling of endogenous serum phosphorylated peptides by titanium (IV) immobilized mesoporous silica particles enrichment and MALDI-TOFMS detection. *Anal Chem* 81, 94-104
36. Iliuk, A., Liu, X. S., Xue, L., Liu, X., and Tao, W. A. (2012) Chemical visualization of phosphoproteomes on membrane. *Mol Cell Proteomics* 11, 629-639
37. Pan, L., Iliuk, A., Yu, S., Geahlen, R. L., and Tao, W. A. (2012) Multiplexed quantitation of protein expression and phosphorylation based on functionalized soluble nanopolymers. *J Am Chem Soc* 134, 18201-18204
38. Hermjakob, H., and Apweiler, R. (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert review of proteomics* 3, 1-3

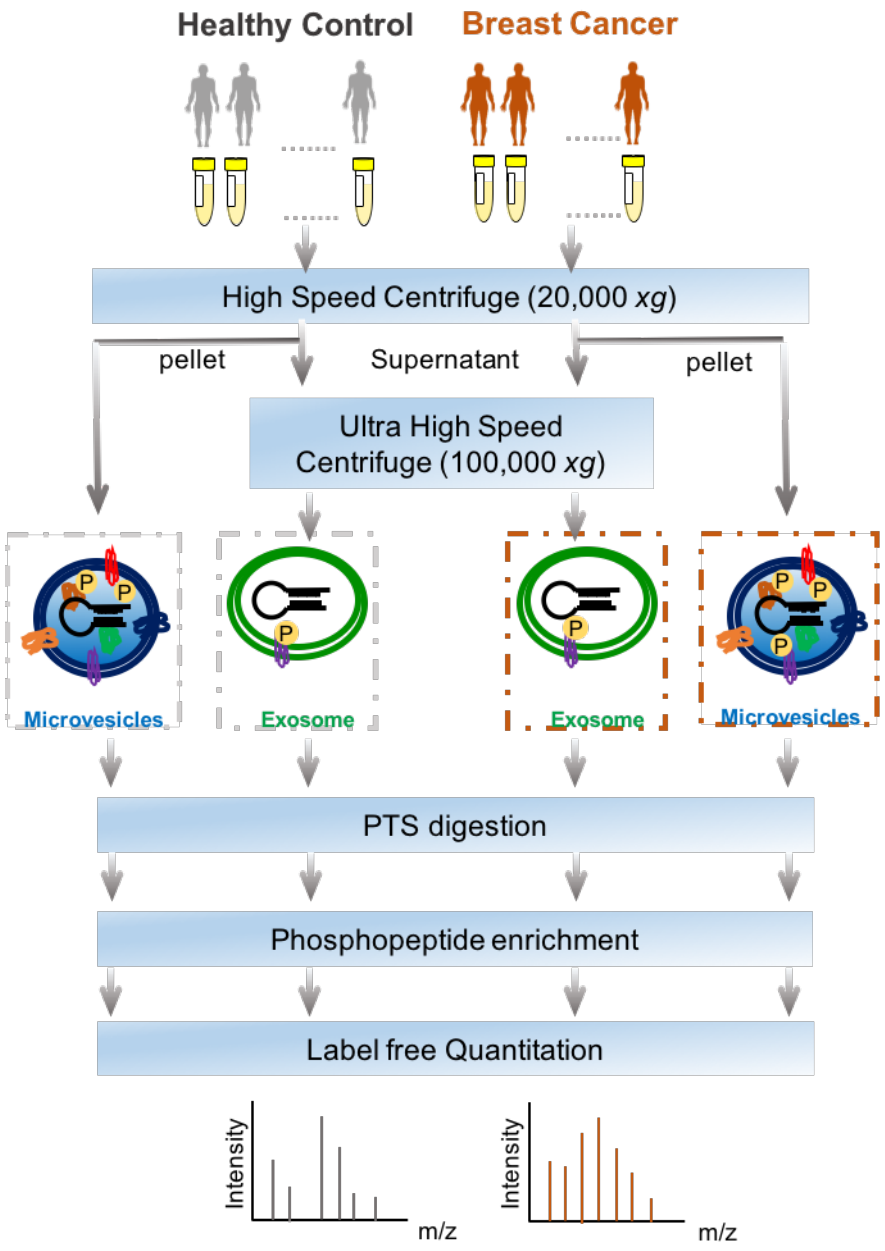

Figure 1.1 The workflow of EVs phosphoproteomics

The workflow for EVs phosphoproteomics of plasma samples from patients with breast cancer and healthy controls. EVs including microvesicles and exosomes were isolated through sequential high-speed centrifugation, followed by protein extraction, phase transfer surfactant digestion, and phosphopeptide enrichment for LC-MS analyses.

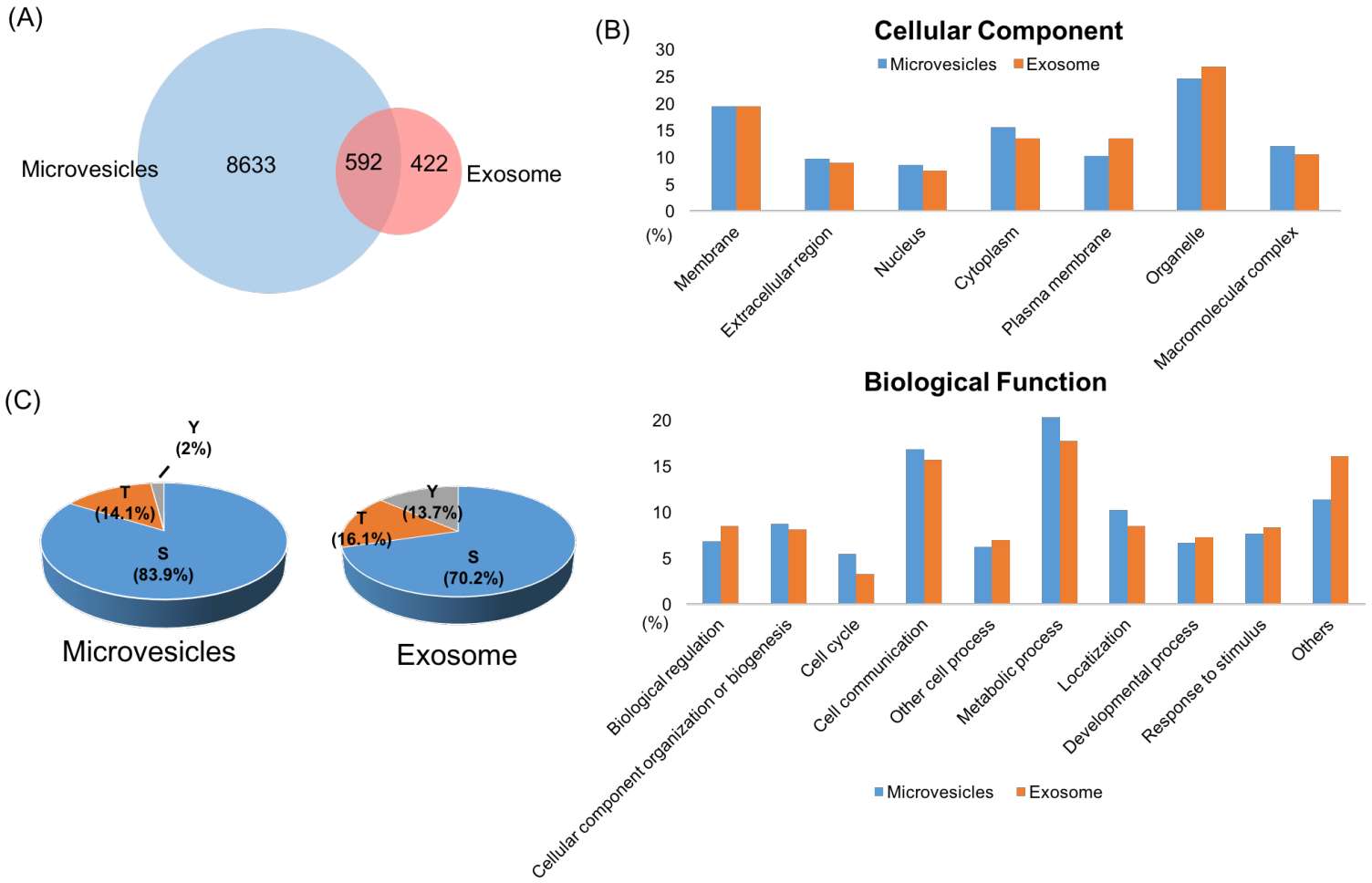

Figure 1.2 The identification result of EVs phosphoproteomics

(A) The Venn diagram showing the number of unique phosphopeptides identified in microvesicles and exosomes. (B) Classification of the identified phosphoproteins based on cellular component and biological function. (C) The distribution of serine/threonine/tyrosine (S/T/Y) phosphopeptides in microvesicles and exosomes.

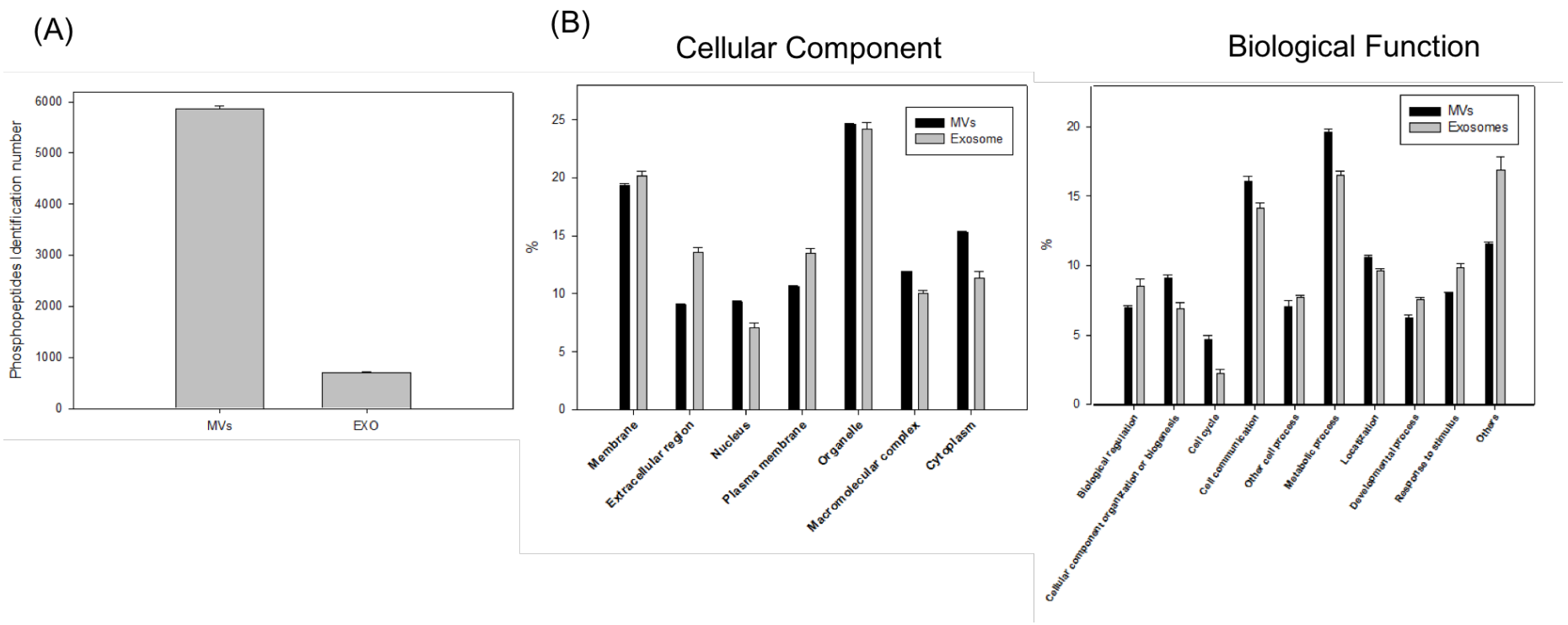

Figure 1.3 The identification comparison between microvesicles and exosomes.

(A) The bar chart showing the number of unique phosphopeptides identified in microvesicles and exosomes. The values indicated the mean identification numbers of technical replicates, the error bar shows the SD between replicates. (B) Classification of the identified phosphoproteins based on cellular component and biological function. The values indicated the mean of technical replicates; the error bar shows the SD between replicates.

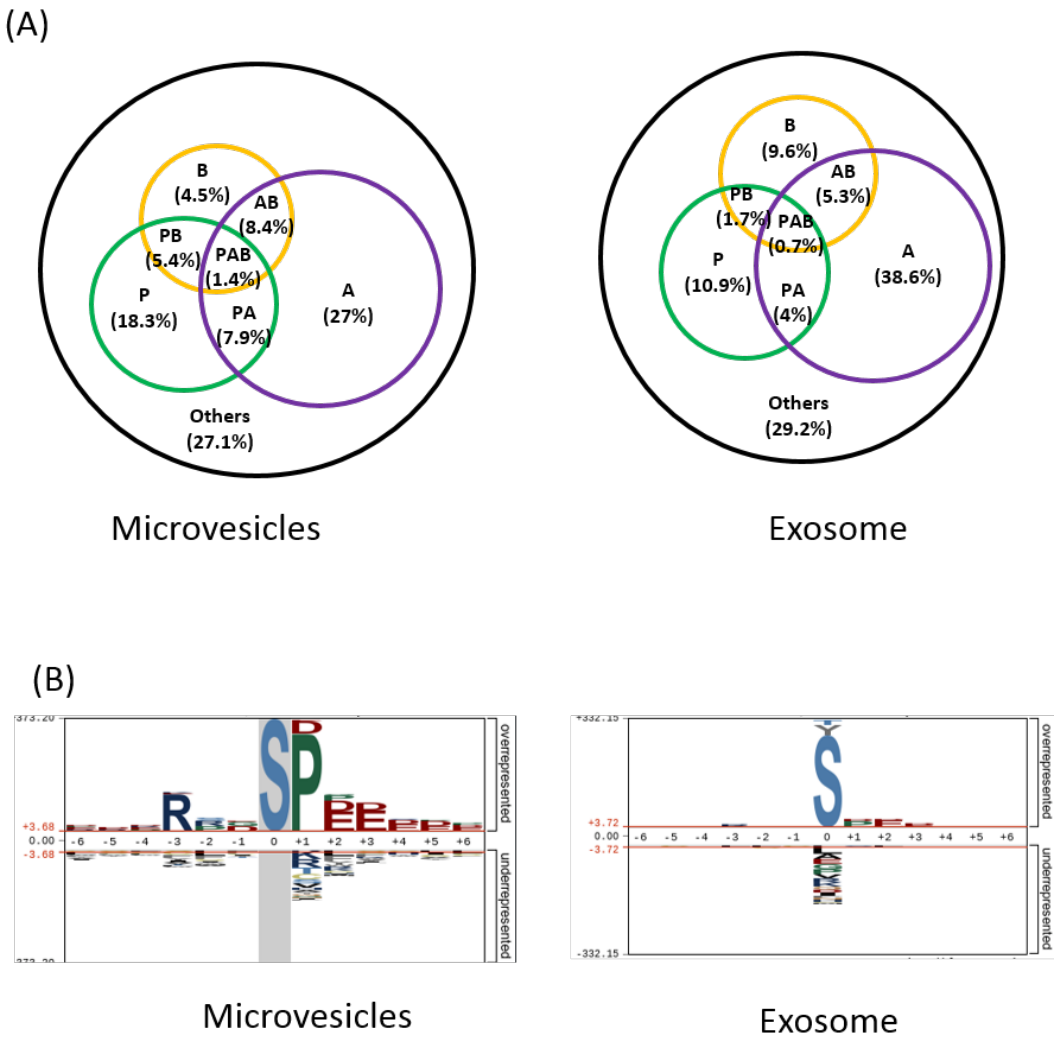

Figure 1.4 The classification and motif analysis of phosphosites.

(A) Classification of phosphosites based on kinase specificities (P, proline-directed; A, acidophilic; B, basophilic; others). (B) The summary of motifs were extracted from the sequence windows of identified probability >0.75 phosphorylation sites by pLogo.

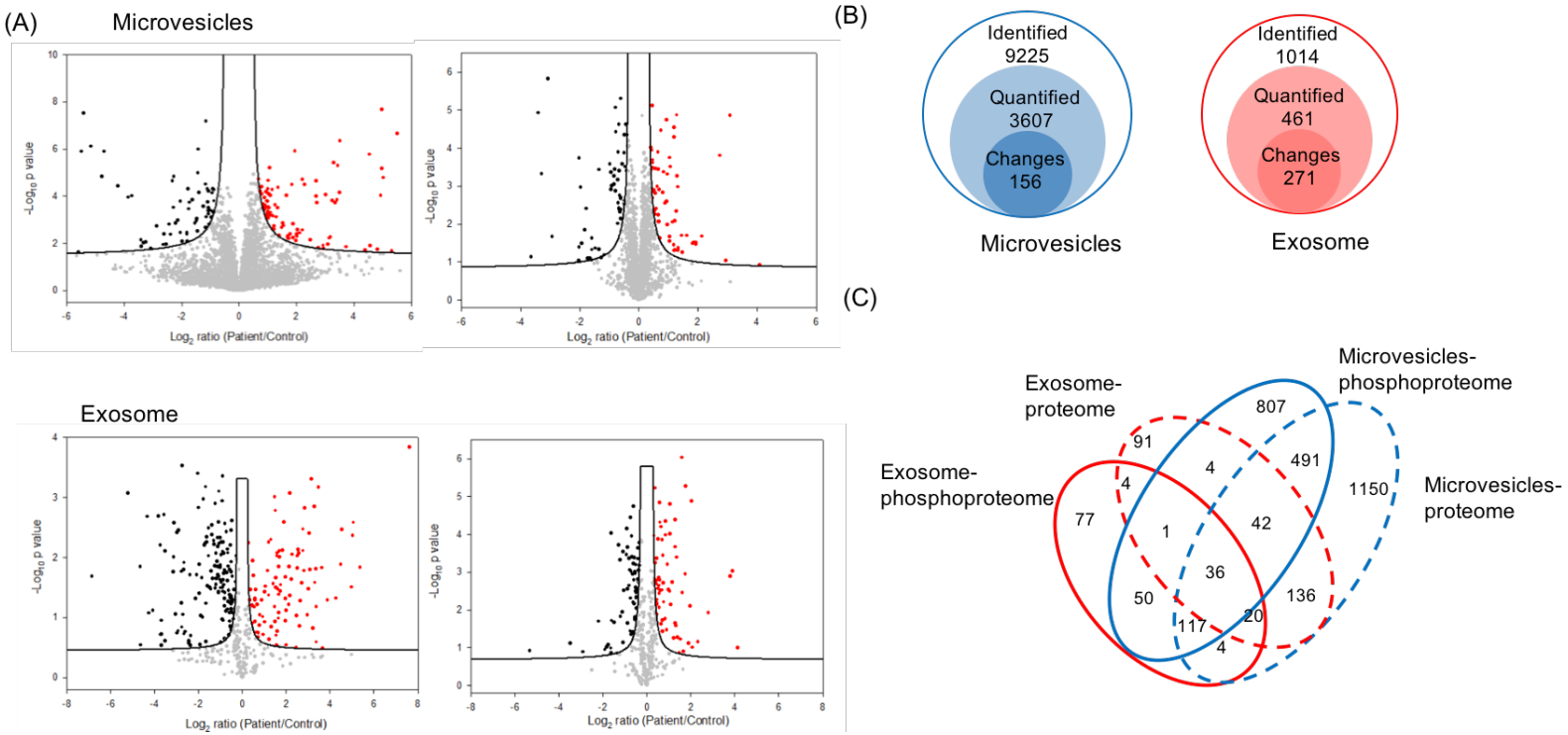

Figure 1.5 The Quantitation result of EVs phosphoproteomics between breast cancer and healthy control

(A) The volcano plots representing the quantitative analyses of the phosphoproteomes (Left) and proteomes (Right) of microvesicles and exosomes in patients with breast cancer vs. in healthy controls. Significant changes in proteins and phosphosites in breast cancer that were identified through a permutation-based FDR t test (FDR = 0.05; S0 = 0.2), based on three technical replicates. The significant up-regulated proteins and phosphosites are colored in red, and down-regulated are colored in black. (B) The numbers of identified phosphopeptides (class 1), quantified phosphosites (class 2), and significantly changed phosphosites (class 3) in label-free quantification. See supplementary figures and Dataset S1 for more detailed information. (C) The Venn diagram showing the protein overlap between phosphoproteomes and proteomes in microvesicles and exosome.

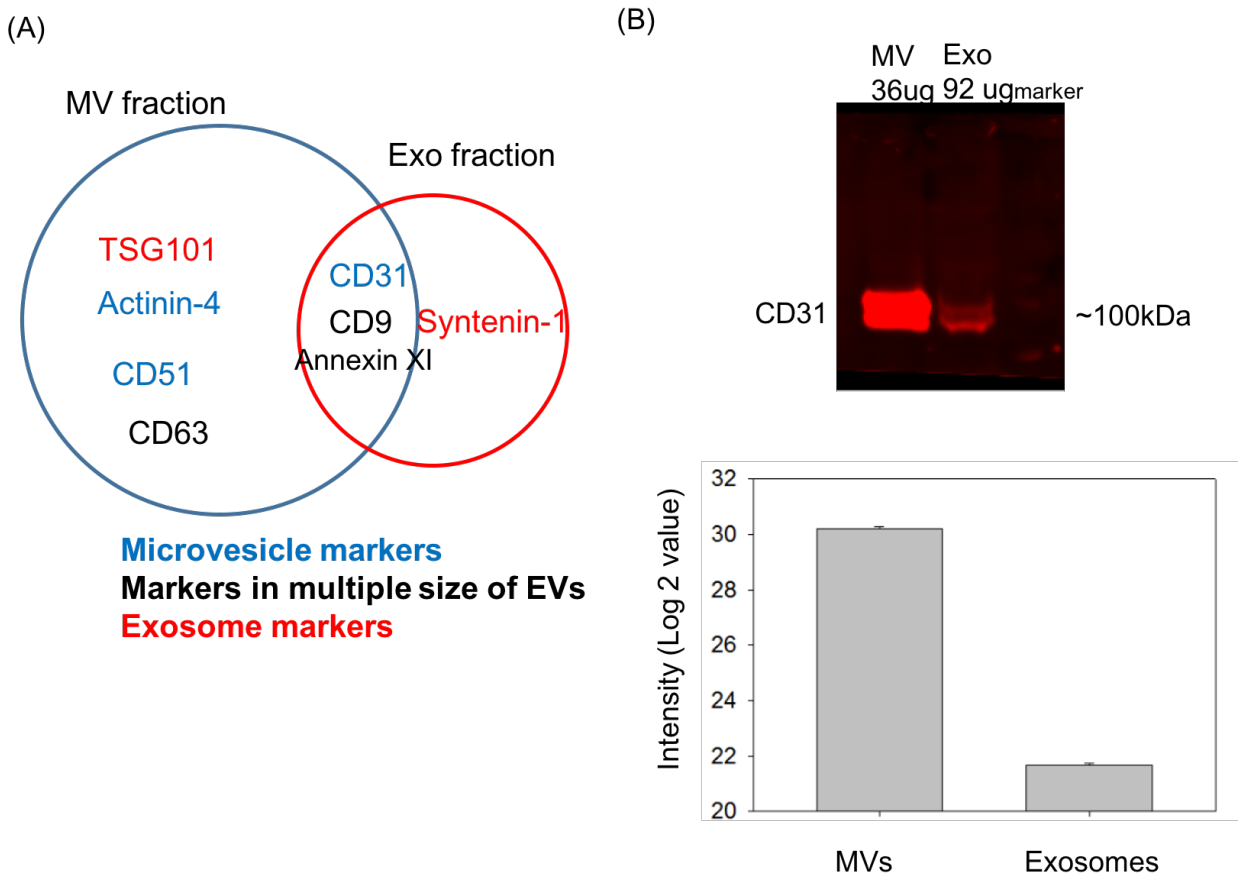

Figure 1.6 Examination of EVs isolation from plasma.

(A) The Venn diagram showing the common EVs markers present in MVs and exosome fractions through proteomic analyses. (B) Western blotting (WB) and MS data showing the purity of EV isolation. Two EV fractions were collected and analyzed by WB using antibody against CD 31, which is considered an endothelial-derived microvesicles marker. A total of 36 μg protein was used in MV fraction, and considering exosomes may possibly contain some plasma proteins, around 2.5-fold of protein amount of exosome fraction was used. MS data were extracted from two EV fractions, and the bar chart showed the intensity mean value with error bar of control and patient replicates.

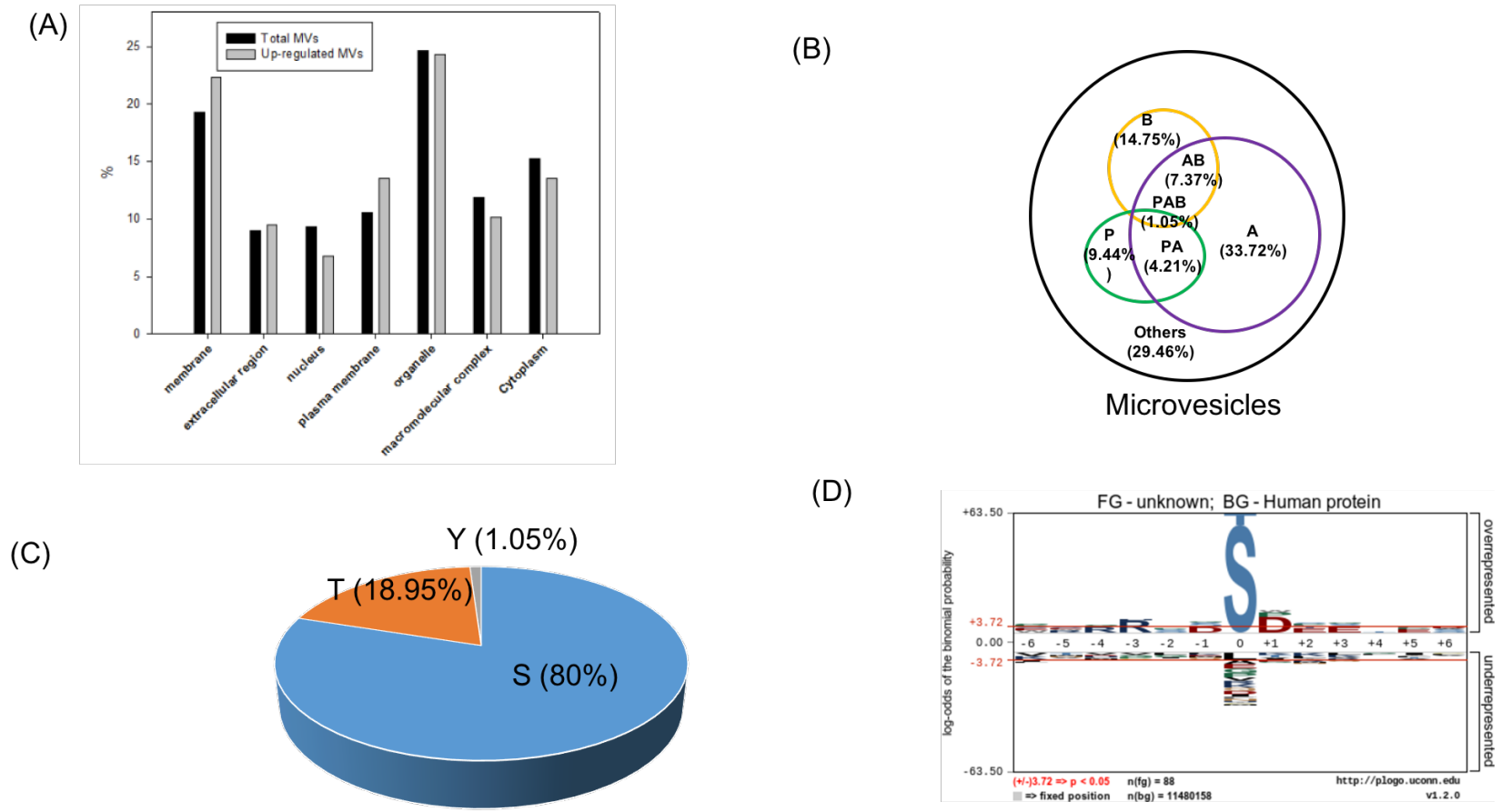

Figure 1.7 Comparison of MV phosphopeptides that showed an increase in patients with cancer.

(A) Comparison of cellular components of MV phosphopeptides that showed an increase in patients with cancer, with those of total phosphopeptides identified in MV. (B–D) Motif and the distribution of S/T/Y phosphopeptides that showed increase in patients with cancer in microvesicles.

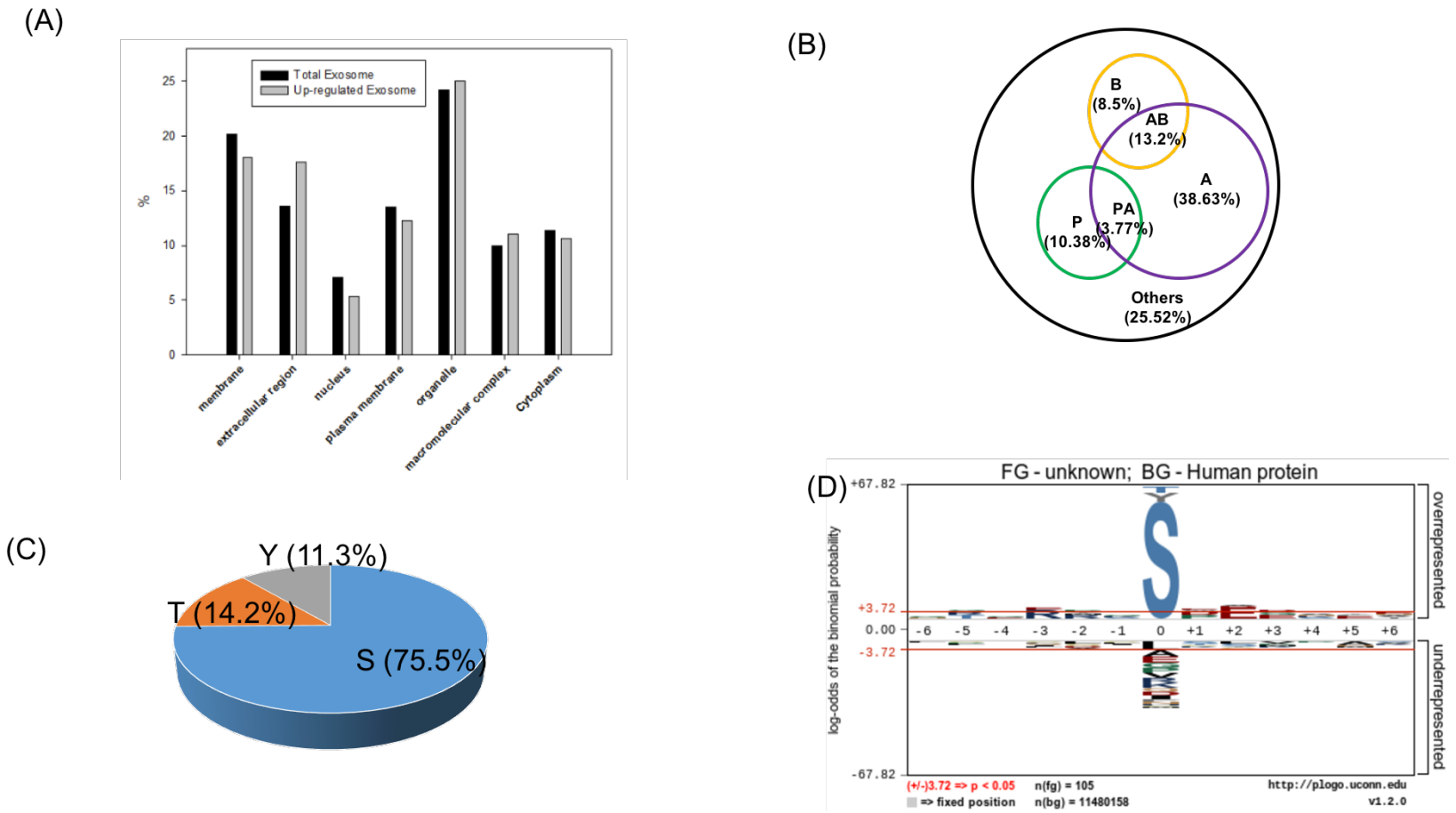

Figure 1.8 Comparison of exosome phosphopeptides that showed an increase in patients with cancer.

(A) Comparison of cellular components of exosome phosphopeptides that showed increase in patients with cancer with those of total phospho- peptides identified in exosome. (B–D) Motif and the distribution of S/T/Y phosphopeptides that showed increase in patients with cancer in exosomes.

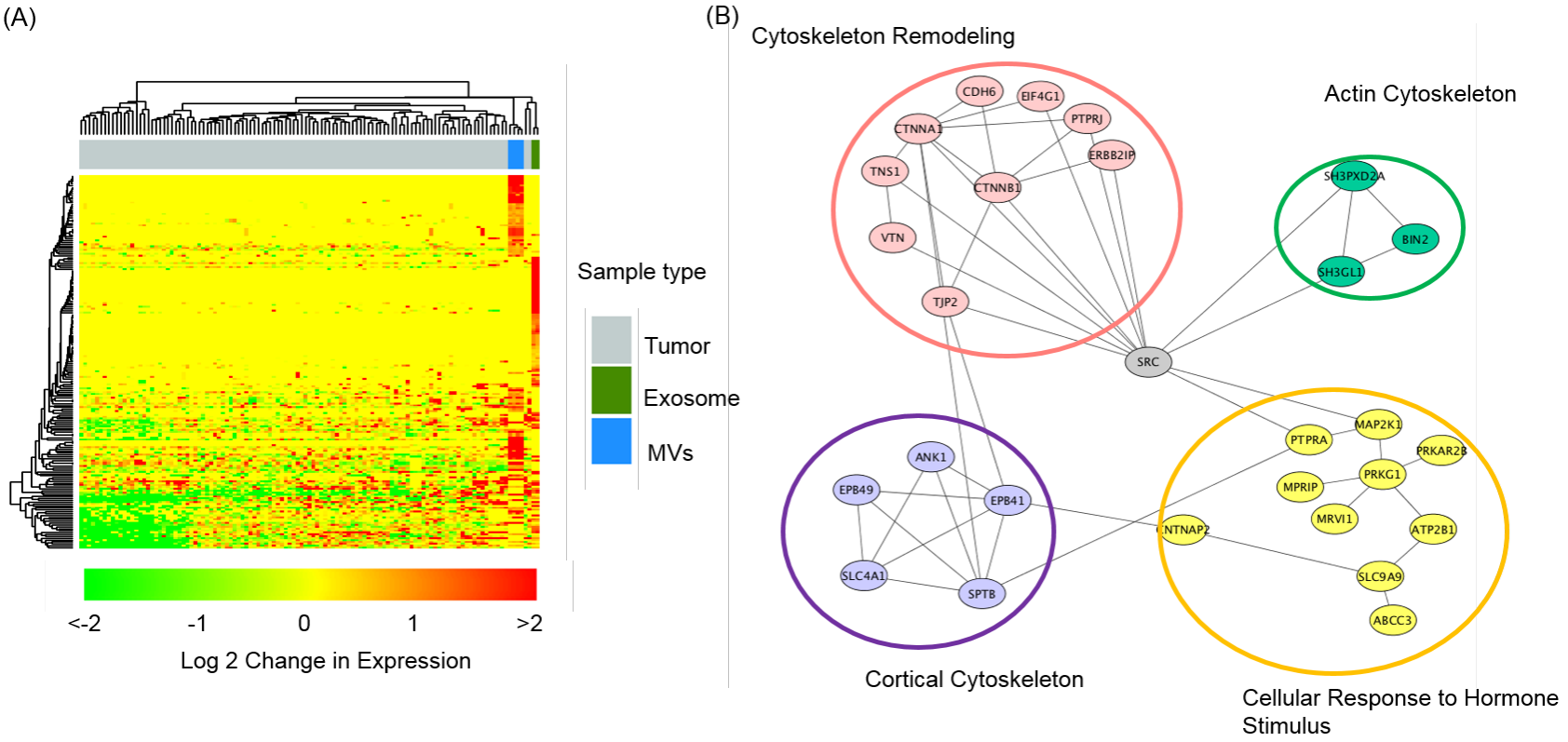

Figure 1.9 Networking analysis of up-regulated phosphoproteins

(A) The hierarchical clustering analysis of up-regulated phosphopeptides conveys the overlap between EVs in this study and breast cancer tissues by Mertins et al. (20). The top bars show the clustering of different samples, and gray represents the tumor samples analyzed by Mertins et al., whereas blue bars are replicates of MV analysis and cobalt green are exosome analyses in this study. The fold change is shown in log 2 value. (B) The STRING network analysis of up-regulated phosphoproteins in EVs.

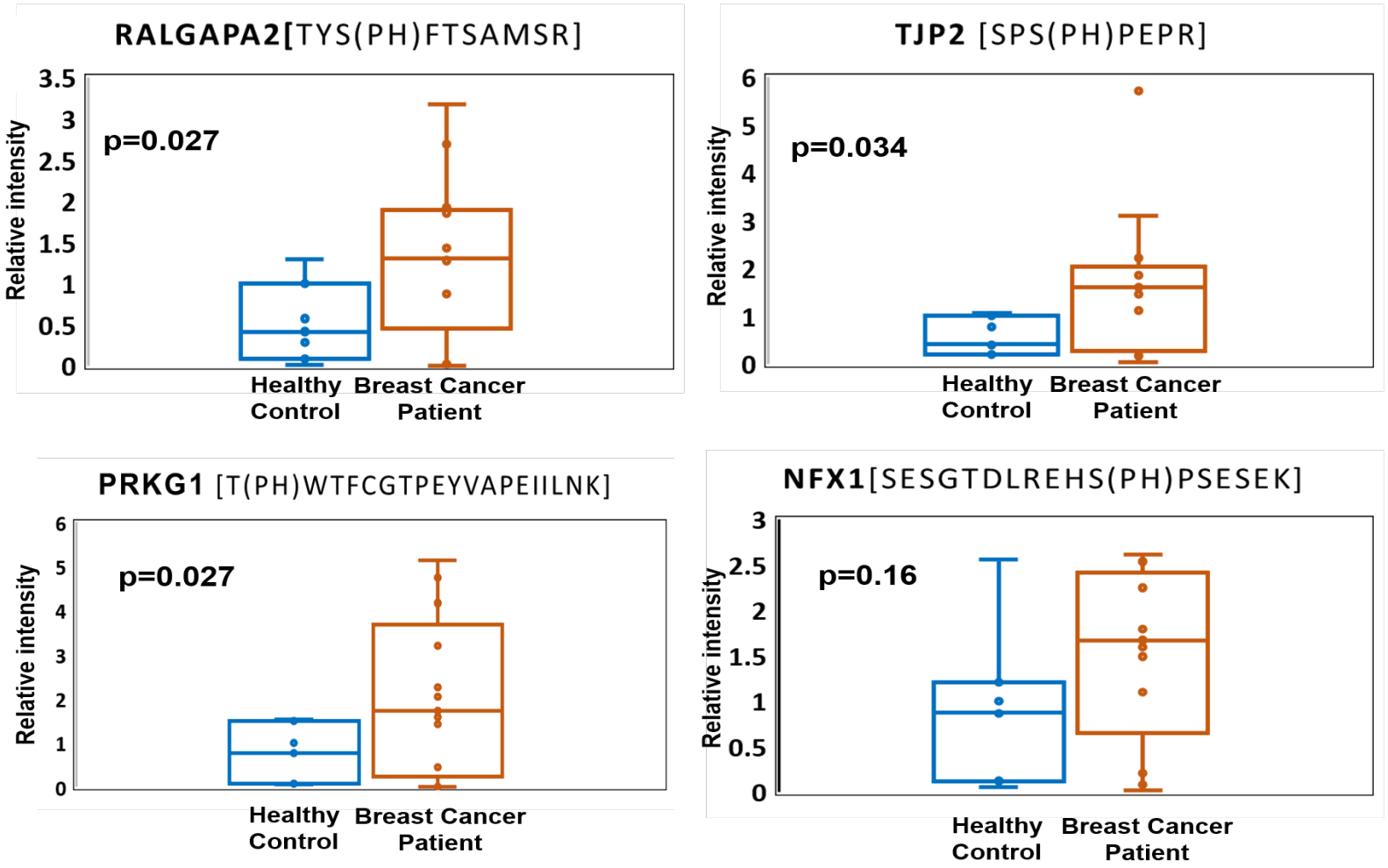

Figure 1.10 Four potential markers were validated in 13 patients with breast cancer and seven healthy individuals, using PRM. Three potential markers, RALGAPA2, PRKG1, and TJP2, show significant difference ($P < 0.05$) in patients with breast cancer compared with healthy control

CHAPTER 2 A PIPELINE FOR DISCOVERY AND VERIFICATION OF GLYCOPROTEINS FROM PLASMA-DERIVED EXTRACELLULAR VESICLES AS BREAST CANCER BIOMARKER

2.1. Summary

Glycoproteins comprise more than half of current FDA-approved protein cancer markers but the development of new glycoproteins as disease biomarkers has been stagnant. Here we present a pipeline to develop glycoproteins from extracellular vesicles (EVs) through integrating A pipeline for discovery and verification of glycoproteins from plasma-derived extracellular vesicles as breast cancer biomarkers quantitative glycoproteomics with a novel reverse phase glycoprotein array, and then apply it to identify novel biomarkers for breast cancer. EV glycoproteomics show promise in circumventing the problems plaguing current serum glycoproteomics and allowed us to identify hundreds of glycoproteins that have not been identified in serum. We identified 1,453 unique glycopeptides representing 556 glycoproteins in EVs, among which 20 were verified significantly higher in individual breast cancer patients. We further applied a novel glyco-specific reverse phase protein array to quantify a subset of the candidates. Together, this study demonstrates the great potential of this integrated pipeline for biomarker discovery.

2.2. Introduction

The emerging liquid biopsy underscores our unyielding goal of achieving non-invasive disease diagnosis through blood tests(1). With most proteins present in the blood being glycoproteins and aberrant glycosylation occurring in many diseases (2), it is not surprising that most common FDA approved clinically utilized biomarkers for cancer diagnosis and monitoring of malignant progression are glycoproteins. Examples are prostate-specific antigen (PSA) for prostate cancer, and carcinoembryonic antigen CEA for colon cancer (3). However, plasma or serum proteomes contain a dynamic range of 12 orders of magnitude in protein concentration, even a 95% reduction of major components such as human serum albumin still leaves a dynamic range of 12 orders of magnitude in concentrations in the sample. As a result, analyzing glycoproteins in blood-derived

plasma or serum to search for new biomarkers continues to face major challenges in terms of analytical sensitivity and depth (4, 5). With increasing evidence about their important roles in cell-cell communication and relevance in the transmission of pathogenic and signaling molecules in diseases, extracellular vesicles (EVs) have been exploited as attractive sources for biomarker discovery and disease diagnosis (6-8). Currently, most studies on EVs focus on mRNA and miRNA transfer, the role of proteins in EVs, in particular their post-translational modifications (PTMs) has been rarely exploited (9, 10). PTMs increase the functional diversity of the proteome and influence almost all aspects of cell biology and pathogenesis. Thus, many PTMs are routinely tracked as disease markers and used as molecular targets for developing target-specific therapies, such as the glycoproteins mentioned above. Given that the extracellular vesicles are membrane-encapsulated packages, EVs are believed to carry a large assortment of resident cell-surface glycoproteins (11). In theory, the glycoproteome of EVs should reflect their cellular origins and functions. A recent study has verified that the altered N-glycoproteome of urinary extracellular vesicles is associated with prostate cancer (12). Importantly, analyzing the glycoproteome in EVs instead of plasma or serum could eliminate the interference from highly abundant plasma components to a large extent, thus providing a wide dynamic range of detection and enabling the discovery of low-level glycoproteins at high sensitivity (as low as ng/mL) (13). We present here an integrated pipeline that profiles glycoproteins from EVs through quantitative glycoproteomics using pooled and individual samples and then validated several targets using a novel reverse phase glycoprotein array, termed polymer-based reverse phase glycoprotein array (polyGPA) (14). Since there are few glycosylation-specific antibodies available, verifying glycoproteins as biomarkers in clinical settings has remained a huge challenge. Although mass spectrometry (MS) has been the driving force in profiling glycans and glycoproteomes for biomarker research (15, 16), it is often necessary to enrich either glycoproteins or glycopeptides (17, 18) prior to MS analyses. With its typical requirements of fair amount of sample, multiple steps for sample preparation, and the commitment of a high performance instrument, MS-based glycoproteomics is typically used for in-depth profiling of glycoproteins during the discovery stage. With this pipeline, we identified 1,453 EV N-glycopeptides representing 556 N-glycoproteins, among which 20 EV glycoproteins showed significant elevation on 21 unique glycosylation sites in breast cancer patient samples. We applied polyGPA to further validate 5 glycoproteins with samples from another cohort of patient and healthy individuals. We demonstrate here the universal performance of this pipeline and its

value in discovering and validating glycoproteins in EVs as novel disease marker.

2.3. Experiment design

2.3.1. Plasma sample

The Indiana University Institutional Review Board approved the use of human plasma samples. In the global glycoproteomics experiment, blood samples were collected from 6 healthy females and from 18 breast cancer patients that obtained through the IU Simon Cancer Center. In the individual verification and PolyGPA validation experiment, blood sample from another 15 healthy controls and 41 breast cancer patients were collected and obtained through Susan G. Komen Tissue Bank and IU Simon Cancer Center. Plasma samples were collected by standard protocol, in brief, plasma sample processing was initiated within 30 min of blood draw to an ethylenediaminetetraacetic acid (EDTA) containing tube. Samples were spun for 30 min at 3500 rpm to remove all cell debris and platelets.

2.3.2. Extracellular Vesicles Isolation

The EVs isolation and digestion were performed according to the reported protocol (10). A total of 5.5 ml pooled plasma samples were collected from both healthy individuals and patients diagnosed with breast cancer for the global glycoproteomics experiment as technical replicates. For the individual verification and polyGPA validation experiments, 0.8ml of plasma was used. Plasma samples were centrifuged at 20,000 xg at 4 oC for 1hr. Pellets were washed with cold PBS and centrifuged again at 20,000 xg at 4 oC for 1 hr, the pellets were microvesicles. Supernatant of the first centrifugation was further centrifuged at 100,000 xg at 4 oC for 1hr. Pellets were washed with cold PBS and centrifuged at 100,000 xg for 1hr again. The pellets from ultra-high speed centrifugations were exosome.

2.3.3. Protein Digestion

The digestion was performed with phase transfer surfactant aided (PTS) digestion (19). Extracellular vesicles were solubilized in lysis buffer containing 12mM sodium deoxycholate (SDC), 12mM sodium lauroyl sarcosinate (SLS) and phosphatase inhibitor cocktail (Sigma-

Aldrich, St. Louis, MO) in 100mM Tris-HCl, pH8.5. Proteins were reduced and alkylated with 10 mM tris-(2-carboxyethyl)phosphine (TCEP) and 40 mM chloroacetamide (CAA) at 95 °C for 5 min. Alkylated proteins were diluted to 5 fold by 50mM triethylammonium bicarbonate (TEAB) and digested with Lys-C (Wako, Japan) in a 1:100 (w/w) enzyme-to-protein ratio for 3 hr at 37 °C. Trypsin was added to a final 1:50 (w/w) enzyme-to-protein ratio for overnight digestion. The digested peptides were acidified with trifluoroacetic acid (TFA) to final concentration of 0.5% TFA, and 250ul of Ethyl acetate was added to 250ul digested solution. The mixture was shaken for 2 min, then centrifuged at 13,200 rpm for 2 min to obtain aqueous and organic phases. The aqueous phase was collected and desalted using a 100 mg of Sep-pak C18 column (Waters, Milford, MA).

2.3.4. Glycoproteomics Enrichment

The glycopeptide enrichment was performed according to the reported protocol (14). Desalted peptides were oxidized with 10 mM sodium periodate in 50% ACN, 0.1% TFA at room temperature with shaking in the dark for 30 minutes. Excess sodium periodate was quenched by using 50 mM sodium sulfite for 15 minutes at room temperature with shaking in the dark. The samples were mixed with 50 µL/100 µL hydrazide magnetic beads for individual and pooled samples respectively, and incubated with vigorous shaking at room temperature overnight for the coupling reaction. Magnetic beads were washed sequentially with 400 µL/800 µL of 50% ACN, 0.1% TFA and 1.5 M NaCl for individual and pooled samples respectively, three times per solution for 1 minute per wash for the removal of non-coupled peptides. Beads were rinsed once with 100 µL/200 µL of 1x GlycoBuffer 2 (NEB) for individual and pooled samples respectively, and incubated with 3 µL/4 µL of PNGase F (NEB) in 100 µL/200 µL for individual and pooled samples respectively. N-glycans were cleaved by PNGase F. After desalting, the released former N-glycopeptides were analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).

2.3.5. LC-MS/MS Analysis

The glycopeptides were dissolved in 4 µL of 0.3% formic acid (FA) with 3% ACN and injected into an Easy-nLC 1000 (Thermo Fisher Scientific). Peptides were separated on a 45 cm in-house

packed column (360 µm OD × 75 µm ID) containing C18 resin (2.2 µm, 100Å, Michrom Bioresources) with a 30 cm column heater (Analytical Sales and Services) set to 50 °C. The mobile phase buffer consisted of 0.1% FA in ultra-pure water (buffer A) with an eluting buffer of 0.1% FA in 80% ACN (buffer B) run over either with a 45 min or 60 min linear gradient of 6%-30% buffer B at flow rate of 250 nL/min. The Easy-nLC 1000 was coupled online with a Velos Pro LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific). The mass spectrometer was operated in the data-dependent mode in where The 10 most intense ions were subjected to collision-induced dissociation (CID) fragmentation (normalized collision energy (NCE) 30%, AGC 3e4, max injection time 100 ms) for each full MS scan (from m/z 350-1500 with a resolution of 30,000 at m/z 400).

2.3.6. Data Processing

The raw files were searched directly UniprotKB database version Jan2015 with no redundant entries using MaxQuant software (version 1.5.6.1) (20) with the Andromeda search engine. Initial precursor mass tolerance was set to 20 p.p.m. and the final tolerance was set to 6 p.p.m., and ITMS MS/MS tolerance was set at 0.6 Da. Search criteria included a static carbamidomethylation of cysteines (+57.0214 Da) and variable modifications of (1) oxidation (+15.9949 Da) on methionine residues, (2) acetylation (+42.011 Da) at N-terminus of protein, and (3) deamidation (+0.984Da) on asparagine residues were searched. Search was performed with Trypsin/P digestion and allowed a maximum of two missed cleavages on the peptides analyzed from the sequence database. The false discovery rates of proteins, peptides and phosphosites were set at 0.01. The minimum peptide length was six amino acids, and a minimum Andromeda score was set at 40 for modified peptides. The glycosylation sites were selected based on the matching to the N-X-S/T (X not Pro) motif. A site localization probability of 0.75 was used as the cut-off for localization of glycosylation sites. All the peptide spectral matches and MS/MS spectra can be viewed through MaxQuant viewer.

2.3.7. Quantitative Data Analysis

All data was analyzed using the Perseus software (version 1.5.4.1) (21). For quantification of both proteomic and glycoproteomic datasets, the intensities of peptides and glycosites were

derived from MaxQuant, and the missing values of intensities were replaced by normal distribution with a downshift of 1.8 standard deviations and a width of 0.3 standard deviations. The significantly increased glycosites or proteins in patient samples were identified by their p-value from a two sample t-test with a permutation-based FDR cut-off 0.05 with S0 set on 0.2 for all of data sets. In the individual glycoproteomics data, the intensities were first normalized by subtracting the median of total intensity, missing values were imputed by normal distribution of each individual samples with a downshift of 1.8 standard deviations and a width of 0.3 standard deviations. The imputed data set was further normalized by z-score within each dataset, and the p-value was calculated by two sample t-test.

2.3.8. Periodate oxidation of plasma EVs

Human plasma microvesicle pellets from healthy and breast cancer-diagnosed individuals were resuspended with 30 µL of 2% SDS in 100 mM sodium acetate, pH 5.5. Solution was heated for 10 minutes at 95 °C to lyse microparticles. Protein concentration was measured by the bicinchoninic acid (BCA) assay. The proteins were then oxidized by 10 mM sodium periodate with shaking in the dark at room temperature for 30 minutes. The excess sodium periodate was quenched by 50 mM sodium sulfite with shaking in the dark for 15 min. The oxidized sample was then denatured in 2% SDS and 2% 2-mercaptoethanol with boiling for 5 minutes.

2.3.9. PolyGPA

The synthesis of PolyGPA reagent was performed according to reported protocol (14). Nitrocellulose membranes were incubated with the diluted polyGPA reagent overnight at 4 °C. The coated membranes were air-dried. Prepared oxidized samples were printed on each membrane using a microarray printing pin (Arrayit® SMP15B). The membrane was washed with 4% SDS in TBST for three times and then TBST once, 5 min per wash, and then blocked with 3% BSA in TBST and probed with a primary protein antibody. Membranes were washed three times with TBST, 5 minutes per wash, and incubated with the corresponding secondary antibody linked with HRP for tyramide-based signal amplification. Then, membranes were incubated with 5 µM biotinyl tyramide in 0.003% H_2O_2 in 100 mM borate buffer (pH 8.5) for 10 min in the dark. The

membranes were washed with TBST 3 times, 5 minutes per wash, and then probed with IRDye® 680RD Streptavidin (LI-COR Biosciences). Membranes were washed 3 times with TBST, 5 minutes per wash and 2 times with DI water. Finally, membranes were scanned using an infrared imaging system (LI-COR Odyssey®) and the fluorescent signals were recorded and quantified using Image Studio (LI-COR Biosciences). After quantification, data was exported to R 3.4 for further statistical analysis, in brief the mean Intensity signals where used to perform a Mann–Whitney U test to compare the Intensities among the control and patient groups.

2.3.10. Dynamic Light Scattering (DLS)

To characterize the size of EVs, DLS were performed by Malvern Nano-S Zetasizer, at Birck Nanotechnology Center, Purdue University. Due to the Brownian motion of the particle, the velocity distribution of nano-particle movement can be analyzed by measuring dynamic fluctuations of light scattering intensity, which yields the particle diameter by Stokes-Einstein equation indirectly. After 20 K and 100 K centrifugation, MVs and exosomes were isolated. The EVs pellets were resuspended in 1000ul and 100ul of PBS buffer, respectively. The background was set as PBS buffer with the refraction index at 1.33 equilibrated at 25℃.

2.4. Result

2.4.1. Identification of 1,453 unique N-glycopeptides from plasma EV

An overview of EV glycoprotein biomarker pipeline and its application to the identification of potential breast cancer biomarkers is illustrated in Fig.2.1. Global quantitative Nglycoproteomics was carried out with EVs, including microvesicles (MVs) and exosomes, using both pooled and individual samples from healthy and patient plasma, to generate a candidate biomarker list. Plasma samples were collected and pooled from healthy individuals (n= 18) and from patients diagnosed with breast cancer (n=18). MVs and exosomes were isolated from human plasma through high speed and ultra-high speed centrifugation, respectively. The isolation specificity was evaluated using dynamic light scattering (DLS) (Fig.2.2A), immunoassay with an EV marker antibody, and mass spectrometry (MS) (Figure 2.3). The DLS data indicated that most MVs isolated after 20K

centrifugation are in the range of 100-1000 nm while exosomes isolated by 100K centrifugation are in the range of 30-100nm. MS analyses identified several protein markers only in microvesicles or exosomes, but at the same time a few surface markers were identified in both microvesicles and exosomes, indicating either current markers for exosome and microvesicles are not totally specific or the differential centrifugation method is not entirely specific. Western Blotting was carried out with the antibody against CD31 which is considered an endothelial derived microvesicles marker and the data showed that CD31 was indeed mainly identified in microvesicles. After isolation, EVs were lysed, proteins were extracted and enzymatically digested with LysC and trypsin, followed by the hydrazide chemistry to enrich pre-oxidized glycopeptides. N-glycopeptides were recovered using PNGase F and analyzed by nanoflow LC-MS/MS. For each glycopeptide sample, three technical replicates were performed and label free quantitation was performed to measure glycopeptides in EV samples in the plasma of control and breast cancer patient samples.

We identified 1,453 unique glycopeptides, including 1,337 from microvesicles and 447 from exosomes, representing 526 and 164 glycoproteins in MV and exosomes, respectively (Fig.2.2B). Gene ontology analysis of the glycoproteins indicated a significant portion of the identified glycoproteins are from membrane, extracellular region, and organelles (Fig.2.2C). Overall, similar cellular components were observed for MV and exosomes. There is also significant overlap of identified glycopeptides and glycoproteins in MV and exosomes. With only 30 glycoproteins being unique in exosomes, we reasoned that it is not critical to differentiate glycoproteins in MV from those in exosomes for disease biomarker discovery and therefore all following data collected in MVs and exosomes in this study were combined and analyzed as EV N-glycoproteomes.

The current data reported here represents one of the largest N-glycoproteomic datasets using serum or plasma as the source. For direct comparison, we carried out a conventional N-glycoproteomic study using the breast cancer plasma samples. The conventional workflow with plasma samples resulted in a larger portion of high abundant plasma glycoproteins while EV glycoproteomics identified more glycoproteins in low abundance (Fig.2.4A). We further examined the identified EV N-glycoproteins against previous reported serum/plasma glycoproteins. Strikingly, about one quarter (126) of glycoproteins have not been previously reported as serum/plasma glycoproteins (Fig. 2.4B). The data supports our hypothesis that EVs are an ideal source to identify novel glycoproteins as potential disease biomarkers.

2.4.2. Cancer-specific glycoproteins in EV

Label-free quantitation of glycopeptides was performed to identify a list of glycoproteins changing in breast cancer. We quantified 1106 unique glycosites and identified 77 glycopeptides with a significant difference in abundance in breast cancer patients versus healthy controls (Fig.2.5A). The difference in glycopeptides may be a result of changes in protein expression or changes of glycosylation on specific sites. To distinguish these factors, we also performed label-free quantitation of total EV proteomes. We identified 1,996 proteins, only 177 of which were also identified with glycopeptide enrichment. Therefore, analyses of the glycoproteme contributes to a deeper coverage of the EV proteome. Quantitative analyses of EV proteomes revealed strikingly similar expression of most proteins in healthy individuals and cancer patients (Fig.2.5B). In comparison, there are a larger number of glycopeptides with significant changes in patient samples, indicating that these glycosylation differences between cancer patients and healthy individuals are not due to changes in protein expression, and thus reflect true cancer patient-specific glycosylation. We then carried out label-free quantitative EV glycoproteomics with individual plasma samples from 18 patients with breast cancer and 10 healthy controls. Glycoproteins with significantly increased glycosylation in patient samples were identified by the p-value from a two sample t-test with a permutation-based FDR cut-off 0.05 with S0 set on 0.2. The imputed data set was further normalized by z-score for the heatmap analysis and together, we identified a total of 20 glycoproteins specific in patients with 21 unique glycosylation sites (P-value <0.05) (Fig.2.5C).

2.4.3. Verification of specific glycoprotein changes in cancer patients via polyGPA

We reason that breast cancer is extremely heterogeneous and instead of identifying a single diagnostic biomarker, the identification of a panel of candidate glycoproteins that reflect the onset and progression of breast cancer would offer better prognostic value. Validation of biomarkers has been carried out using antibody-based Sandwich assays such as ELISA or targeted quantitative MS methods like selected reaction monitoring (SRM) and multiple reaction monitoring (MRM). However, there are virtually no existing antibodies specific for glycosylated proteins. On the other hand, the development of SRM/MRM assays requires a great deal of efforts including the high cost of synthetic stable isotope labeled peptides, in particular here formerly N-glycosylated

peptides. Thus, in an effort to verify increased glycoproteins in specific cancer patients, we applied a novel reverse phase protein array specific for glycoproteins to quantify individual EV glycoproteins in plasma from breast cancer patients and healthy individuals. We have recently developed a three-dimensionally functionalized reverse phase protein array, polyGPA, to validate glycoproteins in high throughput. PolyGPA uses hydroxyamino dendrimer-modified nitrocellulose to covalently capture pre-oxidized glycans on glycoproteins, followed by on-membrane detection using the same validated antibodies as in typical reverse phase protein arrays. Although no glycosylation specific antibody or lectin is used, any change in polyGPA signal is attributed to the change in overall glycosylation of targeted glycoprotein. In addition, we demonstrated that polyGPA's sensitivity is much higher than RPPA (over 10-fold signal increase) for the same protein concentration, likely due to improved orientation of glycoproteins during their glycan binding to the polyGPA membrane, exposing more epitopes for increased overall signal. We prioritized the glycoproteins for further verification by polyGPA through their biological relevance to cancer in previous studies and availability of their antibodies which are validated by Human Protein Atlas (HPA) project for high specificity. Among the glycoproteins that show significant increase in breast cancer patients (Fig.2.5C), some are known plasma/serum glycoproteins while others have never been detected from blood. Interestingly, 70% of the glycoproteins on the list have previously been identified from cancer tissues (Figure 2.7)(22), highlighting the important feature of this biomarker strategy which did not require an invasive biopsy but rather used EVs as the source to identify biomarkers previously reported in cancer tissue studies. We selected 6 EV glycoproteins, a membrane protein Lymphocyte antigen 6 complex locus protein G6f (LY6G6F), a multimeric plasma glycoprotein von willebrand factor (VWF), CD147/basigin (BSG), Complement C1q subcomponent subunit A (C1QA), Angiopoietin-1 (ANGPT1/Ang1), and Cadherin-6 (CDH6) for further verification with another cohort of plasma samples from 28 breast cancer patients and 10 healthy controls. EVs were isolated from plasma samples, lysed, pre-oxidized and each individual sample was printed onto the polyGPA membranes and unfunctionalized membranes as in regular RPPA. Specific protein antibodies were then used to detect and quantify endogenous LY6G6F, VWF, BSG, C1QA, ANGPT1, and CDH6 signals in individual samples. As shown in Figure 2.6, measurements by polyGPA showed much better sensitivity because of significantly reduced sample complexity after the enrichment of glycoproteins on the functionalized membrane and better orientation of

glycoproteins for epitope detection by the antibodies. This enhanced sensitivity proved to be critical for the detection of proteins with much lower abundances, such as BSG, C1QA, ANGPT1, and CDH6, and their protein signals were barely detectable in RPPA (Figure 2.6c-f). Five out of six glycoproteins, except CDH6, showed statistically significant specificity ($p<0.05$) for breast cancer. The quantitative measurements with polyGPA and RPPA also allowed us to identify whether glycosylation elevation is due to changes in protein expression or changes in glycosylation. Significant elevation in both polyGPA and RPPA for LY6G6F. The increase in breast cancer patients was clearly observed in polyGPA for VWF, but the difference is small in RPPA (the distinction is largely due to one outlier; Figure 5b), indicating that that the glycosylation elevation in cancer patients is likely due to changes in patient-specific glycosylation. As stated above, due to low abundance, BSG, C1QA, ANGPT1 and CDH6 could only be quantified by polyGPA, further highlighting its uniqueness and high sensitivity for clinical samples.

2.5. Discussion

Glycosylated proteins are one important class of proteins that play important roles in a wide range of cellular functions and have also been utilized for disease diagnosis. Development of new glycoproteins as potential biomarkers, however, has struggled due to the lack of good tools. The purpose of this study was to continue our efforts to develop novel glycoproteins as potential disease biomarkers by proposing new strategies and new analytical platforms. We tested the hypothesis that, to overcome the great complexity of protein glycosylation at the presence of thousands of proteins in serum in which a number of highly abundant serum proteins are glycoproteins, glycoproteins from EVs are valuable sources for biomarker discovery and disease diagnosis. Here, we reported in-depth analyses of N-glycoproteomes in plasma EVs and demonstrated the feasibility of developing glycoproteins as potential breast cancer biomarkers. With multiple high abundant glycoproteins that prevent us from exploring disease-relevant, typically low abundant glycoproteins in blood, this method relies on glycoproteins EVs to efficiently identify many glycoproteins that are difficult to detect using existing methods.

This study also addresses a major issue in the development of glycoproteins for biomarker discovery, i.e., how to validate specific glycoproteins in high throughput. Without glycospecific antibodies, SRM/MRM appeared as the only choice but it requires considerable efforts including the synthesis of isotopic labeled formerly glycosylated peptides. Instead, we introduced polyGPA

as an alternative and novel high throughput method for simple, sensitive quantification of glycoproteins in array format. Using glyco-specific, 3-dimensional functionalized membrane to capture glycoproteins followed by detection using high quality antibodies, the new platform allowed us to measure glycoproteins in multiple clinical samples in parallel. Here we developed a novel biomarker discovery pipeline that focuses on glycoproteins from plasma-derived EVs and integrates high performance LC-MS/MS for candidate discovery with novel glycoprotein-specific RPPA for v. We applied the pipeline to identify EV glycoproteins as novel breast cancer biomarkers. Using data-dependent LC-MS/MS-based EV N-glycoproteomics, we identified 1,453 unique N-glycopeptides in the plasma EVs from breast cancer samples, representing 556 glycoproteins that include not only known plasma proteins spanned several orders of magnitude of abundance, but also non-plasma proteins that were identified only from tissues previously. Among them, 20 glycoproteins were quantified with significantly elevated level in breast cancer patients and we further validate 5 glycoproteins with separate cohort of breast cancer patients and healthy controls. The 5 validated glycoproteins all have been directly linked to or implicated with cancer according to previous studies. LY6G6F (G6f) is a type I transmembrane protein and putative cell-surface receptor encoded by a gene in the MHC. Its phosphorylation has been previously related to downstream signaling pathways including Ras-MAP kinase pathway (23). VWF is a multimeric plasma glycoprotein and was previously discovered to be highly enriched in mesenchymal stem/stromal cells-derived EVs (24). Besides its essential role in hemostasis, there are growing studies connecting it to cancer (25), such as its modulation on angiogenesis and apoptosis(26) and in tumor metastasis (27). BSG (CD147 or basigin) is also a transmembrane glycoprotein that is highly expressed by various cancer cells such as malignant melanoma cells (28). The full length of BSG was identified in microvesicles shedding from lung carcinoma cells (29). BSG is strongly related to cancer progression, enhancing cancer proliferation and VEGF production (30). It was reported that it promotes tumor cell glycosylation through facilitating lactate transport (28). Complement C1q has recently been discovered to act as tumor-promoting factor by facilitating adhesion, migration and proliferation of cancer cells as well as angiogenesis and metastasis (31). Angiopoietin-1 (ANGPT1) has been discovered in multiple human breast cancer cell lines such as MCF-7 and has been shown to play an important role in tumor angiogenesis (32). Verification of glycoproteins by polyGPA is a unique element of our pipeline, providing a simple and relatively high throughput method to prioritize a list of candidates meriting

further validation with larger, heterogeneous patient cohorts. However, the limitations of polyGPA for clinical validations need to be noticed. First, like other RPPA, its applications are highly dependent on the availability of validated, high quality antibodies for any novel candidate. Second, polyGPA only measures the overall glycosylation in a protein. For a glycoprotein with multiple glycosylation sites, polyGPA may not be sensitive enough to a glycosylation change on a specific site. As shown in Fig. 2.8, side-by-side measurement by polyGPA and LC-MS/MS of the same plasma samples from patients and healthy controls revealed in general attenuated difference on polyGPA compared to the difference observed by MS. For example, the relative intensity of ANGPT1 in a breast cancer patient and a healthy individual is almost equal, while MS detected glycosylation elevation in the patient only on site 122. Since ANGPT1 has at least five N-glycosylation sites, it is conceivable that the glycosylation on the individual site might have been elevated drastically in patient samples but the overall glycosylation level has minimal change.

2.6.Data Access

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) with Project accession number PDX00757 via the PRIDE partner repository (33).

2.7. Reference

1. Yong, E. (2014) Cancer biomarkers: Written in blood. *Nature* 511, 524-526
2. Pinho, S. S., and Reis, C. A. (2015) Glycosylation in cancer: mechanisms and clinical implications. *Nature reviews. Cancer* 15, 540-555
3. Blixt, O., and Westerlind, U. (2014) Arraying the post-translational glycoproteome (PTG). *Curr Opin Chem Biol* 18, 62-69
4. Hanash, S. M., Pitteri, S. J., and Faca, V. M. (2008) Mining the plasma proteome for cancer biomarkers. *Nature* 452, 571-579
5. Geyer, P. E., Kulak, N. A., Pichler, G., Holdt, L. M., Teupser, D., and Mann, M. (2016) Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* 2, 185-195

6. Melo, S. A., Luecke, L. B., Kahlert, C., Fernandez, A. F., Gammon, S. T., Kaye, J., LeBleu, V. S., Mittendorf, E. A., Weitz, J., Rahbari, N., Reissfelder, C., Pilarsky, C., Fraga, M. F., Piwnica-Worms, D., and Kalluri, R. (2015) Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature* 523, 177-182
7. Gonzales, P. A., Pisitkun, T., Hoffert, J. D., Tchapyjnikov, D., Star, R. A., Kleta, R., Wang, N. S., and Knepper, M. A. (2009) Large-scale proteomics and phosphoproteomics of urinary exosomes. *Journal of the American Society of Nephrology : JASN* 20, 363-379
8. Boukouris, S., and Mathivanan, S. (2015) Exosomes in bodily fluids are a highly stable resource of disease biomarkers. *Proteomics Clin Appl* 9, 358-367
9. Moreno-Gonzalo, O., Villarroya-Beltri, C., and Sanchez-Madrid, F. (2014) Post-translational modifications of exosomal proteins. *Front Immunol* 5, 383
10. Chen, I. H., Xue, L., Hsu, C. C., Paez, J. S., Pan, L., Andaluz, H., Wendt, M. K., Iliuk, A. B., Zhu, J. K., and Tao, W. A. (2017) Phosphoproteins in extracellular vesicles as candidate markers for breast cancer. *Proc Natl Acad Sci U S A* 114, 3175-3180
11. Gerlach, J. Q., and Griffin, M. D. (2016) Getting to know the extracellular vesicle glycome. *Molecular bioSystems* 12, 1071-1081
12. Saraswat, M., Joenvaara, S., Musante, L., Peltoniemi, H., Holthofer, H., and Renkonen, R. (2015) N-linked (N-) glycoproteomics of urinary exosomes. [Corrected]. *Mol Cell Proteomics* 14, 263-276
13. Sok Hwee Cheow, E., Hwan Sim, K., de Kleijn, D., Neng Lee, C., Sorokin, V., and Sze, S. K. (2015) Simultaneous Enrichment of Plasma Soluble and Extracellular Vesicular Glycoproteins Using Prolonged Ultracentrifugation-Electrostatic Repulsion-hydrophilic Interaction Chromatography (PUC-ERLIC) Approach. *Mol Cell Proteomics* 14, 1657-1671
14. Pan, L., Aguilar, H. A., Wang, L., Iliuk, A., and Tao, W. A. (2016) Three-Dimensionally Functionalized Reverse Phase Glycoprotein Array for Cancer Biomarker Discovery and Validation. *J Am Chem Soc* 138, 15311-15314
15. Ruhaak, L. R., Miyamoto, S., and Lebrilla, C. B. (2013) Developments in the identification of glycan biomarkers for the detection of cancer. *Mol Cell Proteomics* 12, 846-855
16. Zhang, Y., Jiao, J., Yang, P., and Lu, H. (2014) Mass spectrometry-based N-glycoproteomics for cancer biomarker discovery. *Clin Proteomics* 11, 18

17. Krishnamoorthy, L., and Mahal, L. K. (2009) Glycomic analysis: an array of technologies. *ACS Chem Biol* 4, 715-732
18. Zhang, H., Li, X. J., Martin, D. B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* 21, 660-666
19. Masuda, T., Sugiyama, N., Tomita, M., and Ishihama, Y. (2011) Microscale phosphoproteome analysis of 10,000 cells from human cancer cell lines. *Anal Chem* 83, 7698-7703
20. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367-1372
21. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13, 731-740
22. Hill, J. J., Tremblay, T. L., Fauteux, F., Li, J., Wang, E., Aguilar-Mahecha, A., Basik, M., and O'Connor-McCourt, M. (2015) Glycoproteomic comparison of clinical triple-negative and luminal breast tumors. *J Proteome Res* 14, 1376-1388
23. De Vet, E. C., Aguado, B., and Campbell, R. D. (2003) Adaptor signalling proteins Grb2 and Grb7 are recruited by human G6f, a novel member of the immunoglobulin superfamily encoded in the MHC. *Biochem J* 375, 207-213
24. Eirin, A., Zhu, X. Y., Puranik, A. S., Woollard, J. R., Tang, H., Dasari, S., Lerman, A., van Wijnen, A. J., and Lerman, L. O. (2016) Comparative proteomic analysis of extracellular vesicles isolated from porcine adipose tissue-derived mesenchymal stem/stromal cells. *Sci Rep* 6, 36120
25. Franchini, M., Frattini, F., Crestani, S., Bonfanti, C., and Lippi, G. (2013) von Willebrand factor and cancer: a renewed interest. *Thrombosis research* 131, 290-292
26. Terraube, V., Pendu, R., Baruch, D., Gebbink, M. F., Meyer, D., Lenting, P. J., and Denis, C. V. (2006) Increased metastatic potential of tumor cells in von Willebrand factor-deficient mice. *J Thromb Haemost* 4, 519-526
27. Terraube, V., Marx, I., and Denis, C. V. (2007) Role of von Willebrand factor in tumor metastasis. *Thrombosis research* 120 Suppl 2, S64-70

28. Kanekura, T., and Chen, X. (2010) CD147/basigin promotes progression of malignant melanoma and other cancers. *J Dermatol Sci* 57, 149-154
29. Sidhu, S. S., Mengistab, A. T., Tauscher, A. N., LaVail, J., and Basbaum, C. (2004) The microvesicle as a vehicle for EMMPRIN in tumor-stromal interactions. *Oncogene* 23, 956-963
30. Ferrara, N. (2009) Vascular endothelial growth factor. *Arterioscler Thromb Vasc Biol* 29, 789-791
31. Bulla, R., Tripodo, C., Rami, D., Ling, G. S., Agostinis, C., Guarnotta, C., Zorzet, S., Durigutto, P., Botto, M., and Tedesco, F. (2016) C1q acts in the tumour microenvironment as a cancer-promoting factor independently of complement activation. *Nature communications* 7, 10346
32. Metheny-Barlow, L. J., and Li, L. Y. (2003) The enigmatic role of angiopoietin-1 in tumor angiogenesis. *Cell Res* 13, 309-317
33. Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 41, D1063-1069

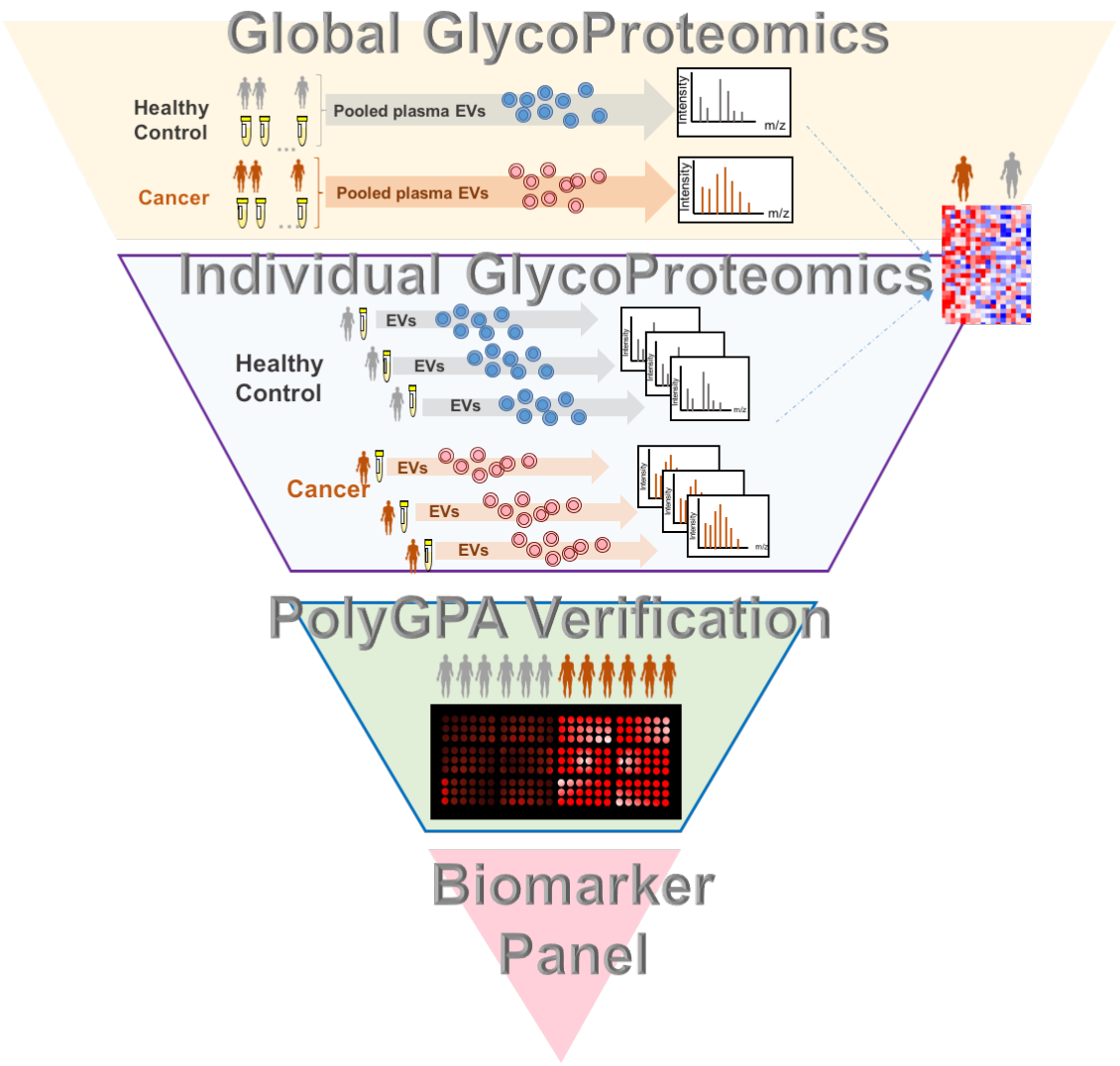

Figure 2.1. Workflow of a pipeline based on plasma EV glycoproteomics for biomarker discovery.

Microvesicles and exosomes were isolated through sequential high-speed centrifugation, followed by protein extraction, phase transfer surfactant digestion, and glycopeptide enrichment using hydrazide chemistry for LC-MS analyses. For global glycoproteomics analyses, 18 cancer and 6 control samples were pooled to create a preliminary list of increased glycosylated proteins. Proteomic analyses on 18 individual breast cancer and 10 healthy controls were performed to further verify the preliminary candidate biomarker list. Finally, Verification of potential biomarkers was performed using polyGPA.

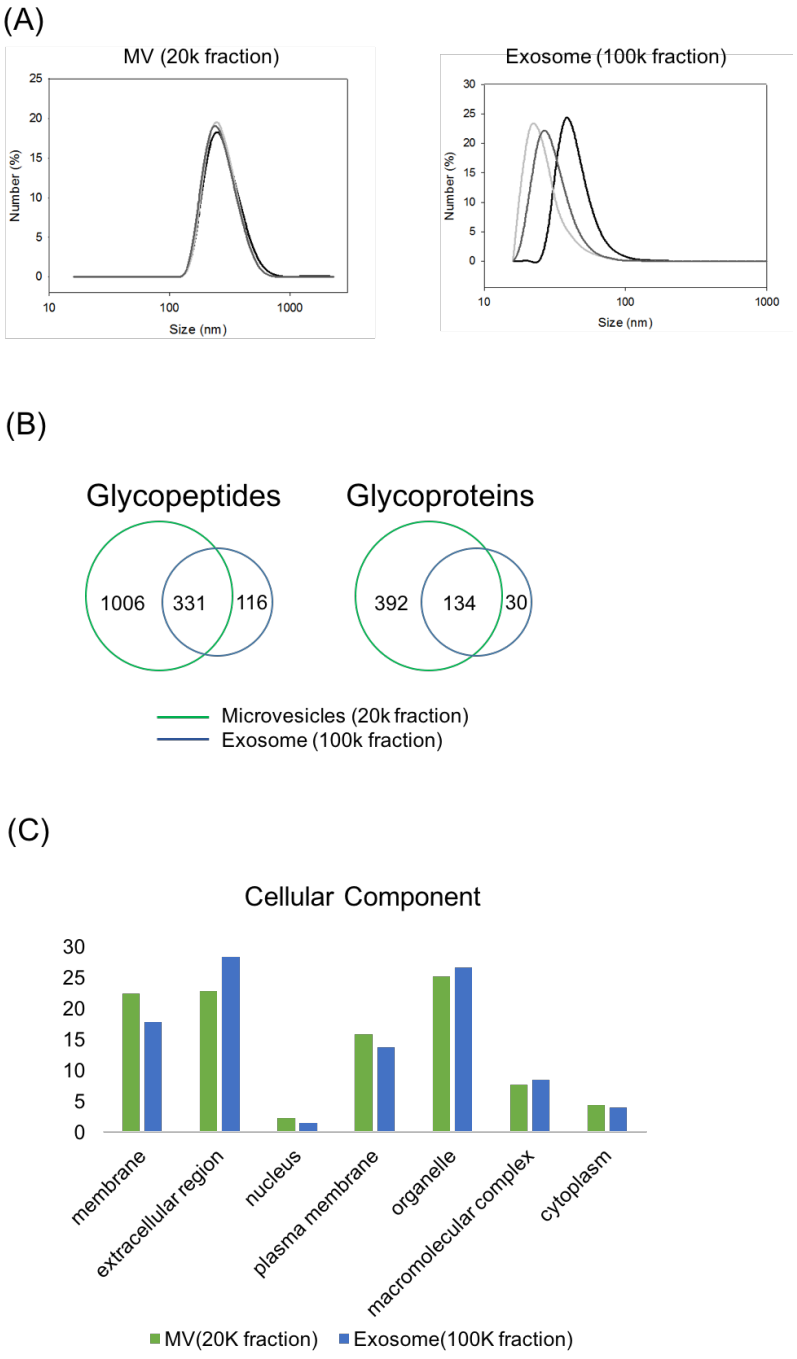

Figure 2.2 Characteristic analysis of glycoproteins in plasma-derived EVs.

(A) The size distribution of EVs isolated from two high-speed centrifugations measured by DLS. Each line corresponds to one acquired result from a single sample; (B) Venn diagram showing the glycopeptides and glycoproteins identification overlap between microvesicles and exosome. (C) Classification of the identified glycoproteins in EVs based on their cellular component.

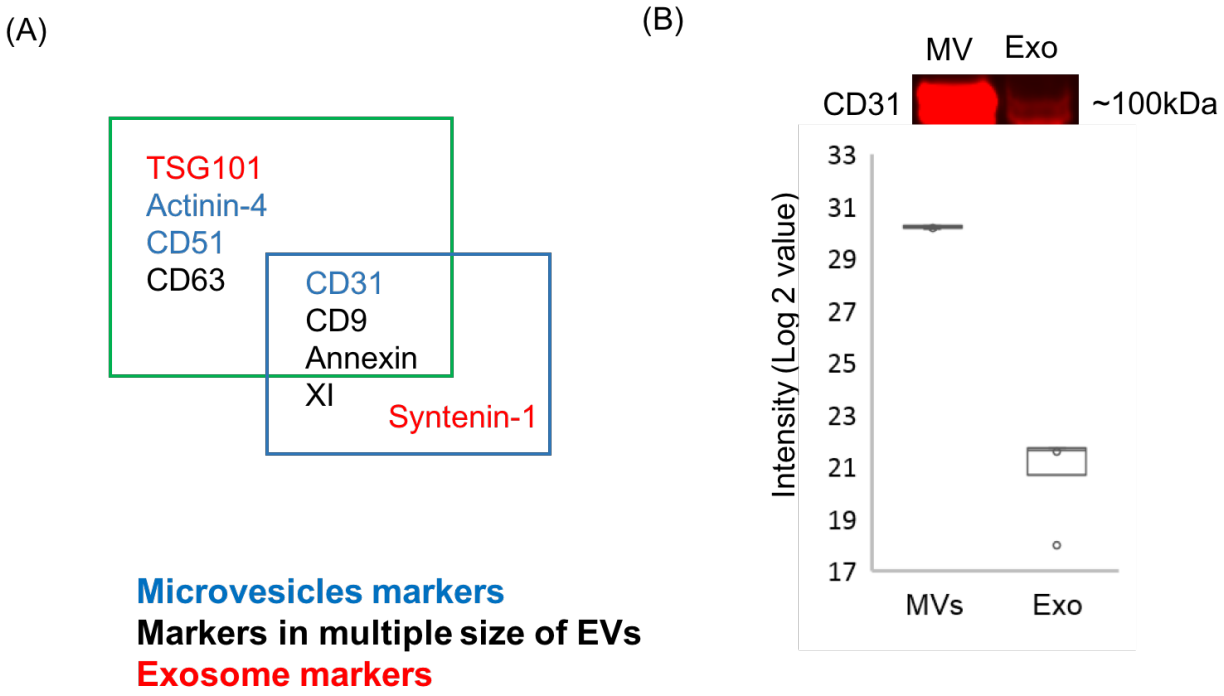

Figure 2.3 Purity of EVs isolation by high-speed centrifugation.

(A) The Venn diagram showing the common EVs markers present in MVs and exosome fractions through proteomic analyses. (B) Western blotting (WB) and MS data showing the purity of EV isolation. Two EV fractions were collected and analyzed by WB using antibody against CD 31, which is considered an endothelial-derived microvesicle marker. A total of 36 μg protein was used in MV fraction, and considering exosomes may possibly contain some plasma proteins, around 2.5-fold of protein amount of exosome fraction was used. MS data were extracted from two EV fractions, and the bar chart showed the intensity mean value with error bar of control and patient replicates.

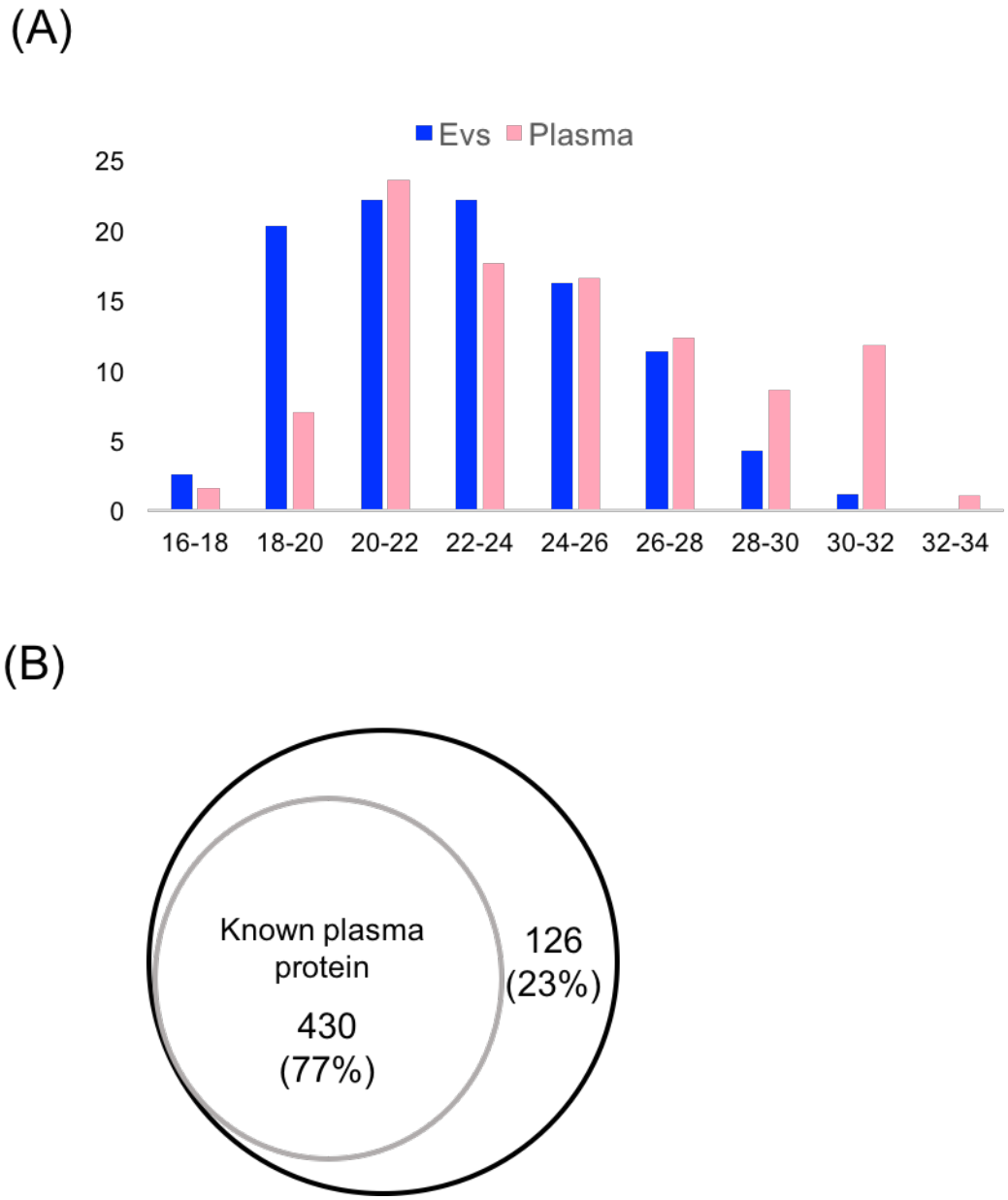

Figure 2.4 Comparison of glycoproteins in plasma and plasma-derived EVs.
(A) EVs and plasma proteins classification according to their intensities and spectral counts. (B) Venn diagram showing the overlap of the number of unique glycoproteins identified in EVs in this study compared to known plasma proteins.

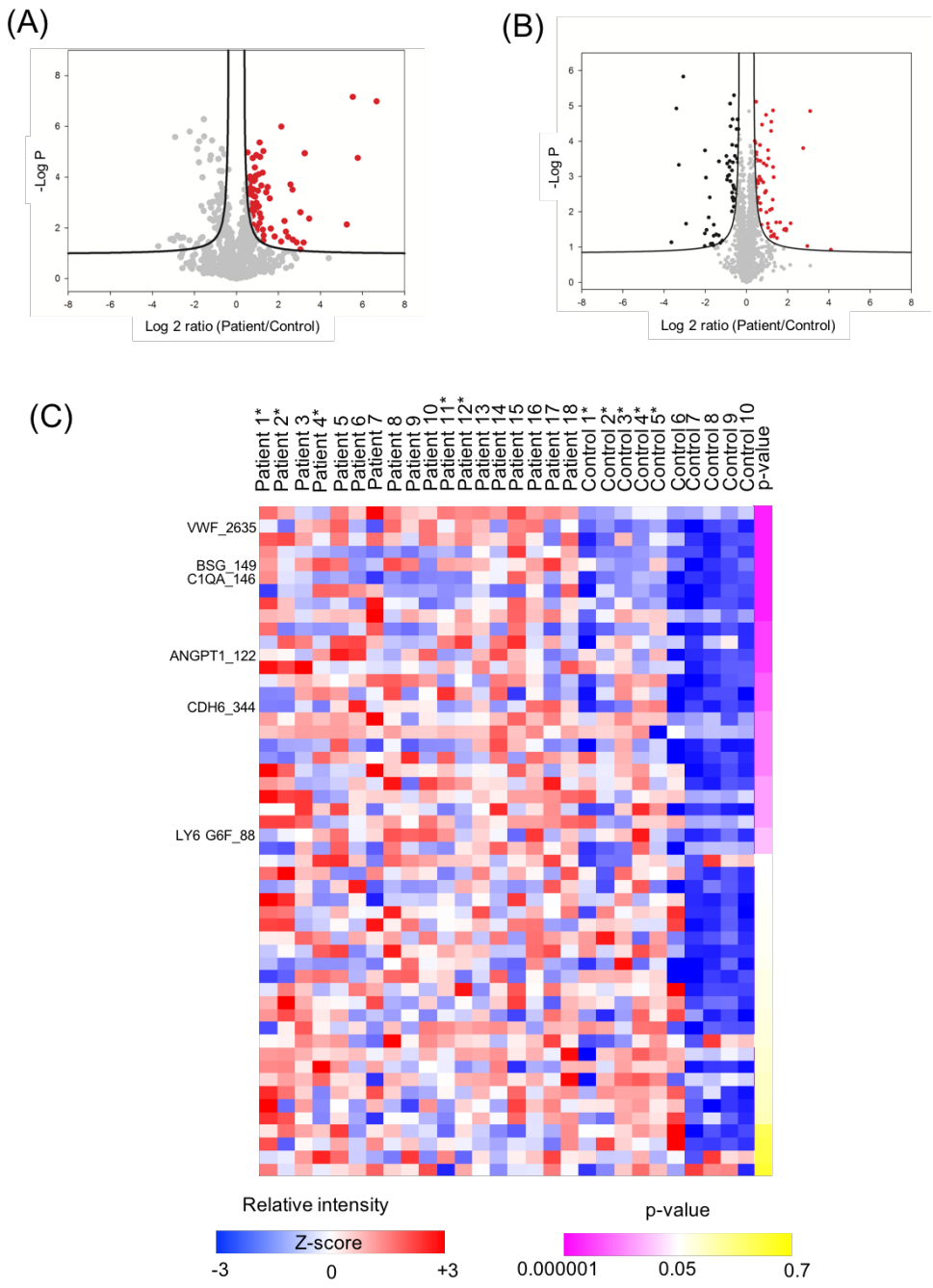

Figure 2.5Quantitative analysis of EV N-glycoproteomics between breast cancer and healthy controls.

(A) For global glycoproteomics, 18 breast cancer and 6 healthy controls were pooled to create a preliminary list of statistically increased glycosylated proteins. Volcano plot representing the quantitative analysis of the glycoproteomes of microvesicles in breast cancer patients v.s. healthy controls in left figure and proteomics analysis in (B). See supplementary figure for exosome

quantitation result. Significant changes in proteins and glycosites in breast cancer were identified through a permutation-based FDR test (FDR=0.05;S0=0.2) based on three technical replicates. The significant up-regulated proteins and glycosites are colored in red, and down-regulated are colored in gray on the left part of the volcano plot; (C) Quantitative glycoproteomics were performed on individuals to verify the preliminary list found in global glycoproteomics, and pvalue represents the significance of comparing individual patients and controls. In total, 18 patients and 10 healthy controls were examined in first verification experiment, 5 out of 18 patients and 6 out of 10 healthy controls were used in both global first individual verification glycoproteomics experiment (asterisk marked).

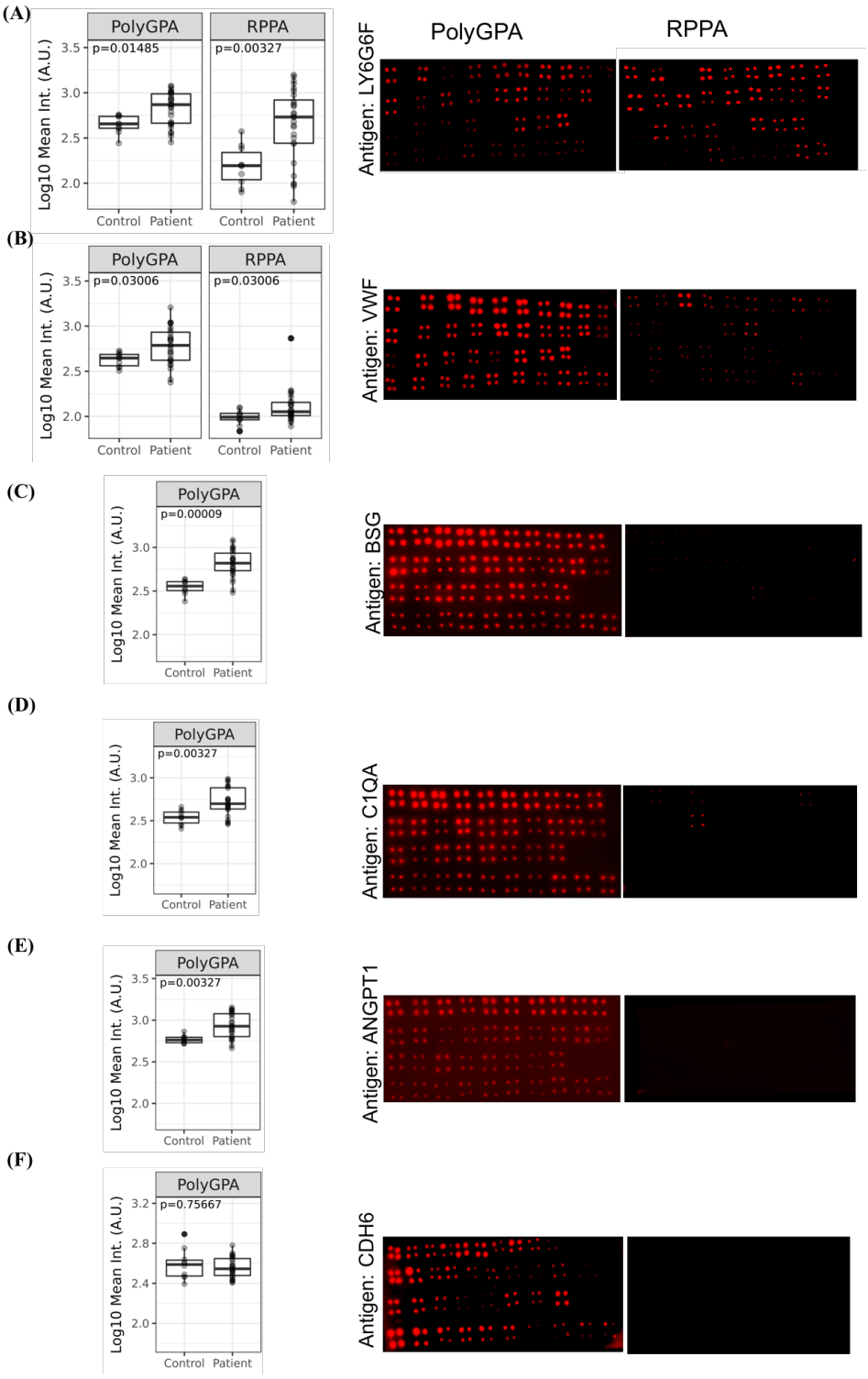

Figure 2.6Verification of selected targets in plasma EVs by PolyGPA.

Quantification of endogenous (a)LY6G6F, (b)VWF, (c)BSG, (d) C1QA, and (e) ANGPT1 (f) CDH6 in plasma EVs. For each membrane, top three rows were printed with 28 breast cancer

samples (first two rows with 10 samples and the third row with 8 samples) and the fourth row with 10 healthy control samples, each with 4 prints per individual sample. For quantitation of signals in polyGPA, the mean intensity of 4 prints per individual was used and the distribution of log10 (intensity) is depicted in the left pane.

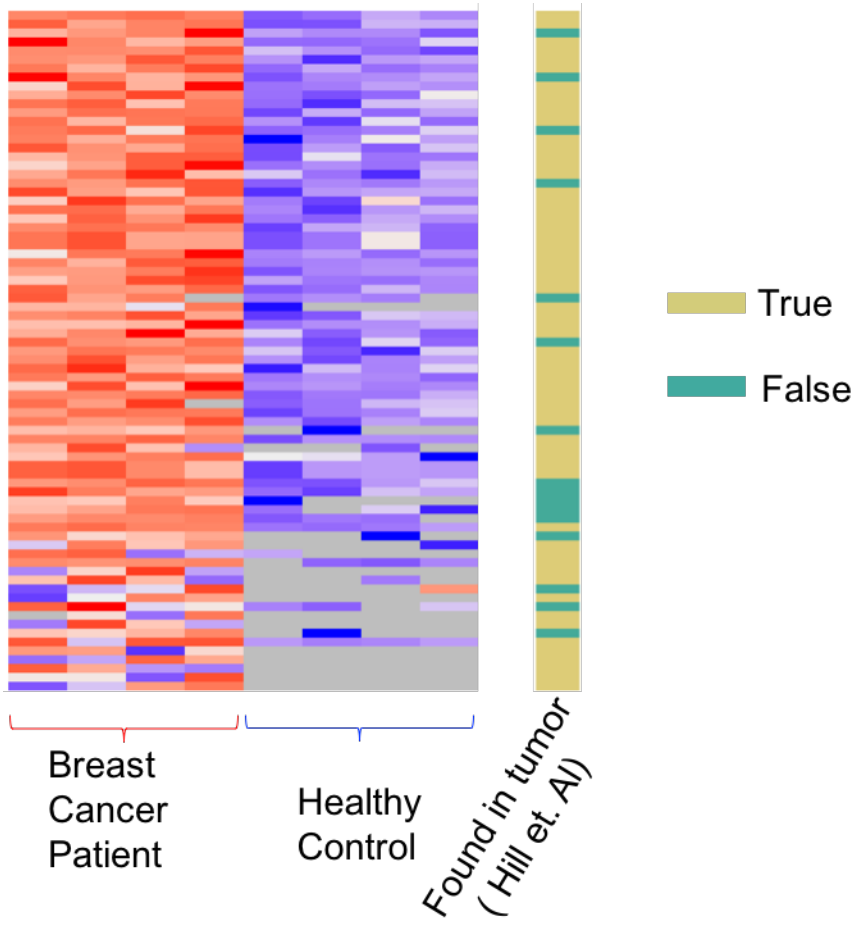

Figure 2.7 The hierarchical clustering analysis of up-regulated glycoproteins conveys the overlap between EVs in this study and breast cancer tissues by Hill et. al.

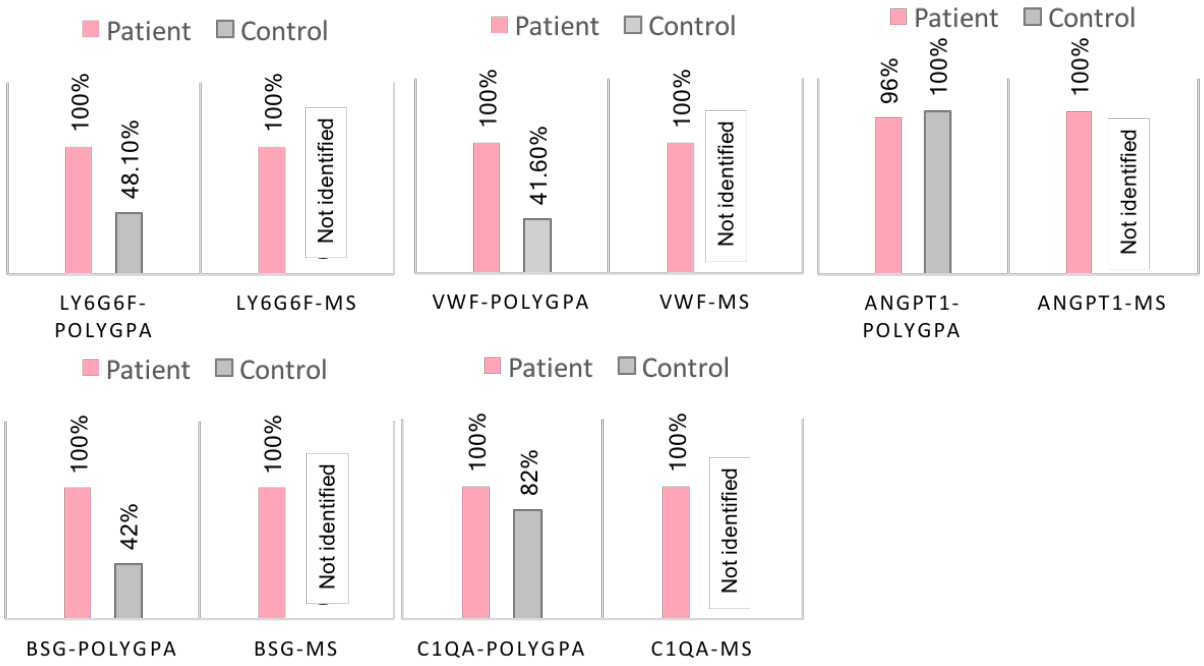

Figure 2.8 The quantitation results between polyGPA and label-free quantitation by MS of five glycoprotein candidates.

The EVs from the same patients and controls were collected for both individual N-glycoproteomics and polyGPA analysis.

CHAPTER 3 DISCOVERY OF PHOSPHORYLATION, GLYCOSYLATION, AND ACETYLATION PROTEINS IN EXTRACELLULAR VESICLES AS BIOMARKERS FOR BREAST CANCER SUBTYPES

3.1 Summary

Breast cancer is a complex disease that can be majorly classified into four molecular subtypes, Luminal A/B, Her 2 positive and Triple negative. With a wide variety of pathologic features and biological behaviors, the diagnosis or prognosis of specific subtypes is critical for applying the appropriate treatment. Here, we present a novel strategy for developing serial PTM-omics in plasma-derived EVs as biomarkers to discriminate different subtypes in breast cancer, which is able to identify 11824, 192, 1259 and 805 of unique pS/T, pY phosphorylation, N-glycosylation and acetylation peptides respectively in EVs, isolated from plasma samples. Using label-free quantitative PTMs-omics, several PTMs sites showed significantly higher in certain subtypes, and PCA further confirms that the expression profile of each PTMs are also different. In addition, several targets are verified in each subtype by using parallel monitor reaction approach. Together, this study demonstrates the great potential of this strategy for developing the biomarkers for different subtypes in breast cancer.

3.2 Introduction

Breast cancer is the most common neoplasm among women in United State. With the characteristic of highly heterogeneous, it includes a number of biologically distinct substances with specific pathologic features and biological behaviors(1, 2). There are different risk factors for different breast cancer subtypes, such as outcome, respond to therapies and histopathological features(3-5). Therefore, the diagnosis clarification of breast cancer among clinically relevant subtypes is required. Besides immunochemistry markers, some hallmarks have also been identified by Weinberg et al.(6) including “sustaining proliferative signaling”, “activating invasion and

metastasis", "resisting cell death", "evading immune destruction" and "reprogramming of energy metabolism".

Breast tumors are divided into four basic subgroups according to IHC markers (ER, PR, Her2), i.e., Luminal A: [ER+][PR+][Her2-], Luminal B: [ER+][PR+][Her2+], Her2 positive (Her2+) : [ER-][PR-][Her2+] and triple negative (TN): [ER-][PR-][Her2-]. In general, Her2+ and TN have relatively poorer prognosis than Luminal A and B cancers. For the Luminal A and B subtypes, ER is known to be the most important biomarker for breast cancer classification and plays crucial roles in breast carcinogenesis(7). Her 2+ subtypes with Her2 gene amplification and usually tend to grow faster, but often can be successfully treated with Her2 targeted therapies(8). Triple-negative breast cancer is hormone receptor and Her 2 negative, and this type of breast cancer is more common associated with BRCA1 mutation(9).

Notably, both hormonal and growth receptor promote the tumor progression through abnormal phosphorylation events. Studies indicated that the HER2 activation was consistent with EGFR and HER3 phosphorylation and downstream signaling activation (10). Moreover, Cuenca-Lopez et. al found that Androgen receptor is present in TN breast cancer and its expression correlates with activated receptor tyrosine kinases, such as EGFR, PDGFRβ and Erk1/2 (11). Taken together, phosphorylation is a key PTMs in finding the diagnostic biomarkers and therapeutic targets. Since many related genes are tyrosine receptor kinase, looking for tyrosine phosphorylation would be promising to provide us more valuable information.

Another PTMs acetylation has also been associated with promoting breast cancer metastasis, enhances the promoting role of AlB1 in breast cancer and been implied as another hormone therapies targets (12). Furthermore, researchers pointed out that enzymes and proteins involved in lysine acetylation are deregulated in cancer (13), and growing evidences of lysine acetylation links to cancer related metabolism and signaling pathways (14).

With the fact that most of proteins in plasma are glycosylated, glycoproteins are popular targets for disease biomarkers, and actually there are more than half of FDA-approved biomarkers for cancer diagnosis are glycoproteins. In breast cancer, IO et. al had reveal that the differences in expression of genes related the process of glycosylation exist between breast cancer subtypes. Another two studies also indicated the aberrant glycosylation in Her 2 positive breast cancer (15, 16). Furthermore, the aberrant glycosylation has been suggested as the biomarker for cancers (17).

As important features of PTMs proteins in breast cancer are described above, however, very few of them have been developed as biomarker for disease diagnosis or prognosis. Development of PTMs proteins as disease biomarkers from biofluids is challenging mainly due to the wide dynamic range of protein abundance in blood. With a few high abundant proteins representing over 95% of the mass in blood, few PTMs proteins in plasma/serum can be identified with stable and detectable concentration.

The discovery of extracellular vesicles in past decade revealed their important roles in cellular functions in tumorgenesis and metastasis, and it has presented them as intriguing sources for biomarker discovery and disease diagnosis. The growing evidences have suggested that tumor secreted EVs have potential on reflecting its cell origin and function, and can be identified before the physical detection of tumor, making them a promising candidate for early stage cancer. In this study, we collected the PTMs information including serine/threonine phosphorylation, tyrosine phosphorylation, acetylation and glycosylation from three major breast cancer subtypes (Luminal A and B, Her2 positive, Triple-negative) by using serial PTMs enrichment in extracellular vesicles approach. We compared the profile of each PTMs in each subtype and concluded a panel of potential PTMs markers for each subtype. In sum, the EV PTMs approach has demonstrated the feasibility of using different PTMs to distinguish different subtypes in breast cancer, and with the great potential of applying to other type of cancers.

3.3 Experimental procedure

3.3.1 Plasma sample

The Iowa University Institutional Review Board approved the use of human plasma samples. In the global PTM-ome experiment, blood samples were collected from 20 healthy females obtained through Susan G. Komen Tissue Bank and from 20 of each subtypes breast cancer patients that obtained through the University of Iowa biobank. Plasma samples were collected by standard protocol, in brief, plasma sample processing was initiated within 30 min of blood draw to an ethylenediaminetetraacetic acid (EDTA) containing tube. Samples were spun for 30 min at 3500 rpm to remove all cell debris and platelets.

3.3.2 Extracellular vesicles isolation

The EVs isolation and digestion were performed according to the reported protocol (18). A total of 5 ml pooled plasma samples were collected from both healthy individuals and patients diagnosed with breast cancer for the global PTMs experiment as technical replicates. Plasma samples were centrifuged at 20,000 xg at 4 oC for 1hr. Pellets were washed with cold PBS and centrifuged again at 20,000 xg at 4 oC for 1 hr, the pellets were microvesicles. Supernatant of the first centrifugation was further centrifuged at 100,000 xg at 4 oC for 1hr. Pellets were washed with cold PBS and centrifuged at 100,000 xg for 1hr again. The pellets from ultra-high speed centrifugations were exosome. Two separate isolated EVs were combined during sample lysis.

3.3.3 Protein digestion

The digestion was performed with phase transfer surfactant aided (PTS) digestion (19). Extracellular vesicles were solubilized in lysis buffer containing 12mM sodium deoxycholate (SDC), 12mM sodium lauroyl sarcosinate (SLS) and phosphatase inhibitor cocktail (Sigma-Aldrich, St. Louis, MO) in 100mM Tris-HCl, pH8.5. Proteins were reduced and alkylated with 10 mM tris-(2-carboxyethyl)phosphine (TCEP) and 40 mM chloroacetamide (CAA) at 95 °C for 5 min. Alkylated proteins were diluted to 5 fold by 50mM triethylammonium bicarbonate (TEAB) and digested with Lys-C (Wako, Japan) in a 1:100 (w/w) enzyme-to-protein ratio for 3 hr at 37 °C. Trypsin was added to a final 1:50 (w/w) enzyme-to-protein ratio for overnight digestion. The digested peptides were acidified with trifluoroacetic acid (TFA) to final concentration of 0.5% TFA, and 250ul of Ethyl acetate was added to 250ul digested solution. The mixture was shaken for 2 min, then centrifuged at 13,200 rpm for 2 min to obtain aqueous and organic phases. The aqueous phase was collected and desalted using a 100 mg of Sep-pak C18 column (Waters, Milford, MA).

3.3.4 Tyrosine phosphopeptides enrichment

Desalted peptides were resuspended in 50mM Tris-HCl pH 7.5, The samples were added with 20uL PT66 beads (Sigma-Aldrich, St. Louis, MO) and incubated with rotation overnight at 4^{o}C. The PT66 beads were washed sequentially with Lysis buffer (50mM Tris-HCl, 50mM NaCl,

1%NP40 pH7.5) and water, three times per solution for 10 mins rotation to wash off non-specific binding. Tyrosine phosphopeptides were sequential eluted twice by 0.1%TFA and once with 0.1%TFA/50%ACN . The eluent was dried under vacuum and then subjected to polymac enrichment.

3.3.5 Lysine acetylation peptides enrichment

Immunoaffinity enrichment of lysine acetylated peptides from EVs was performed using the PTMScan protocol as described previously with some modification. In brief, 20ul of lysine acetylation antibody conjugated beads were washed extensively with PBS. The Flow-through from tyrosine phosphopeptides were mixed with lysine acetylation antibody beads and incubated for 2hr at 4oC. The beads were washed twice with IAP buffer (50 mM MOPS, pH 7.2, 10 mM sodium phosphate, 50 mM NaCl) and three times with water. Peptides were eluted from beads with 0.15% TFA (sequential elutions of 55 µl followed by 50 µl, 10 min each elution at room temperature). Eluted peptides were desalted by SDB-XC stage tip and eluted with 40% acetonitrile in 0.1% TFA. Eluted peptides were dried under vacuum. The flow-through were desalted by SDB-XC stage tip and dried under vacuum.

3.3.6 Polymac phosphopeptides enrichment

Peptides were resuspended in 200 µL of loading buffer containing 1% trifluroacetic acid, and 80% acetonitrile and incubated with PolyMAC-Ti silica beads (Tymora Analytical, IN)(20) for 15 min. The beads were loaded into the tip with frit to remove the flow-through. The beads were washed twice with 200 µL washing buffer containinf 100uM Glycolic acid, 1% TFA, and 50% ACN and once with 80% ACN, using centrifuge at 100 rcf. The phosphopeptides were then eluted from the beads by twice with 50 µL of 400 mM ammonium hydroxide, 50%ACN, using centrifuge at 100 rcf. The eluates were collected and dried under vacuum. The flow-through were dried for glycopeptides enrichment

3.3.7 Glycopeptides enrichment

The glycopeptide enrichment was performed according to the reported protocol (21). Desalted peptides were oxidized with 10 mM sodium periodate in 50% ACN, 0.1% TFA at room temperature with shaking in the dark for 30 minutes. Excess sodium periodate was quenched by using 50 mM sodium sulfite for 15 minutes at room temperature with shaking in the dark. The samples were mixed with 50 µL/100 µL hydrazide magnetic beads for individual and pooled samples respectively, and incubated with vigorous shaking at room temperature overnight for the coupling reaction. Magnetic beads were washed sequentially with 400 µL/800 µL of 50% ACN, 0.1% TFA and 1.5 M NaCl for individual and pooled samples respectively, three times per solution for 1 minute per wash for the removal of non-coupled peptides. Beads were rinsed once with 100 µL/200 µL of 1x GlycoBuffer 2 (NEB) for individual and pooled samples respectively, and incubated with 3 µL/4 µL of PNGase F (NEB) in 100 µL/200 µL for individual and pooled samples respectively. N-glycans were cleaved by PNGase F. After desalting, the released former N-glycopeptides were analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).

3.3.8 LC-MS/MS

The PTMs peptides were dissolved in 4 µL of 0.3% formic acid (FA) with 3% ACN and injected into an Easy-nLC 1200 (Thermo Fisher Scientific). Peptides were separated on a 45 cm in-house packed column (360 µm OD × 75 µm ID) containing C18 resin (2.2 µm, 100Å, Michrom Bioresources) with a 30 cm column heater (Analytical Sales and Services) set to 50 °C. The mobile phase buffer consisted of 0.1% FA in ultra-pure water (buffer A) with an eluting buffer of 0.1% FA in 80% ACN (buffer B) run over either with a 45 min or 60 min linear gradient of 5%-25% buffer B at flow rate of 300 nL/min. The Easy-nLC 1200 was coupled online with a Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer. The mass spectrometer was operated in the data-dependent mode in where the 10 most intense ions were subjected to High-energy collisional dissociation (HCD) fragmentation (normalized collision energy (NCE) 30%, AGC 3e4, max injection time 100 ms) for each full MS scan (from m/z 350-1500 with a resolution of 120,000 at m/z 200).

3.3.9 Data Processing

The raw files were searched directly UniprotKB database version Aug2017 with no redundant entries using MaxQuant software (version 1.5.6.1) (22) with the Andromeda search engine. Initial precursor mass tolerance was set to 20 p.p.m. and the final tolerance was set to 6 p.p.m., and ITMS MS/MS tolerance was set at 0.6 Da. Search criteria included a static carbamidomethylation of cysteines (+57.0214 Da) and variable modifications of (1) oxidation (+15.9949 Da) on methionine residues, (2) acetylation (+42.011 Da) at N-terminus of protein, and (3) phosphorylation(+79.996 Da) on serine, threonine or tyrosine residues for phosphorylation, acetylation (+42.011 Da) on Lysine residue for acetylation and deamidation (+0.984Da) on asparagine residues for glycosylation were searched. Search was performed with Trypsin/P digestion and allowed a maximum of two missed cleavages on the peptides analyzed from the sequence database. The false discovery rates of proteins, peptides and PTMs sites were set at 0.01. The minimum peptide length was six amino acids, and a minimum Andromeda score was set at 40 for modified peptides. The glycosylation sites were selected based on the matching to the N-X-S/T (X not Pro) motif. A site localization probability of 0.75 was used as the cut-off for localization of glycosylation sites. All the peptide spectral matches and MS/MS spectra can be viewed through MaxQuant viewer.

3.3.10 Quantitative Data Analysis

All data was analyzed using the Perseus software (version 1.5.4.1) (23). For quantification of both proteomic and PTM-omic datasets, the intensities of proteins and PTMs sites were derived from MaxQuant, and the missing values of intensities were replaced by normal distribution with a downshift of 1.8 standard deviations and a width of 0.3 standard deviations. The significantly increased PTMs sites or proteins in patient samples were identified by a ANOVA multi-test with a permutation-based FDR cut-off 0.05 for all of data sets. For heatmap, the changed sites or proteins were used, the imputed data set was normalized by z-score within each dataset.

3.4 Result

3.4.1 Identification of 11824, 192, 1259 and 805 unique pS/T, pY phosphorylation, N-glycosylation and acetylation peptides from plasma-derived extracellular vesicles

The workflow of integrating proteomics analysis of PTMs by serial enrichment is illustrated in Figure 3.1. EVs were isolated from human plasma through two steps ultra-high-speed centrifugation. For the initial screen, the plasma samples were collected and pooled from healthy individuals (n=20), patients diagnosed with Luminal A or B (n=20), Her 2+ (n=20) and Triple negative (n=20). After digest of EVs proteins, the desalted peptides were firstly used for tyrosine phosphorylated peptides enrichment by using PT66 antibody. The flow-through of first enrichment directly used for the second step which is acetylation enrichment. Then the flow-through of second enrichment was used for serine, threonine phosphorylated peptides enrichment by Polymac. Finally, the flow-through from Polymac was used for glycopeptides entichment. All of the PTMs peptides were analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) on a high-speed and high-resolution mass spectrometer with technical replicates. Label-free quantitation was performed to determine the differential PTMs proteins in the plasma of control and three subtypes of breast cancer patient samples. The Strategy allowed us to identify 12016 phosphopeptides including 192 tyrosine phosphorylation, 805 acetylpeptides and 1259 glycopeptides, representing 1699, 453 and 495 proteins. Gene ontology analysis shows that different PTMs proteins are distinctively from certain cellular location (Figure 3.2). All of PTMs proteins are significantly enriched from membrane and organelle, phosphoproteins and glycoproteins are distinctively from plasma membrane, acetylproteins and glycoproteins are distinctively from extracellular region, phosphoproteins and acetylproteins are distinctively from cytoplasm. As shown in Figure 3.3A, few protein overlap between each PTMs, and around 40% of each PTMs proteins are uniquely identified in their own run, indicating the enrichment of PTM-ome rescued the low-abundance proteins that usually escaping from shut-gun proteomics (Figure 3.3B).

3.4.2 Cancer specific PTMs peptides in EVs for different subtypes

Label-free quantitation of all PTMs peptides was performed to identify the list of PTMs proteins changing in three subtypes of breast cancer. We quantified 6281, 62, 393 and 1127 unique class1 phosphorylation, tyrosine phosphorylation, acetylation and glycosylation sites respectively. By using ANOVA multi-test, we found that 94 phosphorylation sites have been identified significantly increased in all subtypes, 20 phosphosites specifically increased in luminal A/B, 67 phosphosites increased in only TN and 31 phosphosites increased in Her2+ (Figure 3.4A). In addition, as shown in Figure 3.5A, we identified 13 tyrosine phosphorylation sites with significantly increase in all subtypes versus healthy control, 2 sites in luminal A or B and 3 sites in TN. For acetylation, 25 acetylsites have been identified with significantly increase in all subtypes, 24 sites increased in luminal A/B and 2 sites in TN (Figure 3.6A). For glycosylation, we identified 72 glycosites with a significant increase of abundance in all subtypes, 66, 36 and 14 glycosites increased in Luminal A/B, TN and Her2+, respectively (Figure 3.7A). The difference in all of PTMs sites may be a result of changes in protein expression or changes of glycosylation on specific sites. To distinguish these factors, we also performed label-free quantitation of total EV proteomes. We quantified 2190 proteins, 93% of changed proteins are non-subtype specific. (Figure 3.8) Among all of changed proteins, less than 40% of the changed PTMs proteins were also changed in total protein level, indicating the rest of 60% changed PTMs proteins were either changed due to the modification rather than protein level or cannot be identified due to the low abundancy (Figure 3.3B).

To evaluate whether PTMs-ome can distinct different subtypes from healthy control, we applied principle component analysis (PCA) to show that phosphorylation, acetylation not only distinguish breast cancer from control, but also well-separate different subtypes (Figure 3.4, 3.5, 3.6B). Besides glycosylation can separate breast cancer from healthy control, also indicates that aggressive breast cancers and Luminal A/B have distinct glycoproteomics profiles (Figure 3.7B). To better understand the biological roles of differential PTMs events, we examined all PTMs proteins specific to patients with cancer, using STRING to identify enriched gene ontology categories and signaling networks. It is interested to reveal that ErbB signaling is significantly enriched in phosphorylation (Fig. 3.9), which is consisted with previous studies linking ErbB family signaling to cancer progression in breast cancer. We also found that several metabolic pathways in acetylation (Fig 3.10) and two major networks including cell adhesion and phagosome

are in glycosylation for all subtypes (Fig 3.11), persisting previous studies that acetylation links to controlling cancer cell metabolism (14) and glycosylation plays roles in EVs uptake and cell adhesion (24).

3.5 Discussion

The traditional classification of breast cancer using ER, PR and Her 2 has been frequently challenged by samples with exceptional clinical associations, thus determining the characteristics of four major subtypes has gaining attention. A lot of potential biomarkers such as AR, KI67, CK, BCL2 and TP53.etc are popular studied. However, very rare of them are used as liquid biopsy markers. The main issue is that the abundance of tumor leakage proteins is too low to be detected in blood, so tumor biopsy is still needed for diagnosis of breast cancer and its subtypes.

In this study, we proposed the feasibility of using different PTMs proteins in extracellular vesicles as different breast cancer subtypes markers. In the first screening result, we clearly see the different PTMs-omics patterns between each subtype while the patterns look similar in their global proteomics, indicating that the PTMs can better present the molecular difference between subtypes. By looking at different PTMs, we are able to inspect the potential roles of PTMs proteins in EVs-derived cancer metastasis. For example, phosphorylation in molecular signaling, glycosylation in EVs uptake and acetylation in metabolize. It is interesting that Principle component analysis shows that glycosylation can well separates aggressive breast cancer from non-aggressive breast cancer and control, implying the level of metastasis may be altered by the protein glycosylation level in EVs. We also found that despite the fact that ErbB signaling commonly increased in all subtypes, nuclear factor of activated T-cells, cytoplasmic 2 (NFATC2) involving in Wnt signaling are significantly higher only in TN, which consists with previous report that Wnt signaling in TN is associated with cancer metastasis (25).

3.6 Data Access

The raw data, MS identification lists, and quantitation tables for all proteomic analyses have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository.

3.7 Reference

1. Spitale, A., Mazzola, P., Soldini, D., Mazzucchelli, L., and Bordoni, A. (2009) Breast cancer classification according to immunohistochemical markers: clinicopathologic features and short-term survival analysis in a population-based study from the South of Switzerland. *Ann Oncol* 20, 628-635
2. Tang, P., Wang, J., and Bourne, P. (2008) Molecular classifications of breast carcinoma with similar terminology and different definitions: are they the same? *Hum Pathol* 39, 506-513
3. Desmedt, C., Sotiriou, C., and Piccart-Gebhart, M. J. (2009) Development and validation of gene expression profile signatures in early-stage breast cancer. *Cancer Invest* 27, 1-10
4. Iwamoto, T., and Pusztai, L. (2010) Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? *Genome Med* 2, 81
5. Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010) The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J Pathol* 220, 263-280
6. Hanahan, D., and Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* 144, 646-674
7. Rakha, E. A., Reis-Filho, J. S., and Ellis, I. O. (2010) Combinatorial biomarker expression in breast cancer. *Breast Cancer Res Treat* 120, 293-308
8. Wolff, A. C., Hammond, M. E., Schwartz, J. N., Hagerty, K. L., Allred, D. C., Cote, R. J., Dowsett, M., Fitzgibbons, P. L., Hanna, W. M., Langer, A., McShane, L. M., Paik, S., Pegram, M. D., Perez, E. A., Press, M. F., Rhodes, A., Sturgeon, C., Taube, S. E., Tubbs, R., Vance, G. H., van de Vijver, M., Wheeler, T. M., Hayes, D. F., and American Society of Clinical Oncology/College of American, P. (2007) American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med* 131, 18-43
9. Peshkin, B. N., Alabek, M. L., and Isaacs, C. (2010) BRCA1/2 mutations and triple negative breast cancers. *Breast Dis* 32, 25-33

10. Wulfkuhle, J. D., Berg, D., Wolff, C., Langer, R., Tran, K., Illi, J., Espina, V., Pierobon, M., Deng, J., DeMichele, A., Walch, A., Bronger, H., Becker, I., Waldhor, C., Hofler, H., Esserman, L., Investigators, I. S. T., Liotta, L. A., Becker, K. F., and Petricoin, E. F., 3rd (2012) Molecular analysis of HER2 signaling in human breast cancer by functional protein pathway activation mapping. *Clin Cancer Res* 18, 6426-6435
11. Cuenca-Lopez, M. D., Montero, J. C., Morales, J. C., Prat, A., Pandiella, A., and Ocana, A. (2014) Phospho-kinase profile of triple negative breast cancer and androgen receptor signaling. *BMC Cancer* 14, 302
12. You, D., Zhao, H., Wang, Y., Jiao, Y., Lu, M., and Yan, S. (2016) Acetylation Enhances the Promoting Role of AIB1 in Breast Cancer Cell Proliferation. *Mol Cells* 39, 663-668
13. Gil, J., Ramirez-Torres, A., and Encarnacion-Guevara, S. (2017) Lysine acetylation and cancer: A proteomics perspective. *J Proteomics* 150, 297-309
14. Lin, R., Zhou, X., Huang, W., Zhao, D., Lv, L., Xiong, Y., Guan, K. L., and Lei, Q. Y. (2014) Acetylation control of cancer cell metabolism. *Curr Pharm Des* 20, 2627-2633
15. Ashkani, J., and Naidoo, K. J. (2016) Glycosyltransferase Gene Expression Profiles Classify Cancer Types and Propose Prognostic Subtypes. *Sci Rep* 6, 26451
16. Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., and Geiger, T. (2016) Proteomic maps of breast cancer subtypes. *Nat Commun* 7, 10259
17. Diaz-Fernandez, A., Miranda-Castro, R., de-Los-Santos-Alvarez, N., and Lobo-Castanon, M. J. (2018) Post-translational modifications in tumor biomarkers: the next challenge for aptamers? *Anal Bioanal Chem*
18. Chen, I. H., Xue, L., Hsu, C. C., Paez, J. S., Pan, L., Andaluz, H., Wendt, M. K., Iliuk, A. B., Zhu, J. K., and Tao, W. A. (2017) Phosphoproteins in extracellular vesicles as candidate markers for breast cancer. *Proc Natl Acad Sci U S A* 114, 3175-3180
19. Masuda, T., Sugiyama, N., Tomita, M., and Ishihama, Y. (2011) Microscale phosphoproteome analysis of 10,000 cells from human cancer cell lines. *Anal Chem* 83, 7698-7703
20. Iliuk, A. B., Martin, V. A., Alicie, B. M., Geahlen, R. L., and Tao, W. A. (2010) In-depth analyses of kinase-dependent tyrosine phosphoproteomes based on metal ion-functionalized soluble nanopolymers. *Mol Cell Proteomics* 9, 2162-2172

21. Pan, L., Aguilar, H. A., Wang, L., Iliuk, A., and Tao, W. A. (2016) Three-Dimensionally Functionalized Reverse Phase Glycoprotein Array for Cancer Biomarker Discovery and Validation. *J Am Chem Soc* 138, 15311-15314
22. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367-1372
23. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13, 731-740
24. Yanez-Mo, M., Siljander, P. R., Andreu, Z., Zavec, A. B., Borras, F. E., Buzas, E. I., Buzas, K., Casal, E., Cappello, F., Carvalho, J., Colas, E., Cordeiro-da Silva, A., Fais, S., Falcon-Perez, J. M., Ghobrial, I. M., Giebel, B., Gimona, M., Graner, M., Gursel, I., Gursel, M., Heegaard, N. H., Hendrix, A., Kierulf, P., Kokubun, K., Kosanovic, M., Kralj-Iglic, V., Kramer-Albers, E. M., Laitinen, S., Lasser, C., Lener, T., Ligeti, E., Line, A., Lipps, G., Llorente, A., Lotvall, J., Mancek-Keber, M., Marcilla, A., Mittelbrunn, M., Nazarenko, I., Nolte-'t Hoen, E. N., Nyman, T. A., O'Driscoll, L., Olivan, M., Oliveira, C., Pallinger, E., Del Portillo, H. A., Reventos, J., Rigau, M., Rohde, E., Sammar, M., Sanchez-Madrid, F., Santarem, N., Schallmoser, K., Ostenfeld, M. S., Stoorvogel, W., Stukelj, R., Van der Grein, S. G., Vasconcelos, M. H., Wauben, M. H., and De Wever, O. (2015) Biological properties of extracellular vesicles and their physiological functions. *J Extracell Vesicles* 4, 27066
25. Dey, N., Barwick, B. G., Moreno, C. S., Ordanic-Kodani, M., Chen, Z., Oprea-Ilies, G., Tang, W., Catzavelos, C., Kerstann, K. F., Sledge, G. W., Jr., Abramovitz, M., Bouzyk, M., De, P., and Leyland-Jones, B. R. (2013) Wnt signaling in triple negative breast cancer is associated with metastasis. *BMC Cancer* 13, 537

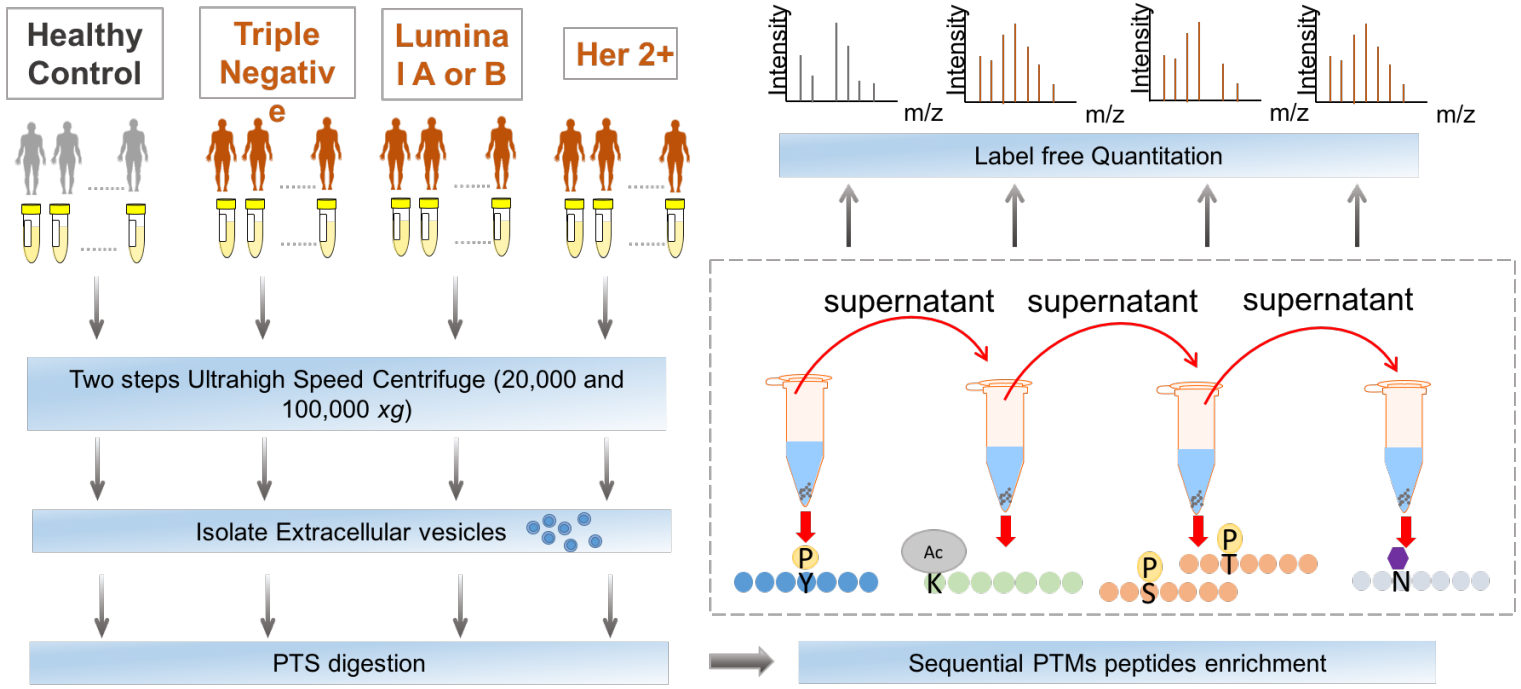

Figure 3.1 The workflow of serial PTMs-omics in extracellular vesicles for biomarker discovery.

The EVs were isolated from plasma by two step ultra-high speed centrifugation. The serial PTMs enrichments were performed by the order of 1) Tyrosine phosphorylation 2) Lysine acetylation 3) Serine/Theronine phosphorylation 4) N-Glycosylation after PTS digestion. Finally, the label-free quantitation was performed.

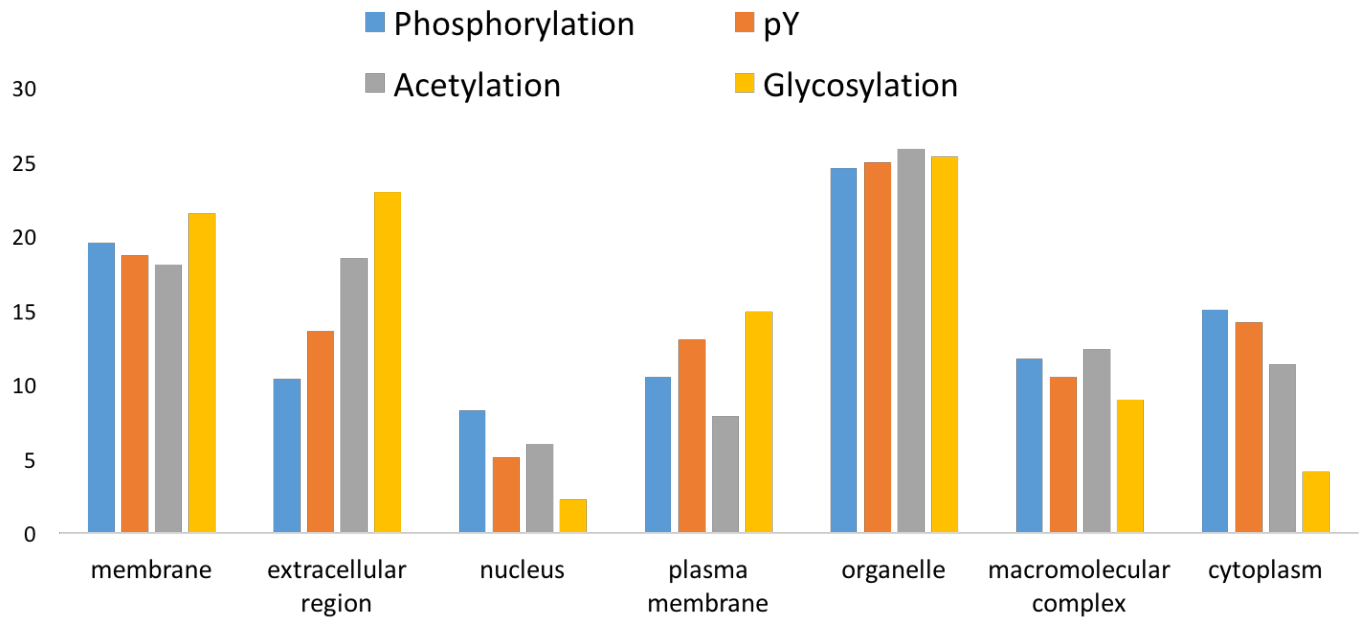

Figure 3.2 The cellular component analysis of identified PTMs proteins in plasma EVs.

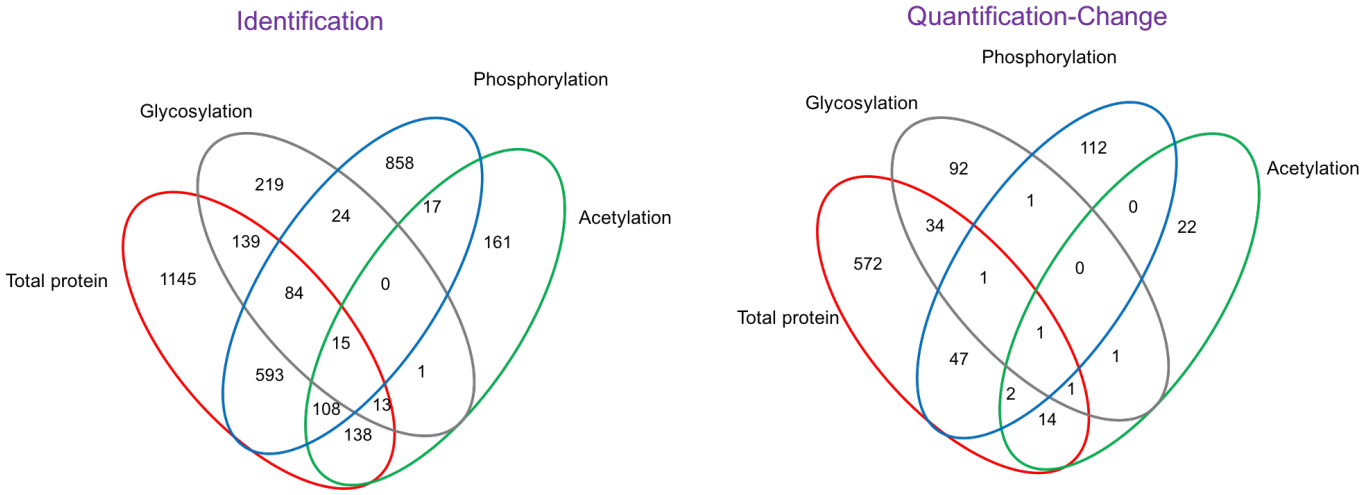

Figure 3.3 The comparison of identification and quantitation result between three modifications and total proteome.

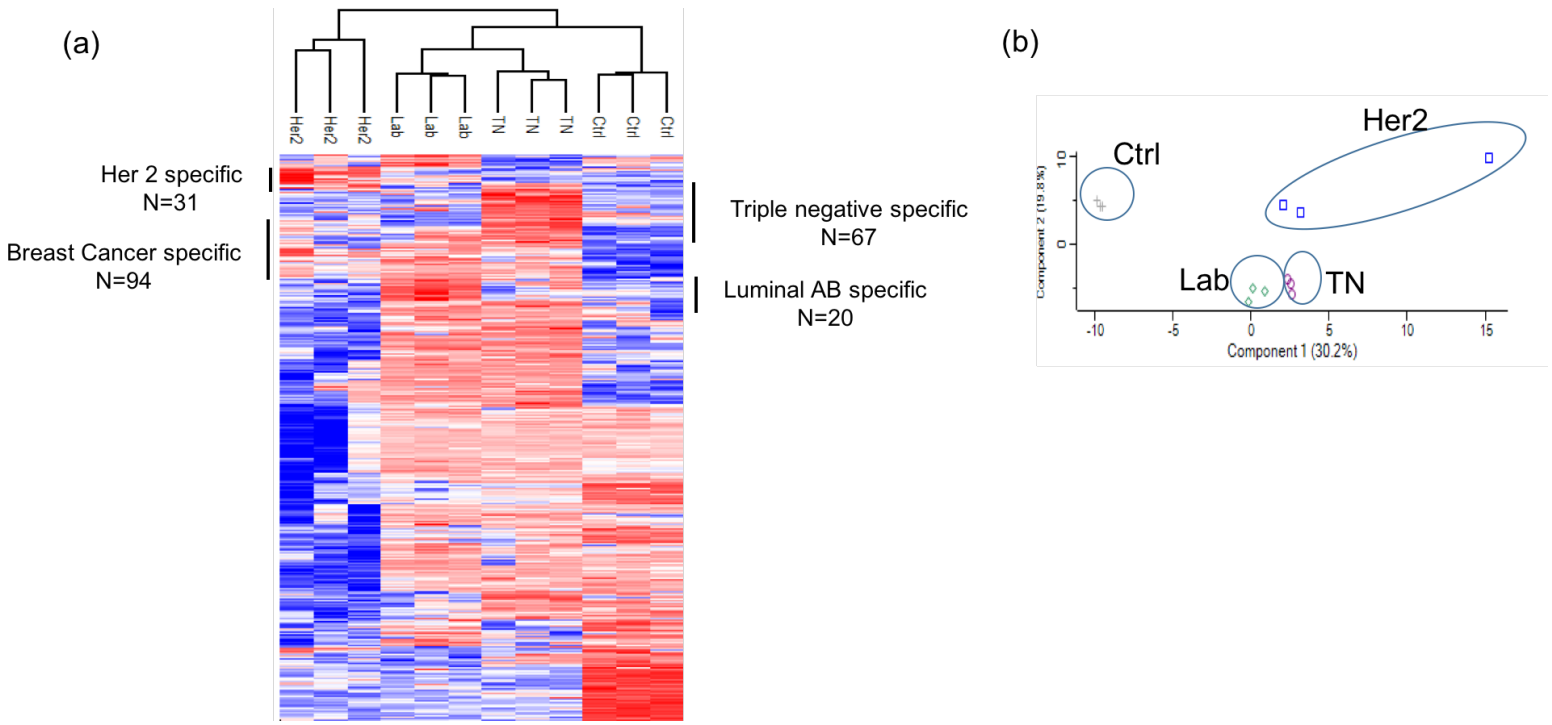

Figure 3.4 The quantitative phosphoproteomics analysis for three breast cancer subtypes.

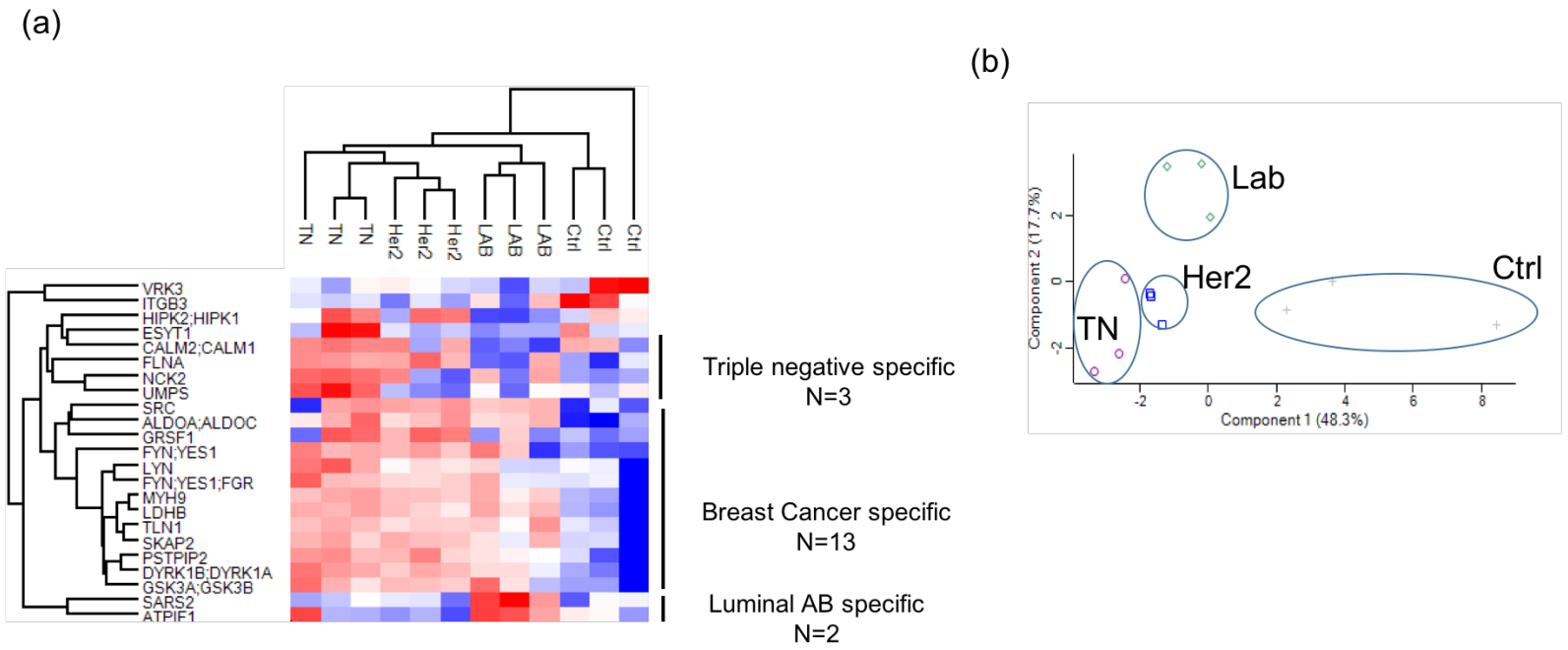

Figure 3.5 The quantitative tyrosine phosphoproteomics analysis for three breast cancer subtypes.

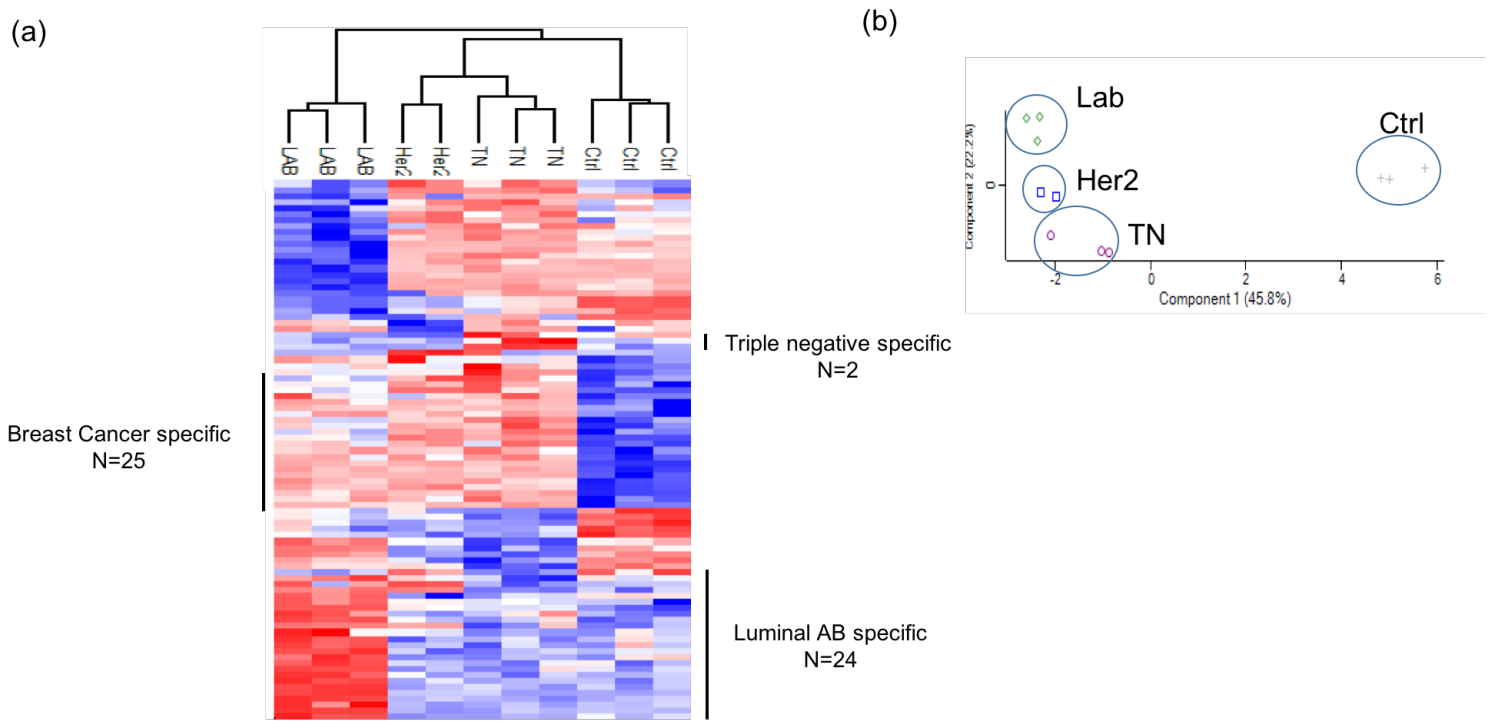

Figure 3.6 The quantitative acetylproteomics analysis for three breast cancer subtypes.

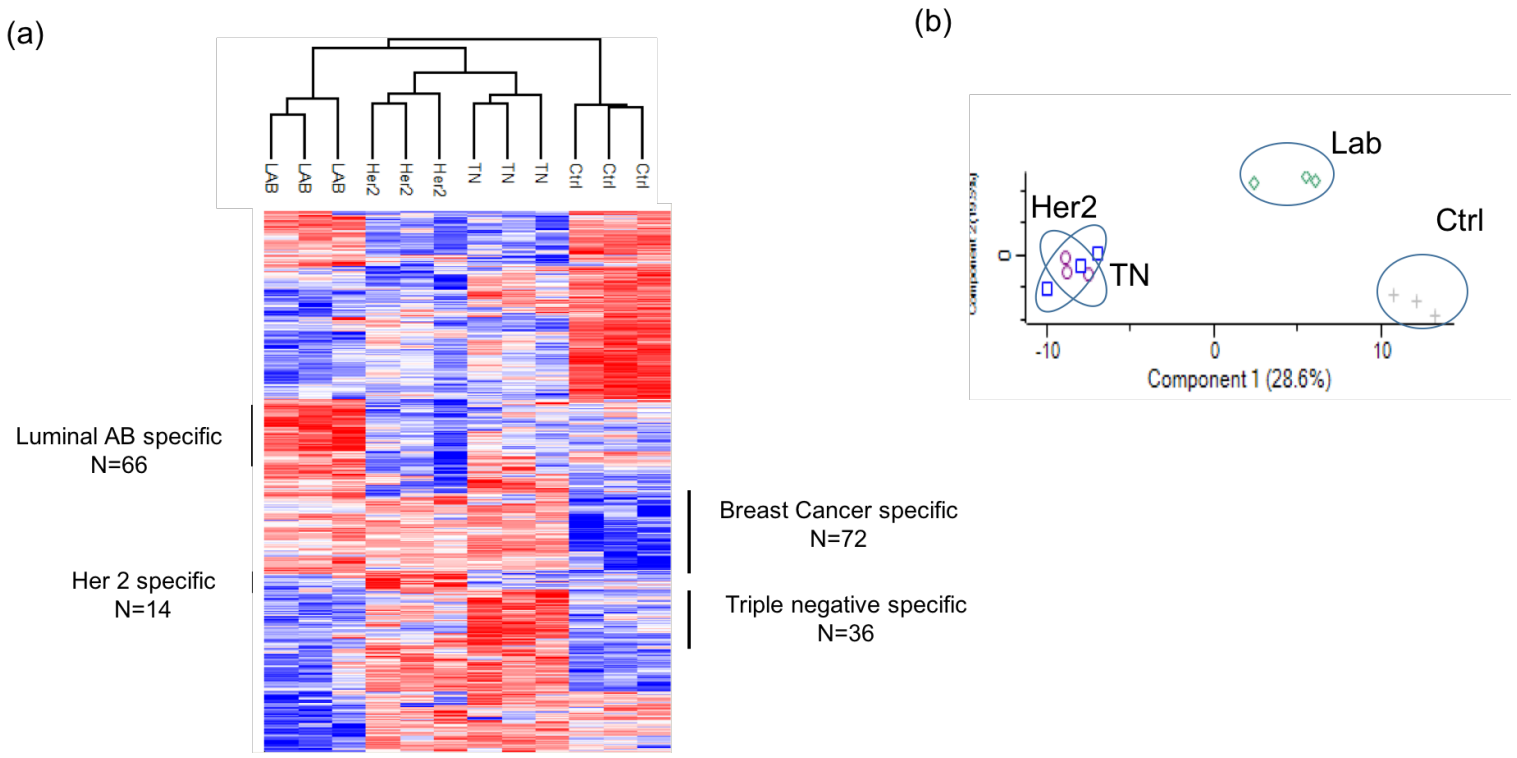

Figure 3.7 The quantitative glycoproteomics analysis for three breast cancer subtypes.

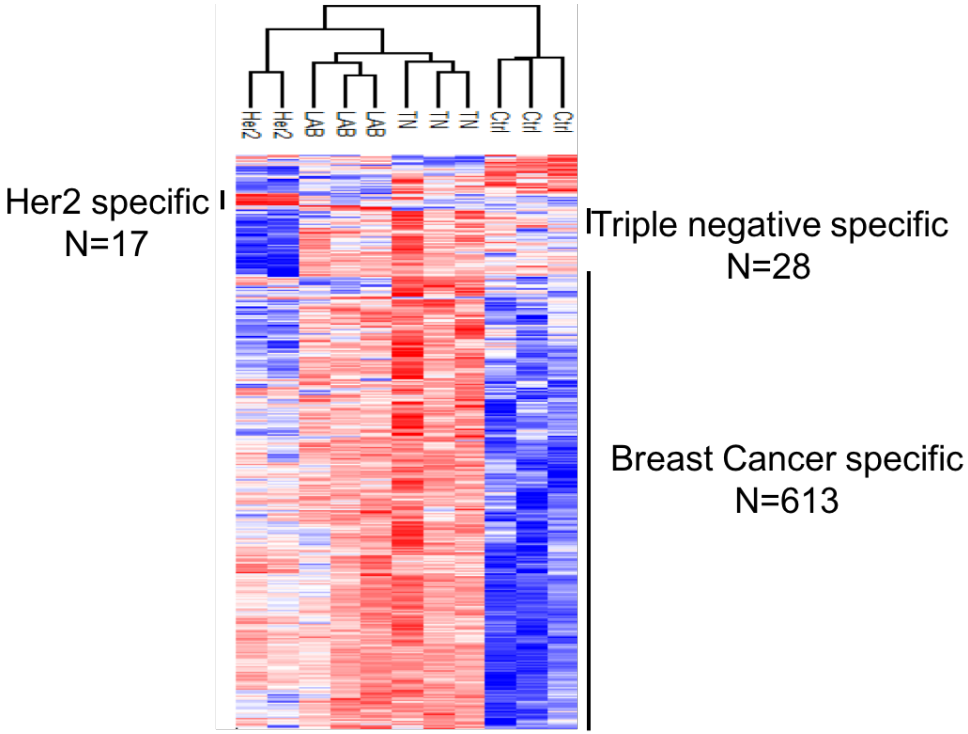

Figure 3.8 The quantitative proteomics analysis for three breast cancer subtypes.

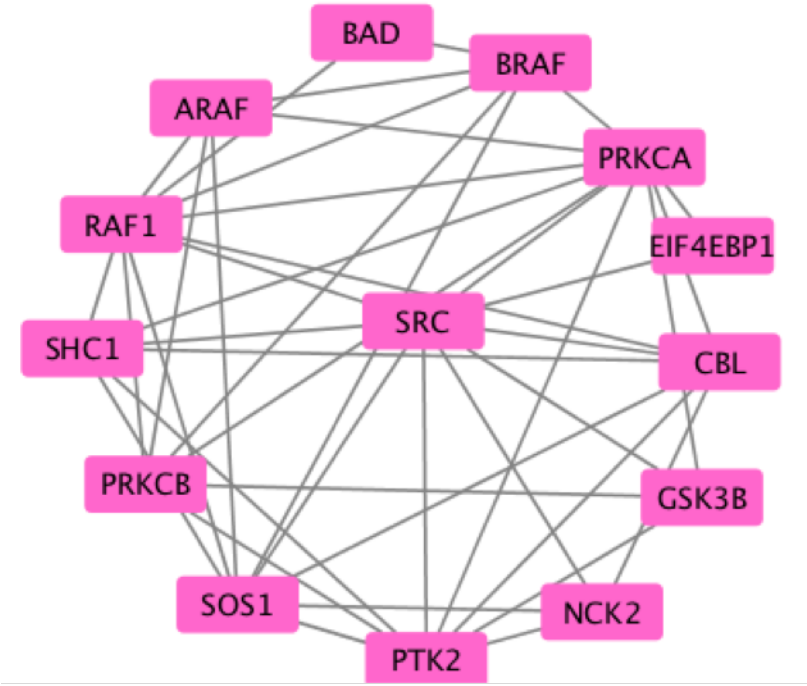

Figure 3.9 The common enriched networking in phosphoproteomics for Luminal A/ B , Her 2 positive and triple negative.

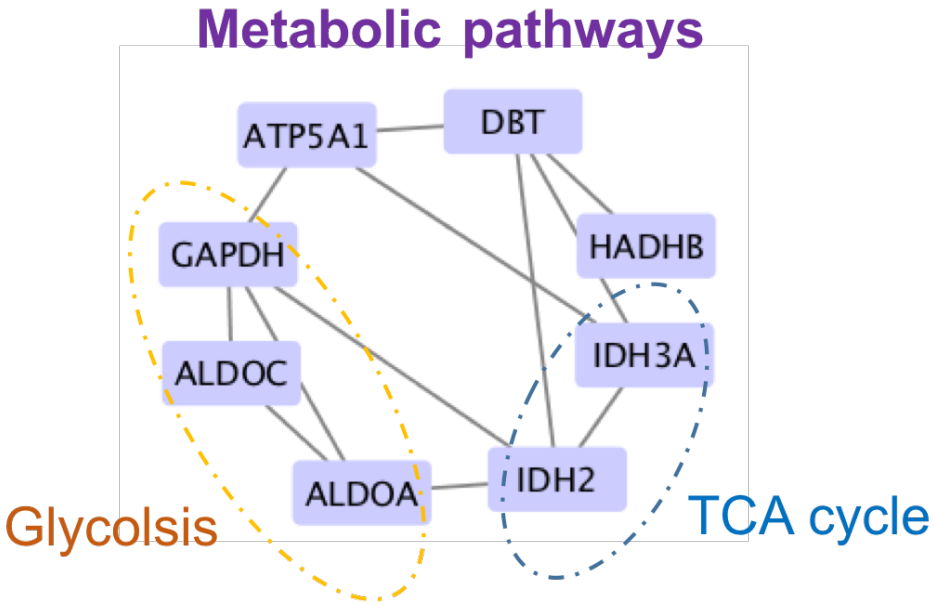

Figure 3.10 The common enriched networking in acetylproteomics for Luminal A/ B , Her 2 positive and triple negative.

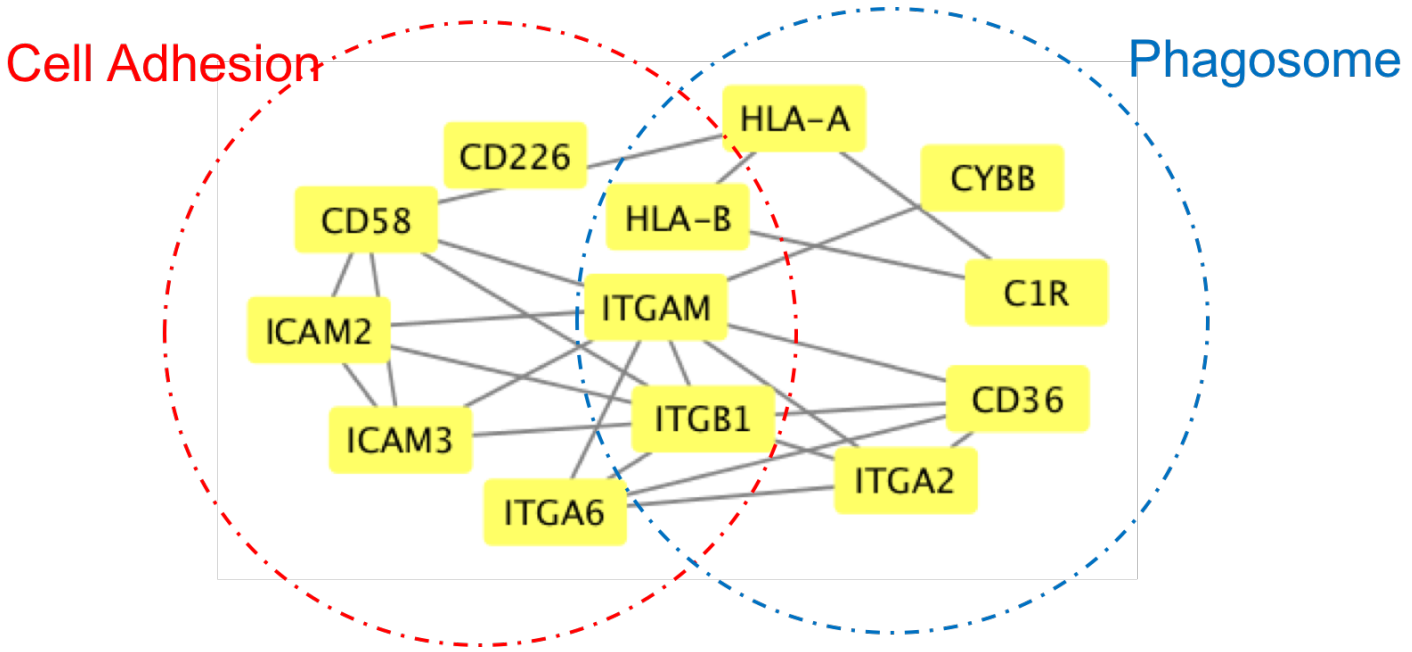

Figure 3.11 The common enriched networking in glycoproteomics for Luminal A/ B , Her 2 positive and triple negative.

PUBLICATIONS

1. **I-Hsuan Chen,** Liang Xue, Chuan-Chih Hsu, Juan Sebastian Paez Paez, Li Pan, Hillary Andaluz, Michael K. Wendt, Anton B. Iliuk, Jian-Kang Zhu, and W. Andy Tao* "**Phosphoproteins in extracellular vesicles as candidate markers for breast cancer",** *Proc. Natl. Acad. Sci. U.S.A.* **2017**, 114, 3175-3180.
2. **I-Hsuan Chen**, Hillary Andaluz, J. Sebastian Paez, Xiaofeng Wu, Michael K. Wendt, Li Pan Anton B. Iliuk, Ying Zhang*, and W. Andy Tao* "**Discovery and verification of glycoproteins from plasma-derived extracellular vesicles as breast cancer biomarkers."** *Analytical Chemistry* **(Accepted)**
3. Maoxiang Qian*, Liyuan Li*, **I-Hsuan Chen**, David Finkelstein, Arzu Onar-Thomas, Melissa Johnson, Christopher Calabrese, Armita Bahrami, Dolores López-Terrada, Jun Yang, Andy Tao, Liqin Zhu* **"Acquisition of Cholangiocarcinoma Traits during Advanced Hepatocellular Carcinoma Development in Mice"** *The American Journal of Pathology* **(In press)**