의학박사 학위논문

Biomarkers Profiling For Aggressive Breast Cancer Using Quantitative Proteomics And Bioinformatics

정량 단백체학 및 생물정보학을 이용한
공격적인 유방암 바이오 마커의 발굴

2022년 2월

서울대학교 대학원
의학과 병리학전공
김 혜 윤

Ph.D. Dissertation of Medicine

정량 단백체학 및 생물정보학을 이용한 공격적인 유방암 바이오 마커의 발굴

Biomarkers Profiling For Aggressive Breast Cancer Using Quantitative Proteomics And Bioinformatics

February 2022

Graduate School of
Seoul National University
College of Medicine
Pathology Major

Hyeyoon Kim

정량 단백체학 및 생물정보학을 이용한 공격적인 유방암 바이오 마커의 발굴

지도 교수 유 한 석

이 논문을 의학박사 학위논문으로 제출함
2021년 10월

서울대학교 대학원
의학과 병리학전공
김 혜 윤

김혜윤의 의학박사 학위논문을 인준함
2022년 1월

위 원 장 _____ (인)
부위원장 _____ (인)
위 원 _____ (인)
위 원 _____ (인)
위 원 _____ (인)

Biomarker Profiling For Aggressive Breast Cancer Using Quantitative Proteomics And Bioinformatics

Han Suk Ryu

Submitting a Ph.D. Dissertation of
Medicine

October 2021

Graduate School of (College of Medicine)
Seoul National University
Pathology Major

Hyeyoon Kim

Confirming the Ph.D. Dissertation written by
Full Name of you
January 2022

Chair _____(Seal)
Vice Chair _____(Seal)
Examiner _____(Seal)
Examiner _____(Seal)
Examiner _____(Seal)

General Abstract

Biomarkers Profiling For Aggressive Breast Cancer Using Quantitative Proteomics And Bioinformatics

Hyeyoon Kim

Department of Pathology, College of Medicine

Seoul National University

Graduate School

Mass spectrometry (MS)-based proteomics covers large-scale molecular and cellular biology at the protein level. Through the identification and quantification of proteins, the proteome analysis can interpret protein sequence, post-transcriptional modification and protein-protein interactions. This allows us to profile new disease biomarkers. From the cell lines to the limited amount of samples (body fluids, fresh frozen tissues, and FFPE tissues), thousands of proteins were discovered simultaneously to detect changes in expression level with disease status. The resulting expression data are complex and ambiguous patterns. Therefore, exquisite bioinformatics algorithms have to be applied to determine these unique biomarker patterns. A proteomic study discovers a list of biomarkers and helps elucidate the biological mechanisms.

In Chapter I, mass spectrometry-based proteomics was performed using breast cancer cells. To discover global proteome changes induced by CD44 expression levels, we regulated CD44 transcription by siRNA in two claudin-low breast cancer cell lines. For deep coverage of proteome, we used tandem mass tag-based MS analysis. We discovered 2736 proteins were upregulated and 2172 proteins were downregulated in CD44-knockdown MDA-MB-231 cells. For Hs 578T CD44-knockdown cells, 412 proteins were upregulated and 443 were downregulated. Informatics (Gene ontology and protein-protein interaction network) analysis demonstrated altered oncogenic cellular processes including proliferation, metabolism, and gene expression regulations. To confirm the changes of biology patterns, functional studies were conducted. As a result, we discovered that CD44-regulated proteome of claudin-low breast cancer cells, revealing changes that mediate cell proliferation and migration.

In Chapter II, label free-based MS proteomic analysis of clinical FFPE tissues. To discover candidate prognosis markers for distant metastasis of breast cancer, 10 no-metastasis, 9 late-metastasis, and 9 early-metastasis patients' primary tumor samples were analyzed. To achieve an in-depth proteome in the minimum of FFPE slides per sample, we performed well-defined proteomic strategies with high-resolution quadrupole Orbitrap LC-MS/MS. We identified a total of 9,455 protein groups using FFPE slides at 1% of the peptide and protein FDR level. Five biomarker candidates were differentially expressed using pair-wise comparison, and correlation network analysis filtered five candidates into two no metastasis specific and one late metastasis specific proteins. In addition, machine learning-based feature selection detected ten early metastasis classifier proteins, and the

system biology method filtered into seven proteins. For external validation, we used published mRNA data of breast primary tumor. Consequently, we suggested seven prognosis protein marker candidates that can help patients who need active treatment.

Keyword: Breast Cancer; Proteomics; Mass spectrometry; Biomarker; CD44; tumor progression; Formalin-fixed paraffin-embedded (FFPE); Distant metastasis; prognosis

Student Number: 2017-20079

* This work is published in Journal of Proteome Research.
Quantitative Proteomics Reveals Knockdown of CD44 Promotes Proliferation and Migration in Claudin-Low MDA-MB-231 and Hs 578T Breast Cancer Cell Lines (Kim, H., Woo, J., Dan, K., Lee, K. M., Jin, M. S., Park, I. A., Ryu, H. S., & Han, D.). Published 2 June 2021/ 10.1021/acs.jproteome.1c00293.

Table of Contents

List of Tables

Chapter I

List of Figures

Chapter I

List of Tables

Chapter II

List of Figures

Chapter II

List of Abbreviations

TNBC: triple negative breast cancer

MS: mass spectrometry

TMT: tandem mass tag

ECAR: extracellular acidification rate

OCR: oxygen consumption rate

DEP: differentially expressed protein

GO: gene ontology

KEGG: Kyoto encyclopedia of genes and genomes

PPI: protein protein interaction

GOBO: Gene expression-based Outcome for Breast cancer Online

CV: coefficient of variation

SD: standard deviation

PCA: principal component analysis

FDR: false discovery rate

FFPE: formalin-fixed paraffin-embedded

WGCNA: Weighted Gene Correlation Network Analysis

iBAQ: intensity-based absolute quantification

KM: Kaplan-Meier

DMFS: distant metastasis free survival

General Introduction

Breast Cancer is the most common women cancer worldwide and highly heterogenous disease on the molecular level [1]. The early detection and treatment reduced mortality rate; however, the heterogeneity of breast tumor could make difficult to clarify disease progression.

Untargeted MS-based proteomics has evolved as a powerful technology that enables the proteome-wide detection and quantification of proteins in complex samples. The proteins are the main functional molecules in the cells, their levels reflect accurately the cellular phenotype and the regulatory processes. The large depth of protein expression profiling data is powerful to understand the pathogenesis of breast cancer [2, 3]. Also, system-wide analyses of differential protein expression represent the entire set of proteins in a biological system and help to identify novel biomarkers.

In the novel biomarkers, single and multivariate markers exist. A single biomarker can help understand how molecular marker regulates tumor heterogeneity and biological function. In breast cancer, a single marker can predict prognosis and response to treatment [4-6]. Moreover, based on the wide variety of individual circumstances (physiological, environmental, and clinical factors) and the complexity of disease-inducing processes, multivariate markers can improve diagnostic power [7-9]. Thus, both single-marker and multi-marker analyses can identify potential biomarkers and improve our understanding of pathogenesis, providing novel therapeutic targets.

In Chapter I, the role of CD44, a cancer stem cell marker, was investigated as a single biomarker. CD44 is not a consistent cancer cell marker in luminal breast cancer subtypes, but it was in several studies to be highly overexpressed and to serve as a cancer stem cell marker in triple-negative subtypes [10]. Among triple- negative subtypes, claudin-low breast cancer represents the most aggressive molecular subtype that is comprised of cells that possess stem cell-like and mesenchymal features [11]. Given these points, to identify CD44-associated downstream proteins and functional regulatory roles in claudin-low breast cancer, we performed an in-depth tandem mass tag-based proteomic analysis of two claudin-low breast cancer cell lines (MDA-MB-231 and Hs 578T) transfected with CD44 siRNA. Bioinformatics analysis revealed molecular characteristics of CD44. Additionally, our results provided relationships between CD44 and CD44-associated proteins and insights into the molecular mechanisms leading to breast cancer progression.

In Chapter II, we developed a novel prognostic score with multiprotein markers for breast cancer distant metastasis patients. Although improvements of early detection in breast cancer makes for a decline in mortality rate, distant metastasis to other organs is the major-cancer related deaths [12]. Until now, the underlying mechanisms of distant-metastasis in breast cancer are currently not well understood. To discover potential biomarkers of early and late distant metastasis, we analyzed early and late distant metastasis of breast cancer using FFPE tissue specimens with a high-resolution mass spectrometry. Through combined supervised and unsupervised data analysis, we can suggest confident candidate distant metastasis regulated proteins. Our proteomic analysis can provide insights into the molecular mechanism leading to distant metastasis.

In chapter I and II, we ultimately intend to discover breast cancer pathogenesis that can help patients avoid unnecessary operations. Overall, using quantitative proteomics and bioinformatics we provided markers with clinical implication in aggressive breast cancer.

Chapter I

Quantitative Proteomics Reveals Knockdown of CD44 Promotes Proliferation and Migration in Claudin-Low MDA-MB-231 and Hs 578T Breast Cancer Cell Lines

Abstract

CD44 is a transmembrane glycoprotein that can regulate the oncogenic process. This is known to be a marker of the claudin-low subtype of breast cancer, as well as a cancer stem cell marker. However, its functional regulatory roles are poorly understood in claudin-low breast cancer. To gain comprehensive insight into the function of CD44, we performed an in-depth tandem mass tag-based proteomic analysis of two claudin-low breast cancer cell lines (MDA-MB-231 and Hs 578T) transfected with CD44 siRNA. As a result, we observed that 2736 proteins were upregulated and 2172 proteins were downregulated in CD44-knockdown MDA-MB-231 cells. For Hs 578T CD44-knockdown cells, 412 proteins were upregulated and 443 were downregulated. Gene ontology and network analyses demonstrated that the suppression of this marker mediates significant functional alterations related to oncogenic cellular processes, including proliferation, metabolism, adhesion, and gene expression regulation. A functional study confirmed that CD44 knockdown inhibited proliferation by regulating the expression of genes related to cell cycle, translation, and transcription. Moreover, this promoted the expression of multiple cell adhesion-associated proteins and attenuated cancer cell migration. Finally, our proteomic study defines the landscape of the CD44-regulated proteome of claudin-low breast cancer cells, revealing changes that mediate cell proliferation and migration. Our proteomics data set has been deposited to the ProteomeXchange Consortium via the PRIDE repository with the data set identifier PXD015171.

Introduction

CD44 is a transmembrane glycoprotein with a key role in cell adhesion to the extracellular matrix (ECM). Its interactions with appropriate ECM ligands, including hyaluronic acid, osteopontin, collagens, and matrix metalloproteinases, promote cell proliferation, cell adhesion, migration, and invasion [13, 14]. Depending on the cell type and growth conditions, several isoforms of CD44 can be generated by alternative splicing. The standard form of CD44 (CD44s) triggers oncogenic signaling, and multiple variant isoforms of CD44 (CD44v) function in promoting tumor progression and development to lymph node metastasis [15-17]. CD44s is expressed on most vertebrate cells, whereas CD44v is expressed in tumor cells or cells associated with the immune response [16, 18, 19].

CD44 expression is upregulated in various cancer cells [20, 21] and is recognized as a molecular marker for tumor-initiating cancer stem cells [22]. Especially, breast cancer cells show abnormal expression of CD44, as well as the heterogeneous expression of its isoforms [15]. Further, several studies have suggested that the role of CD44 is associated with aggressive tumor behavior [23, 24]. Moreover, breast tumors with high CD44 expression are enriched in claudin-low subtype signatures [25-27]. Interestingly, recent immunohistochemistry results showed that CD44-expressing breast cancer patients tended to survive longer than those with negative expression [28]. With this contrasting evidence, there is a constant need to understand the functional roles of CD44 in cellular functions in certain cell subtypes and extracellular conditions. However, the intricate association between CD44 and the underlying biological functions has not been fully disclosed at molecular levels. Although a few genomic studies [29] have been performed, the intracellular proteins

that are regulated by CD44 and the associated signaling pathways remain elusive.

Mass spectrometry (MS)-based proteomics has been recently used to provide comprehensive information about complex biological systems through large-scale protein identification and quantification. Especially, the development of multiplexed quantitation based on isobaric chemical tags, such as tandem mass tag (TMT), has allowed us to obtain quantitative information with deep coverage of the proteome and high multiplexing capacity, as compared to that with other conventional proteomic approaches [30]. Moreover, large-scale proteome data combined with bioinformatics analysis are expected to become a useful tool for the discovery of novel biological functions and networks [31, 32]. Several studies have provided a quantitative proteome map to compare similarities and differences in biological function between two similar systems[33, 34].

In this study, we discovered global proteome changes induced by CD44 knockdown in two triple-negative breast cancer (TNBC) claudin-low cell lines through TMT-based quantitative proteomics. The results provide not only an understanding of the biological functions of CD44 but also common and different functional changes among different breast cancer cell lines. Interestingly, the proteomes associated with CD44 were found to be highly involved in cell proliferation and migration.

Materials and Methods

1. Breast Cancer Cell Lines and Cell Culture

The TNBC cell lines MDA-MB-231 and Hs 578T were maintained at 37 °C in a humidified atmosphere containing 5% CO2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS; Hyclone, Boston, MA) and 1% Pen Strep (Gibco, Rockford, IL).

2. siRNA Transfection

For the knockdown experiment, siRNAs specific for CD44 and negative controls were purchased from Gene Pharma (Shanghai, China). The sequences of sense and antisense oligonucleotides for CD44 siRNA were as follows: sense, 5′-UAU UCC ACG UGG AGA AAA A-3′; antisense, 5′-UUU UUC UCC ACG UGG AAU ACA-3′. The sequences of sense and antisense oligonucleotides for the negative control were as follows: sense, 5′-UUC UCC GAA CGU GUC ACG UTT-3′; antisense, 5′-ACG UGA CAC GUU CGG AGA ATT-3′. For the proteomics analysis, 3 × 106 cells were plated in 6-well culture plates and transfected with 15 nM of negative control or CD44 siRNA using Lipofectamine RNAiMAX (Invitrogen, Carlsbad, CA) according to the manufacturer's procedure. Transfected cells were analyzed after incubation with the siRNA complex for 48 h.

3. Western Blotting

Proteins were extracted from cells using RIPA buffer (Sigma-Aldrich, St. Louis, MO) combined with a protease inhibitor cocktail (Roche, Basel, Switzerland). Lysates of proteins were subjected to SDS-PAGE and were transferred to a PVDF membrane (Millipore, Darmstadt, Germany). The membrane was blocked in 5% skim milk (BD, San Jose, CA) in TBS-T (0.05% (v/v) tween-20 in tris-buffered saline) and incubated

with primary antibodies. After washing in TBS-T, the membrane was incubated with horseradish peroxidase-conjugated secondary antibodies against mouse or rabbit. After washing in TBS-T, protein expression was detected with a biomolecular imager (GE, Boston, MA) using the membranes and an ECL solution (AB Frontier, Seoul, Korea).

4. Viability Assay

For the viability assay, 3 × 104 cells were seeded in five replicates onto 96-well culture plates (SPL, Seoul, Korea) for each group. After a 24-h cultivation, the cells were transfected with negative control and CD44 siRNA. After incubation for 0, 24, and 48 h, the viability of cells was assessed using the CellTiter 96 AQueous One Solution Cell Proliferation Assay (Promega, Madison, WI). After treatment with 20 μL of One Solution reagent containing MTS [3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium], each cell sample was incubated with 150 μL of fresh medium for 2 h at 37 °C with 5% CO2. The absorbance was measured at a wavelength of 490 nm.

5. Proliferation Assay

The proliferation assay was performed using the BrdU Cell Proliferation Assay Kit (Cell signaling Tech, Danvers, MA). For proliferation assays, 3 × 104 cells were seeded in five replicates into 96-well culture plates (SPL, Seoul, Korea) for each group. The cells were incubated overnight at 37 °C with 5% CO2 and transfected with negative control and CD44 siRNA. After incubation for 48 h, the proliferation of each cell group was assessed according to the manufacturer's procedure. The cells were incubated with BrdU for 6 h, and detection of BrdU incorporation was performed. The absorbance was measured at a wavelength of 450 nm.

6. Migration Assay

For the migration assay, transfected cells were re-suspended in serum-free DMEM and 5 × 104 cells were seeded onto the cell culture insert with an 8.0 μm pore membrane (Corning, NY). In the lower chamber, DMEM containing 10% FBS was added. The chambers were incubated at 37 °C in a humidified atmosphere containing 5% CO2. After 24 h, a wet cotton swab was used to wipe the cells from the upper space of chambers, whereas cells remaining on the bottom of the chambers were fixed with 4% formaldehyde in PBS and stained in a dye solution containing 0.1% crystal violet. The cells from three randomly selected high power fields (10×) were counted under a microscope (Nikon, Tokyo, Japan). Migration assays were conducted in triplicate.

7. Cell Metabolism Assay

A glycolytic rate assay kit was used to measure the glycolytic proton efflux rate (glyco PER). The assay was conducted according to the manufacturer's protocol. For the assay, transfected cells were seeded into Seahorse XF96 microplates (Agilent, CA) at a density of 104 cells per well and incubated to adhere overnight at 37 °C with 5% CO2. At 72 h after transfection, the glycolytic rate of each cell group was measured with a Seahorse XFe96 analyzer (Agilent, CA). Cells were treated with 0.5 μM rotenone and 0.5 μM antimycin A to block the mitochondrial activity. To inhibit glycolysis, 50 mM 2-deoxy-d-glucose was finally added. The extracellular acidification rate (ECAR) and the oxygen consumption rate (OCR) were measured. The Seahorse data were normalized to in situ cell counts based on a Cytation 1 system (BioTek, VT). After the glycolysis rate assay was completed, the membrane-permeable Hoechst stain, for nuclei, was injected. The results were analyzed with Agilent Seahorse Wave software (version 2.6.1). Unstimulated OCR and ECAR

values were normalized to the cell number, and the OCR/ECAR ratio was plotted.

8. Cell Lysis and Protein Digestion

Cell pellets were lysed in lysis buffer (4% SDS, 2 mM TCEP, and 0.1 M tris-HCl, pH 7.5) by direct sonication (10%, 3 cycles, 5 s, 2 s). Lysates were heated for 30 min at 95 °C. To minimize interference from the reducing reagent, the concentration of proteins was measured using a reducing agent-compatible BCA assay (Thermo Fisher Scientific, Waltham, MA). Then, each sample containing 300 μg of total protein was precipitated with cold acetone. Samples were re-suspended in denaturation buffer (2% sodium dodecyl sulfate (SDS), 10 mM tris (2-carboxyethyl) phosphine (TCEP), 50 mM chloroacetamide (CAA), and 0.1 M tris-HCl, pH 8.5) and heated for 15 min at 95 °C. The proteins were digested via multi-digestion filter-aided sample preparation according to our previously described process[35].(23) After UA buffer (8 M urea in 0.1 M tris-HCl, pH 8.5) was added to the samples, they were loaded onto a 30 K spin filter. Buffer was exchanged with UA solution twice and 50 mM triethylammonium bicarbonate (TEAB) solution three times by centrifugation. The first round of protein digestion was performed at 37 °C overnight using Trypsin/LysC (Promega, Madison, WI; protein-to-protease ratio = 100:1). The digested peptides were collected to a new collection tube by centrifugation. In the second digestion, remaining proteins in the filter were digested at 37 °C for 3 h using trypsin (protein-to-protease ratio = 200:1). Then, the digested peptides were collected by centrifugation, and an additional elution step was subsequently performed with 50 mM TEAB and 0.5 M NaCl.

9. TMT Labeling and Desalting

Peptide concentrations were measured by a tryptophan fluorescence assay [36]. TMT labeling was applied according to the manufacturer's protocol with some

modifications [37]. Briefly, the TMT reagent (0.8 mg) was dissolved in 100% acetonitrile. After spiking with 500 ng of peptides derived from ovalbumin as an internal standard, 25 μL of the reagent was added to 50 μg of peptide samples along with acetonitrile to give a final concentration of 30% (v/v). For MDA-MB-231 cells, negative control samples were labeled with the tags TMT-126, TMT-128, and TMT-130, whereas CD44 siRNA-transfected samples were labeled with the tags TMT-127, TMT-129, and TMT-131. For Hs 578T cells, negative control samples were labeled with the tags TMT-126, TMT-127, and TMT-128, whereas CD44 siRNA-transfected samples were labeled with the tags TMT-129, TMT-130, and TMT-131. After incubation at room temperature (22 °C) for 1 h, the reaction was quenched with 15.3 μL of 5% hydroxylamine. TMT-labeled samples were pooled at a 1:1:1:1:1:1 ratio. The resulting peptide mixtures were vacuum-centrifuged to dry and subjected to C18 solid-phase extraction (Waters, Milford, MA).

10. Offline High pH Reversed-Peptide (HPRP) Fractionation for Deep Coverage

The TMT-labeled peptide mixtures were subjected to HPRP-HPLC fractionation using an Agilent 1290 bioinert HPLC (Agilent, Santa Clara, CA) equipped with an analytical column (4.6 × 250 mm2, 5 μm) for fractionation. Solvent A consisted of 15 mM ammonium hydroxide in water and solvent B consisted of 15 mM ammonium hydroxide in 90% acetonitrile (ACN). The peptides were separated with a gradient of 5–35% ACN at a flow rate of 0.2 mL/min. In total, 96 fractions were concatenated to mix different parts of the gradient into 12 fractions. The fractions were lyophilized and stored at −80 °C until MS analysis.

11. Mass Spectrometry and Data Analysis

The fractionated peptide samples were analyzed with an LC-MS system (Quadrupole

Orbitrap mass spectrometers, Q-exactive plus, Thermo Fisher Scientific, Waltham, MA) equipped with an Ultimate 3000 RSLC system (Dionex, Sunnyvale, CA) via EASY-Spray LC columns as the electrospray source, and the temperature of the column heater was set to 60 °C. Peptides were separated on a 2-column system with a trap column (3 mm diameter, 1 cm length) and an analytic column (75 μm diameter, 50 cm length) using 0.1% formic acid in water as solvent A and 0.1% formic acid in acetonitrile as solvent B. The samples were separated using a 240 min gradient from 7 to 32% solvent B at a flow rate of 300 nL/min. The survey MS scans were acquired in the range of 350–1650 m/z with a resolution of 70 000 at m/z 200. The Q-exactive was operated in the data-dependent mode using a top 20 method to select up to the 20 most abundant precursor ions with an isolation width of 1.2 m/z. High-energy collisional dissociation scans were acquired with a normalized collision energy of 32 and a resolution of 35 000 at 200 m/z. The maximum ion injection time for the survey scan and MS/MS scan was 20 and 100 ms, respectively.

Raw MS/MS files were processed with Proteome Discoverer version 2.2 (Thermo Fisher Scientific, Waltham, MA), using the SEQUEST HT algorithms against the Human Uniprot protein sequence database (December_2014, 88 657 entries). The database search parameters were as follows: full enzyme digest using trypsin (after KR/-) with up to two missed cleavages allowed; a precursor ion mass tolerance of 20 ppm; a fragment ion mass tolerance of 0.02 Da; dynamic modifications of 15.995 Da for methionine oxidation and 42.011 Da for protein N-term acetylation; and static modifications of 57.021 Da for carbamidomethylation on cysteine residues and 229.153 Da for TMT 6plex on any N-terminus. The co-isolation threshold was set to 50%. Six reporter ion intensities for TMT were corrected for isotopic impurities as provided by the manufacturer. Peptide spectral matches and peptides were confirmed

by a Percolator based on a 1% false discovery rate (FDR). Confidence criteria were set to a 1% FDR at the protein level. The MS-based proteomics data of all identified peptides and protein lists have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository, with the data set identifier PXD015171.

12. Statistical Analysis

Perseus software (version 1.6.2.2) was used for all statistical analyses [38]. For pair-wise comparisons, the statistical analysis was performed based on the logarithmic intensities of TMT-reporter ions. Normalization was performed by width adjustment. Proteins with a permutation-based FDR value <0.05 using a Student's t-test were considered significantly differentially expressed proteins (DEPs).

13. Bioinformatics Analysis

Functional gene classification was performed with the Proteomaps web tool (version 2.0) based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [39]. To visualize functional trends in the proteome, DEPs were uploaded. Each polygon indicates individual proteins, and the size of polygons expresses the proportion of different values of DEPs based on a t-test.

2D annotation enrichment was performed with Perseus software [34]. 2D annotation enrichment for gene ontology (GO) and KEGG terms was disseminated to discover the significant differences between two related systems. The MANOVA test was used for two groups in two dimensions based on the multivariate data (p-value <0.05). For significant annotation terms, a position score was calculated. A position score determines the distribution of protein fold-changes that are part of the annotation term. This score was rescaled to range from −1 to 1. A value near 1 indicates that the annotation term is concentrated with upregulated proteins in the distribution.

Otherwise, a value near −1 indicates that the annotation term is concentrated with downregulated proteins. A pair of scores indicates correlating patterns of enrichment terms between two systems. The 2D annotation enrichment results were summarized and visualized using the REVIGO tool [40].

14. Protein Network Analysis

Lists of significantly downregulated proteins were analyzed using the HumanBase (https://hb.flatironinstitute.org/) module detection function to construct functional networks. The network was generated based on shared k-nearest neighbors and each network was clustered into distinct modules with the Louvain community-finding algorithm [41]. In each module, GO biological process terms were discovered via one-sided Fisher's exact tests. Benjamini–Hochberg corrections were used to calculate q-values. For protein–protein interaction (PPI) analysis, the STRING (version 11.0) database was used [42]. The networks were visualized using Cytoscape (version 3.8.1) [43].

Results

1. Establishment of CD44-Knockdown Cell Line Model

To identify CD44-associated downstream proteins and molecular processes in claudin-low breast cancer, siRNA-mediated knockdown approaches were applied to breast cancer cell lines. Because the claudin-low cell lines are known to express F11R, EPCAM, and Claudin proteins at very low levels, whereas CD44 is expressed at a high level, we first assessed mRNA expression levels of these markers across the 51 breast cancer cell lines using GOBO (Gene expression-based Outcome for Breast cancer Online) gene set analysis (version 1.0.3.) [44]. Expectedly, mRNA expression patterns showed that the Basal B subgroup is associated with a claudin-low subtype and a more stem cell-like phenotype [25, 45] (Figures 1A and 1B). Moreover, we confirmed that the Basal B subgroup included nine breast cancer cell lines (MDA-MB-157, MDA-MB-231, MDA-MB-435, MDA-MB-436, SUM-159PT, SUM-1315, BT-549, Hs 578T, and HBL-100) previously identified as claudin-low [25]. Except for two problematic cell lines (MDA-MB-435 and HBL-100), the characterization of seven claudin-low cell lines is summarized in Table 1. Among these seven cell lines, gene expression profiling [44, 46] and proteome [47] showed that Hs 578T, MDA-MB-231, and MDA-MB-157 cells were closely clustered into the mesenchymal stem-like cell group, as well as the claudin-low group. Since CD44 is well known as a mesenchymal stem cell marker, we choose these cell lines for further analysis. To assess the basal protein expression levels of CD44, we performed western blot analysis of several breast cancer cell lines including three cell lines (Figure 1C). Likewise, the highest CD44 protein expression was found among claudin-low cell

lines. It was of interest to note that the MDA-MB-157, MDA-MB-231, and Hs 578T cell lines predominantly expressed CD44s. Finally, we decided to explore the effects of CD44 knockdown on MDA-MB-231 and Hs 578T cells to investigate the common features generated upon CD44 suppression despite the different origins. After transfection with negative control and CD44 siRNA (30 nM), we confirmed that the siRNA targeting CD44 could reduce the CD44 levels to 32 and 40% of control levels in MDA-MB-231 and Hs 578T cells, respectively (Figure 1D).

Table 1. Summary of claudin-low breast cancer cell lines

Cell line	molecular type	Origin	Disease
MDA-MB-157	Basal B	metastasis	Breast carcinoma
MDA-MB-231	Basal B	metastasis	Breast adenocarcinoma
MDA-MB-436	Basal B	metastasis	Invasive ductal carcinoma
SUM-159PT	Basal B	primary	Pleomorphic breast carcinoma
SUM-1315MO2	Basal B	metastasis	Invasive ductal carcinoma
BT-549	Basal B	primary	Invasive ductal carcinoma
Hs 578T	Basal B	primary	Invasive ductal carcinoma

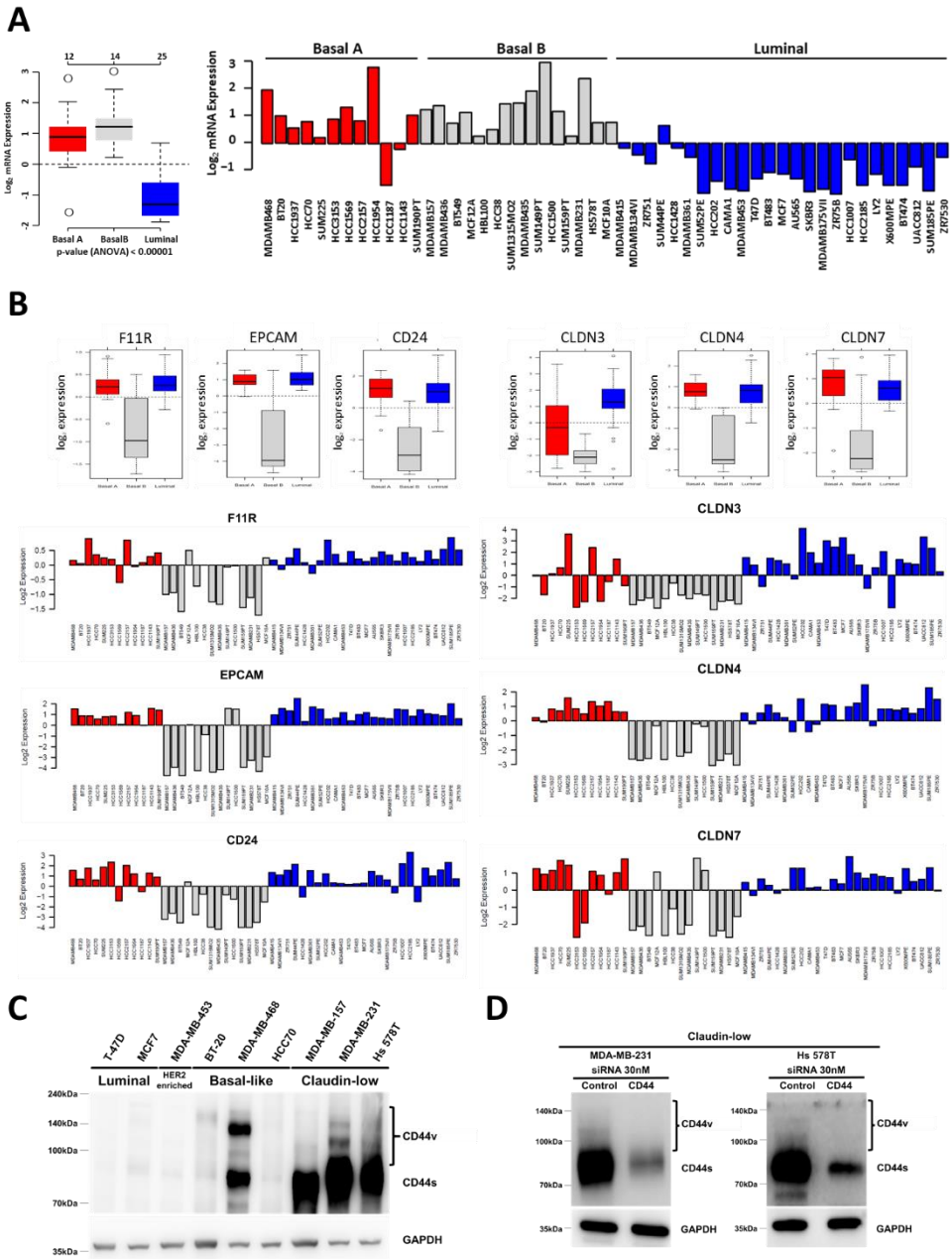


Figure 1. **Basal CD44 expression is prominent in the claudin-low breast cancer subtype.** (A) Relative mRNA expression of CD44 in breast cancer cell lines, classified as basal (gray), claudin-low (red), and luminal (blue). Bar graph on the left represents the expression across the individual 51 cell lines. Box plot displays on the right according to the breast cancer intrinsic subtypes. (B) Relative mRNA expression of Claudin-low subtype markers. (C) Basal CD44 levels in various human breast cancer cell lines. The different cell lines were harvested in log phase and analyzed by western blotting with antibodies against CD44 and GAPDH (as a loading control). The breast cancer subtype of each cell line is indicated on the upper side. (D) CD44 siRNA transfection decreases the levels of CD44 proteins in two different cancer cell lines, analyzed by western blotting.

2. Quantitative Analysis of Differential Protein Expression Following CD44 Knockdown

To identify the downstream proteins and molecular processes induced by the knockdown of CD44, we investigated global proteome profiles employing two TMT 6-plex experiments (Figure 2A). Briefly, cell lysates were digested via filter-aided sample preparation [48], and the peptide samples were labeled with TMT reagents for each cell line. The labeled peptide mixtures were separated into 12 fractions using high-pH reversed-phase fractionation and analyzed in 3 h LC-MS/MS experiments on an Orbitrap mass spectrometer (Q-exactive plus). In this study, we identified 7396 and 6567 protein groups at the protein level with an FDR <1% in MDA-MB-231 and Hs 578T cells, respectively. Considering only proteins observed in all TMT channels and with at least one unique peptide, 7298 and 6511 protein groups were quantified in each cell line (Figure 2B). Of these, 5614 were detected in both cell lines, whereas 1151 and 775 were unique to MDA-MB-231 and Hs 578T cells, respectively. Quantified proteins with one unique peptide included 805 and 890 proteins in MDA-MB-231 and Hs 578T cells, respectively.

In TMT data, we noticed that the level of CD44, serving as a positive control, was downregulated in all CD44 siRNA-treated cells (Figures 2C and 2D). Regarding the reproducibility of quantitative variation, we used a nonhomologous spiked-in internal standard (ovalbumin). The coefficient of variation (CV) for ovalbumin was 4.6 and 5.3% in MDA-MB-231 and Hs 578T cells, respectively (Figure 2E). The results indicated that our experimental procedures accurately identified and quantified the proteins among the 6-channel TMT.

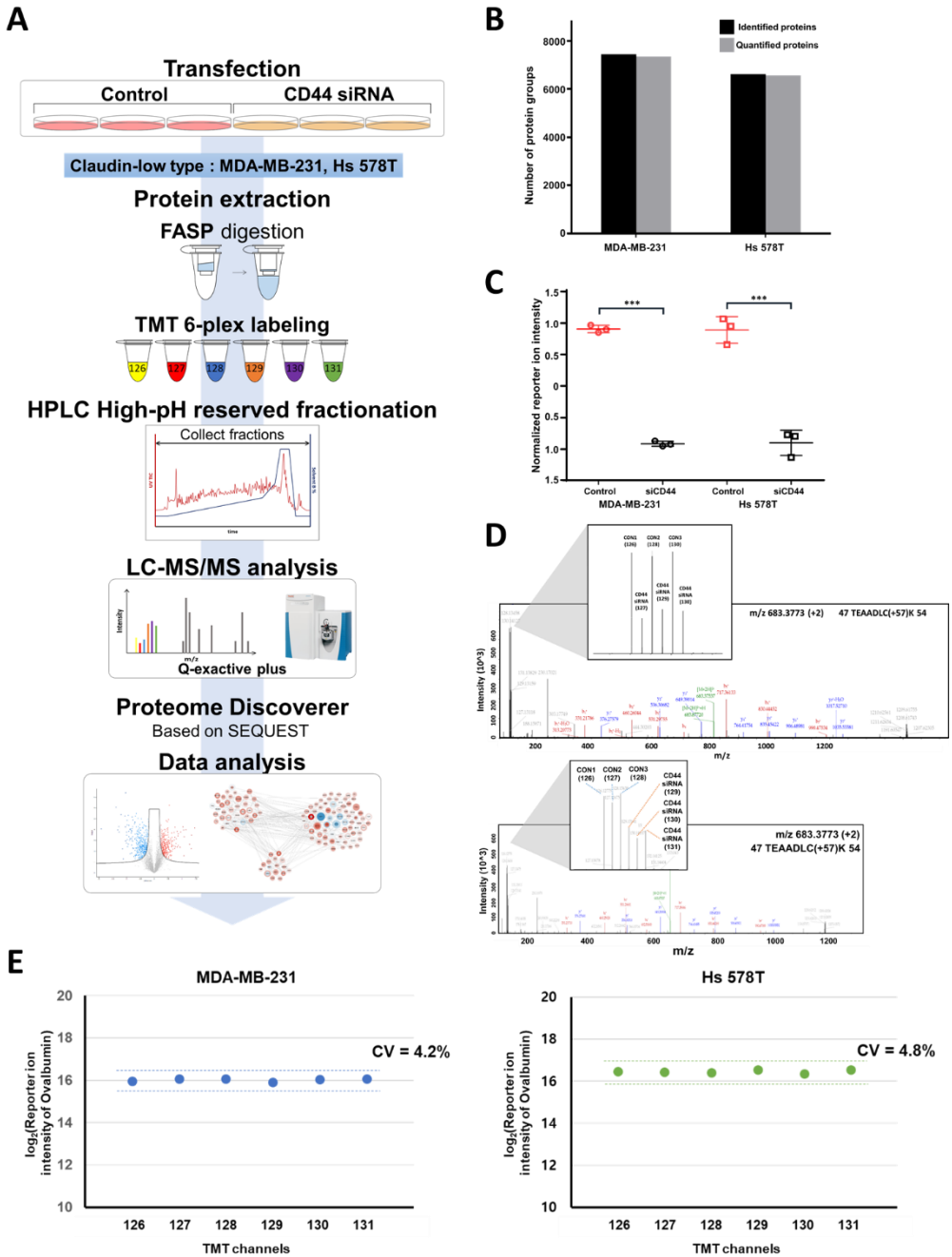


Figure 2 Mass spectrometry-based profiling of claudin-low breast cancer cell lines. (A) Experimental overview: quantitative proteomic analysis of cell lines. (B) Black bar graph showing the number of proteins identified in each cell line. Gray bar graph showing the number of proteins quantified in each cell line. (C) Dot plot showing the z-score-normalized reporter ion intensity of CD44 protein for each condition. Student's t-test was used to calculate p-values (***$p < 0.0005$). (D) Representative MS/MS spectra of TEAADLC peptide in MDA-MB-231 and Hs578T. (E) Distribution of TMT reporter ion intensities for internal standard Ovalbumin peptides in MDA-MB-231(A) and Hs 578T(B) cells.

In addition, the mean CV values between the biological triplicates in each condition for the abundance of overall proteins were below 10% in each cell line. Moreover, approximately 99% of quantified proteins were within the 20% CV cutoff range. CV distribution according to unique peptides showed that our TMT quantification had good reproducibility and precision, regardless of the number of unique peptides used for identification.

Interestingly, principal component analysis (PCA) revealed that negative control- and CD44 siRNA-transfected cells could be clearly discriminated by the proteome profiling despite a high correlation between both conditions (Figure 3A). To identify DEPs between control and CD44-knockdown cells, a two-sample t-test was performed using a permutation-based FDR <5% as a threshold (Figure 3B).

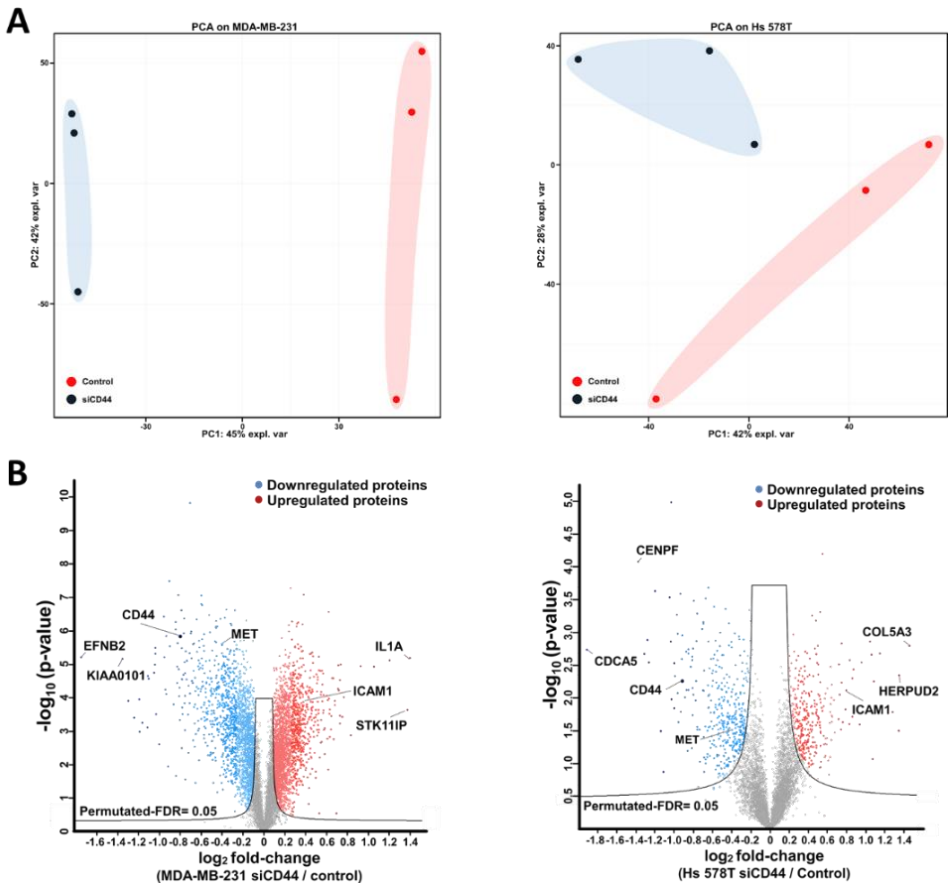


Figure 3 Analysis of the altered proteins with downregulated CD44 levels in MDA-MB-231 and Hs 578T cells. (A) Principal component analysis (PCA) of proteins identified in each cell line. The red spots represent control cases and the black spots are CD44 siRNA-transfected cases. (B) Scatter plots of proteome comparison between CD44 siRNA-transfected and control cases. Significant proteins had an FDR-adjusted p-value less than 0.05. Gray lines indicate significance thresholds of FDR <5%. The highlighted and labeled proteins were validated by western blot.

As a result, 4908 and 855 proteins were statistically significant in MDA-MB-231 (2736 upregulated and 2172 downregulated proteins) and Hs 578T (412 upregulated and 443 downregulated proteins) cells, respectively. To validate the quantitative results based on TMT, western blotting was performed to examine the expression levels of top-ranked DEPs (Figures 4). In addition, we validated ICAM1 and MET as common DEPs between two cell lines. It could be seen that the protein expression trends were in high accordance with proteomic assay results.

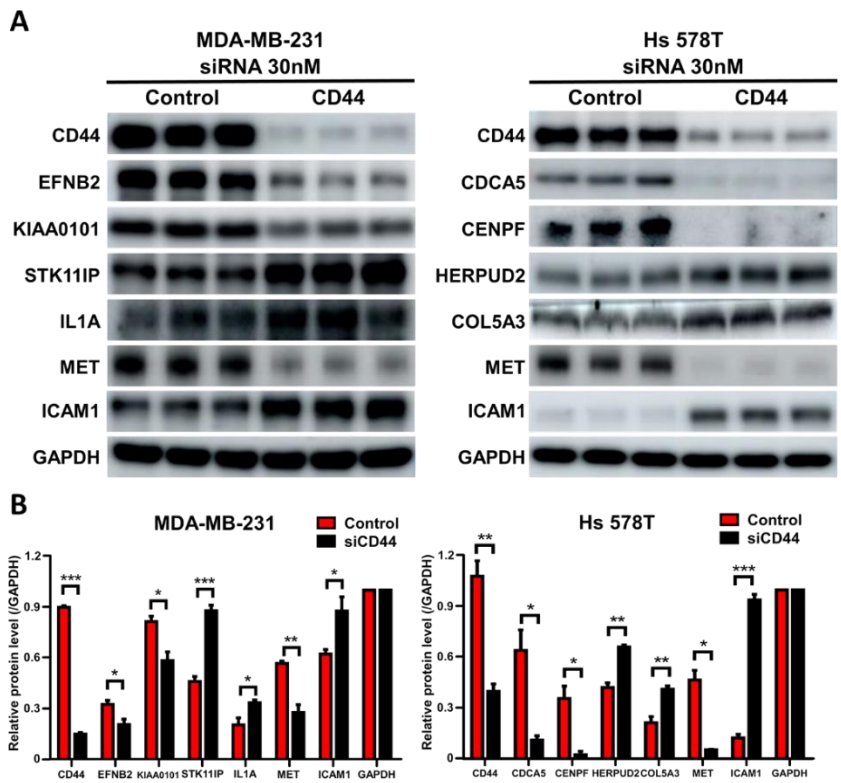


Figure 4 Validation of proteomics results. (A) Western blot validation of top-ranked DEPs in MDA-MB-231 (EFNB2, KIAA0101, STK11IP, and IL1A) and Hs 578T (CDCA5, CENPF, HERPUD2, and COL5A3). ICAM1 and MET were validated as common DEPs between two cell lines. CD44 was used as positive control. GAPDH was used as the loading control. (B) Bar graphs showing the normalized protein levels in control and CD44 siRNA-transfected cells. Red color indicates negative control-transfected cells and black color indicates CD44 siRNA-transfected cells. Results are plotted as means ± standard deviations (SD) of values were obtained from three independent experiments. Student's t-test was used to calculate *p*-values (*$p < 0.05$, **$p < 0.005$, and ***$p < 0.0005$).

3. Biological Functions of Proteins Regulated by CD44

The quantitative proteomic analysis revealed that a decrease in CD44 expression could affect the expression of many proteins. In efforts to discover cellular functions affected by the knockdown of CD44 in claudin-low-type breast cancer cells, we constructed a Proteomap [39] to cluster the DEPs according to their KEGG pathway annotations and found a striking resemblance between the maps of the two cell lines (Figures 5A). Each polygon of the Proteomap corresponds to a single KEGG pathway term, and the size correlates with the ratio between the groups. In both cell lines, upregulated proteins were dominated by higher levels of metabolism-related proteins, whereas downregulated proteins were dominated by spliceosome, chromosome, and cell cycle-related proteins.

Metabolic mechanisms considerably control breast cancer cell fate by switching from oxygen-dependent to oxygen-independent pathways of glycolysis[49, 50]. (38,39) It was revealed that CD44 enhances glycolytic flux rather than mitochondrial respiration[51]. (40) Remarkably, proteins that regulate the TCA cycle were highly enriched only in MDA-MB-231 cells. It was also found that basal metabolic profiles of MDA-MB-231 and Hs 578T cells were quite distinct [52]. Compared to the Hs 578T cells, MDA-MB-231 cells show strong metabolic shifts affecting TCA cycle metabolites.

To investigate the effects of CD44 knockdown on cellular metabolism, we performed a Seahorse glycolytic rate assay. The ECAR and OCR were assessed. We found that CD44 knockdown affects the metabolic phenotype of MDA-MB-231 and Hs 578T cells (Figure 6).

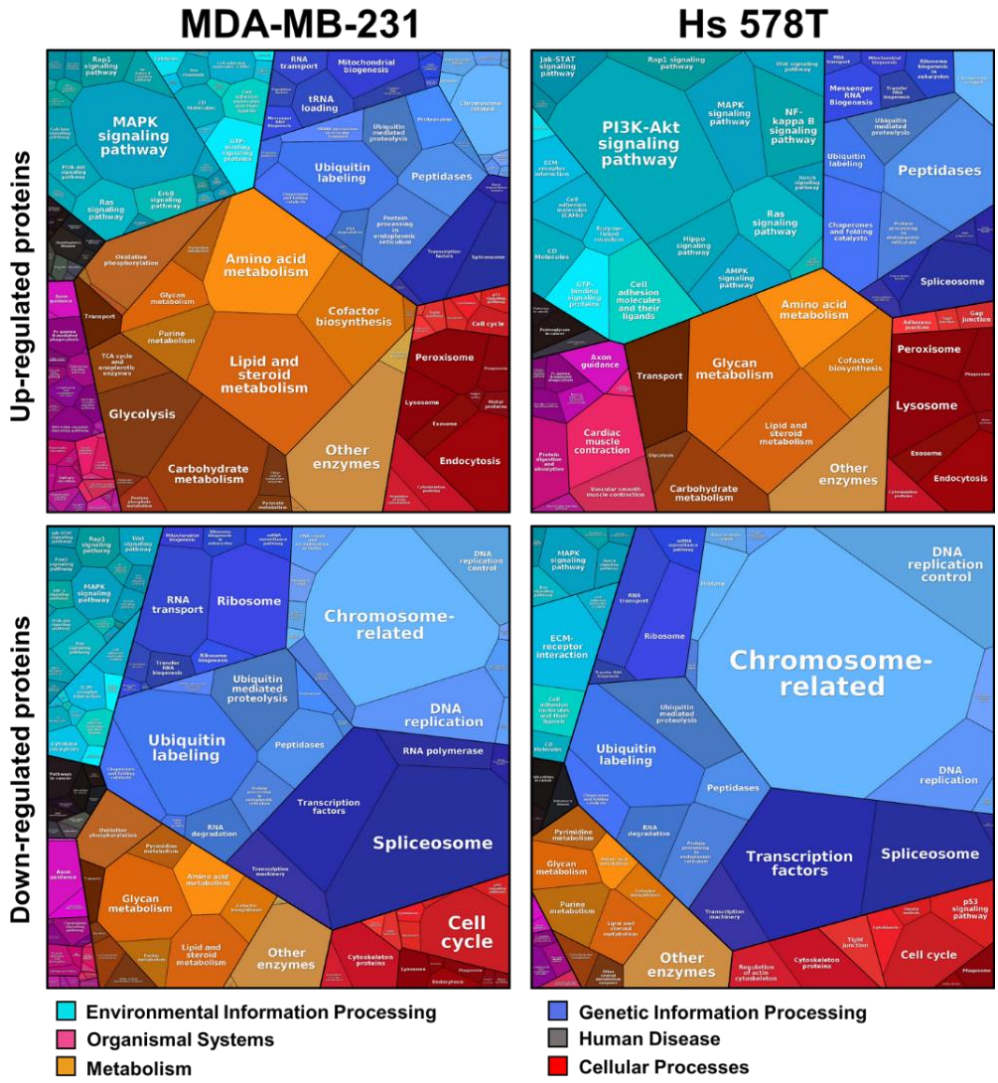


Figure 5 Biological classification of differentially expressed proteins.
Treemap of cellular categories altered by siCD44 transfection illustrated by Proteomaps. The conditions of each cell line are marked on the upper side.

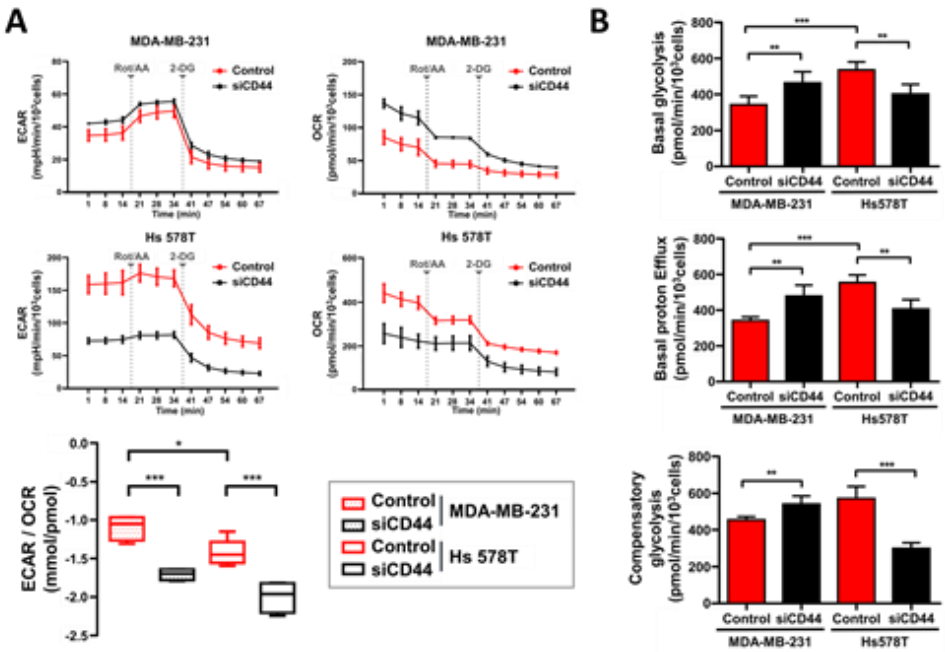


Figure 6 . A metabolic change induced by CD44 knockdown in claudin-low breast cancer cells. Bioenergetics of MDA-MB-231 and Hs578T cells were assessed by Seahorse Flux analyser using Glycolysis rate assay kit. The ECAR and OCR were measured under basal condition. Next, the ECAR and OCR were measured in the presence of rotenone and mitochondrial complex inhibitor antimycin A (Rot/AA) and glycolytic inhibitor 2-DG. The glycolytic inhibitor 2-DG was added to the end.
(A) The effect of CD44 expression level on ECAR, OCR and ECAR / OCR ratio in MDA-MB-231 and Hs578T cells. (B) Effects of CD44 knockdown on the glycolytic rate. The basal glycolysis (top), basal proton efflux (middle), and compensatory glycolysis (bottom) levels in MDA-MB-231 and Hs 578T cells. The data are mean ± SD, n=5 wells per group. *, $P < 0.05$, **,$P < 0.005$ and ***, $P < 0.0005$ compared with control cells.

First, at the basal level, Hs 578T control cells displayed an approximately 1.63-fold higher basal glycolytic rate compared to that of MDA-MB-231 control cells, consistent with results of a previous study [52]. In MDA-MB-231 cells, the knockdown of CD44 stimulated basal glycolysis and compensatory glycolysis, whereas CD44 knockdown in Hs 578T cells decreased basal glycolysis and compensatory glycolysis. Interestingly, in both cell lines, we observed a shift from a high to low ECAR/OCR ratio in CD44 siRNA-transfected cells compared to that in control cells (Figure 6). In accordance with a previous report [51], the low ECAR/OCR ratio indicated that CD44-knockdown cells rely more on oxidative

phosphorylation than on glycolysis. These cell line-specific metabolic dynamics possibly cause phenotype diversity between similar claudin-low breast cancer cells. Moreover, our results indicate that lipid metabolic proteins are increased in MDA-MB-231 and Hs 578T cells. Prior research suggests that lipid metabolic dysregulation correlates with the expression level of CD44 protein [53]. Further, downregulated CD44 expression might enhance the consumption of lipids through fatty acid β-oxidation, which could reduce cancer invasiveness.

Along with the metabolic categories, upregulated proteins from both cell lines had higher proportions of cell adhesion molecules, as well as signaling-related categories including mitogen-activated protein kinase (MAPK), PI3K-AKT, nuclear factor-kB (NF-kB), and RAS signaling pathways. CD44 can influence the activity of various oncogenic signaling pathways [13, 54]. Interestingly, the Proteomap showed that the MAPK signaling pathway was the biggest contributor to this category in MDA-MB-231 cells; however, in Hs 578T cells, the PI3K-Akt signaling pathway was mainly enriched. Previous studies have demonstrated that activation of the signaling pathway depends on the mutation status of cell lines. MDA-MB231 cells harbor KRAS(G13D) and BRAF(G464V) mutations that lead to the hyperactivation of the MAPK/ERK pathway in association with the suppression of Akt phosphorylation [55-57]. Hs 578T cells harbor an HRAS(G12D) mutation that leads to the hyperactivation of both the PI3K/Akt and MAPK/ERK pathways [58]. CD44 regulation affected distinct signaling pathways, indicating that even if it is expressed at a similar level in the two claudin-low breast cancer cell lines, its functions might depend on the difference in the genetic background and origin.

Downregulated proteins were significantly associated with cell proliferation, including the cell cycle, DNA replication, and chromosome-related categories, in

accordance with the well-known roles of CD44 in cell proliferation in breast cancer [59, 60]. Interestingly, a recent study showed that claudin-low cell lines are sensitive to drug compounds that interfere directly with the cell cycle, mitotic spindle function, or chromosome segregation [61, 62]. In addition, Prat et al. [63] showed that claudin-low cell lines have higher expression of proliferation-related genes than other types of cell lines. Because the rate of proliferation of the cell lines is another factor that might influence drug sensitivity, CD44 knockdown could sensitize claudin-low cancer cells to drugs that block the cell cycle.

To help display different and common cellular functions between two similar systems, we employed 2D annotation enrichment analysis [34] using the overlapped DEPs between comparison sets (544 proteins in MDA-MB-231 and Hs 578T cells). Largely, scatter plots showed concordant up- or downregulation of GO terms across the two cell lines. As can be seen in Figure 7, proteins associated with cell adhesion, immune response, metabolic process, and ion transport were upregulated in CD44-downregulated cell lines (Table S8). Moreover, mitochondrial and secreted proteins were significantly upregulated in both cell lines. In contrast, proteins associated with nucleic acid metabolism, like the regulation of transcription, RNA metabolic process, regulation of gene expression, and other related terms, were attenuated, suggesting that CD44 might regulate the expression of these proteins through transcription and RNA processing [64].

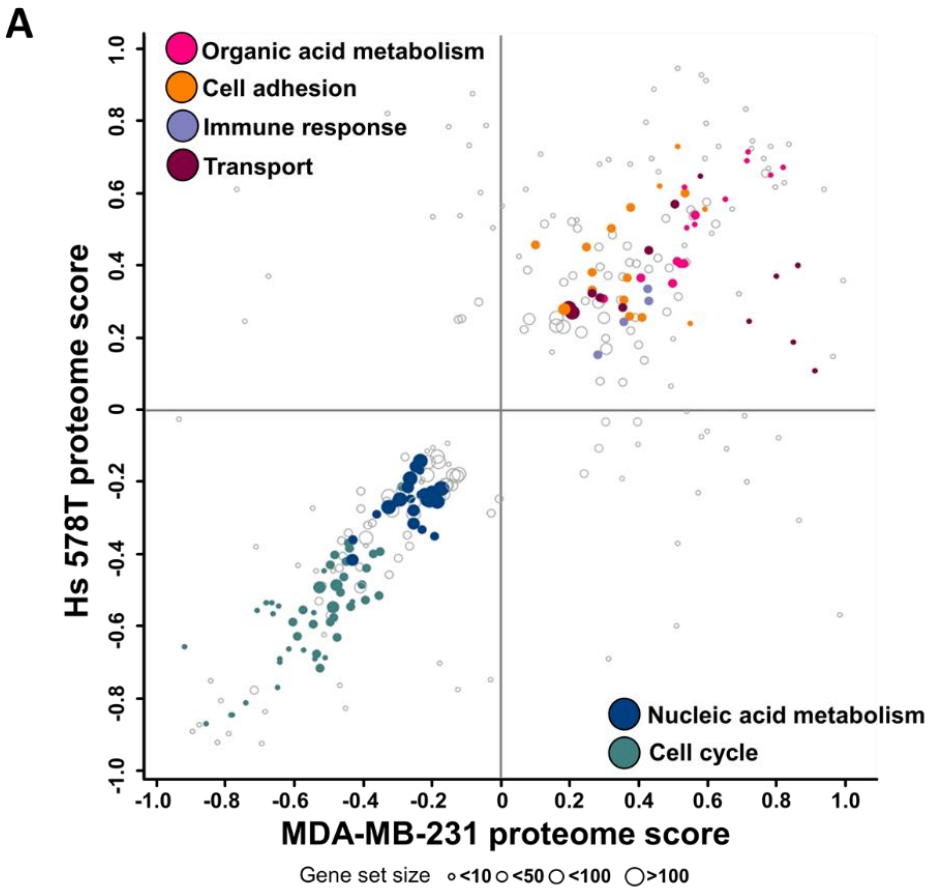


Figure 7 Biological classification of common differentially expressed proteins.
2D annotation distribution. Scatter plots of annotation changes after normalization between the two cell lines. The annotations analyzed were GOBP, GOCC, GOMF, and KEGG enrichment. Circle size is proportional to the size of annotated gene sets.

Table 2 Summary of the Two-dimensional annotation enrichment analysis

Term ID	Term name	Size	*p*-value	MDA-MB-231 Proteome Score	Hs 578T Proteome Score
GO:0022610	cell adhesion	27	0.0036	0.246364	0.455119
GO:0043269	regulation of ion transport	8	0.0054	0.722481	0.246735
GO:0006955	immune response	27	0.0072	0.429902	0.307114

For an unbiased interpretation, we reanalyzed the result of 2D annotation enrichment using the REVIGO tool [40], which is a web server that summarizes the lists of GO terms by finding a representative subset of the terms using a simple clustering algorithm that relies on semantic similarity measures (Figure 8). In agreement with the Proteomap, this analysis revealed that the overexpressed genes were associated with molecular processes involved in cell adhesion, metabolic processes, and mitochondrial function, whereas the downregulated genes corresponded to transcription, translation, and cell cycle, which are associated with proliferative cells. Hence, a reduction in CD44 leads to widespread proteome changes that might control the proliferation and migration of breast cancer cells.

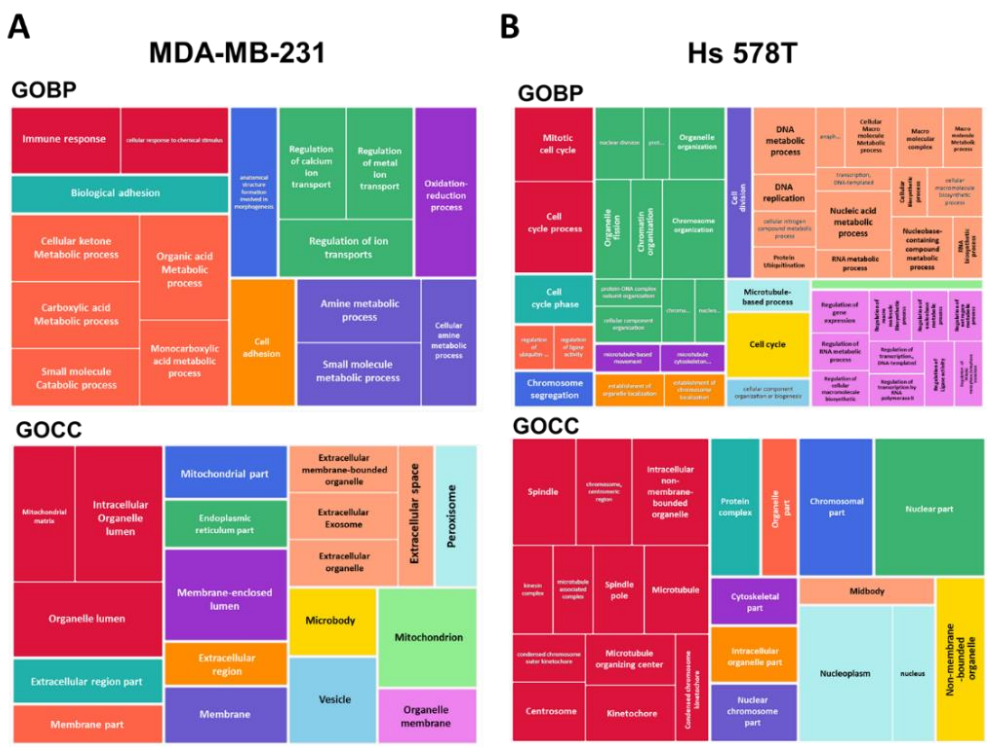


Figure 8 Treemaps of 2D annotation enrichment analysis.
Treemaps were created by REVIGO web server. (A) MDA-MB-231 (B) Hs 578T

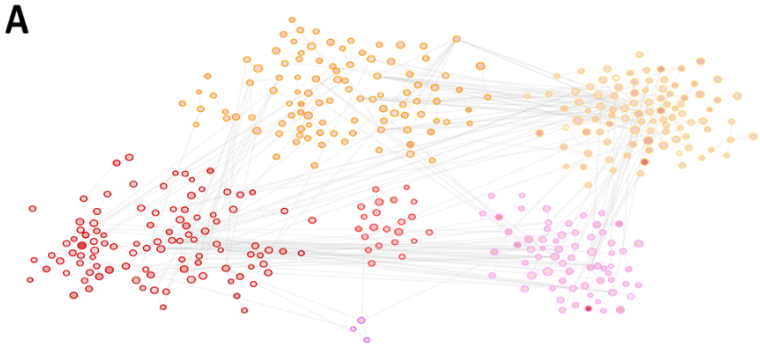
4. Functional Characterization of Genes Regulated by CD44

To gain further insight into CD44 functions in breast cancer cells, we employed large-scale network-based functional prediction clustering analysis [65] with significantly up- and downregulated proteins. For upregulated proteins, we identified six modules in MDA-MB 231 cells (Figure 9A). M1 was associated with the immune response, whereas M2 and M3 were enriched in fatty acid β-oxidation and metabolic process. Modification-related terms were enriched in M4, M5, and M6. In contrast to that in MDA-MB 231 cells, network clusters in Hs 578T cells showed that the cell structure and localization-related terms were enriched in most modules including M1, M3, M4, M5, and M6 (Figure 9B). Splicing-associated terms were identified in M2. Of note, no common modules were observed in both cell lines, possibly due to the low statistical power of annotation enrichment for upregulated proteins. In downregulated proteins, we identified five modules using functional network analysis of MBA-MB-231 cells (Figure 10A). M1 and M2 were enriched in the cell cycle, including cell division, chromosome segregation, and DNA replication. Biological processes related to gene expression processing, such as translation, transcription, and mRNA splicing, were also identified in M4 and M5. In module 3, cell migration-, adhesion-, and morphogenesis-related terms were observed. In the case of Hs 578T cells, we identified six modules (Figure 10B). M1, M2, and M6 were enriched in the cell cycle, whereas gene expression processing-associated terms were found in M3 and M4, and M5 was related to cell migration.

In terms of commonalities, 161 proteins (M1 and M2) in MDA-MB-231 cells and 149 proteins (M1, M2 and M6) in Hs 578T cells were identified as involved in cell cycle-related terms, including chromosome segregation, mitotic nuclear division,

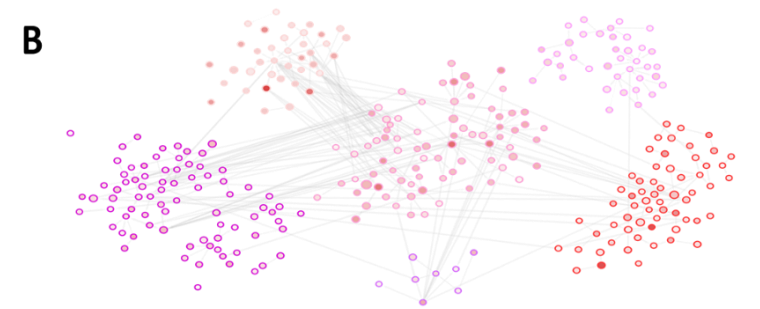
and microtubule cytoskeleton organization. These modules were also comprised of a proliferative signature of cancer cells, including CCNA2, CCNB1, CDC20, and PLK1. Western blotting was performed to confirm quantitative results of these cell cycle-related proteins (Figure 11A). Several studies have found that cell cycle-regulated genes are significantly altered in highly proliferative breast cancer cells and tumors [66]. Interestingly, CD44 was found in the cell migration-related modules M3 in MDA-MB-231 cells and M5 in Hs 578T cells.

A



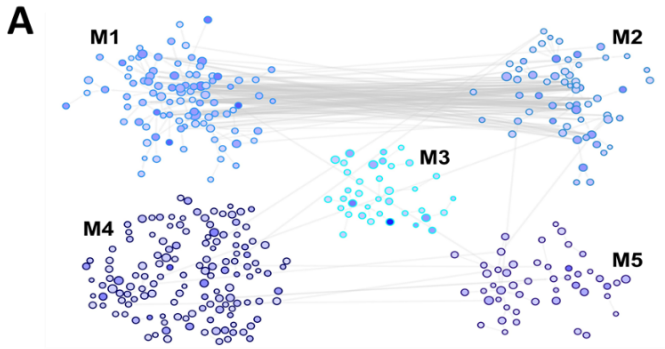
	Cluster genes	Cluster terms	Top terms	q-value
M1	106	282	response to virus	3.426E-05
			innate immune response	0.0001736
			regulation of multi-organism process	0.0001736
M2	114	93	generation of precursor metabolites and energy	0.0001736
			electron transport chain	0.0002687
			fatty acid beta-oxidation	0.0021475
M3	102	70	galactose metabolic process	0.0001736
			fatty acid beta-oxidation using acyl-CoA dehydrogenase	0.0013246
			mitochondrion organization	0.0028506
M4	64	62	positive regulation of histone modification	0.0018650
			positive regulation of chromatin organization	0.0044249
			phosphatidic acid biosynthetic process	0.008378
M5	67	93	tubulin deacetylation	0.0063606
			iron ion import	0.0093950
			protein deacetylation	0.0133081
M6	3	2	peptidyl-tyrosine phosphorylation	0.0128045
			peptidyl-tyrosine modification	0.0129084

B

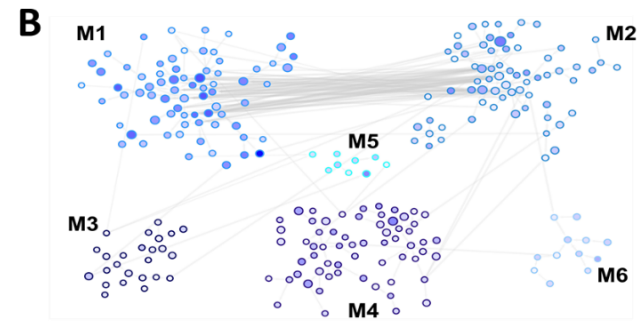


	Cluster genes	Cluster terms	Top terms	q-value
M1	82	318	collagen fibril organization	2.50E-08
			response to wounding	8.42E-08
			extracellular matrix organization	2.01E-07
M2	39	35	mRNA splicing, via spliceosome	0.000492
			RNA splicing	0.001137
			mRNA processing	0.001208
M3	63	126	regulation of protein localization to plasma membrane	0.001105
			regulation of protein localization to cell periphery	0.001308
			clathrin coat assembly	0.004316
M4	85	199	actin cytoskeleton organization	0.001308
			negative regulation of hyaluronan biosynthetic process	0.001615
			extrinsic apoptotic signaling pathway	0.004316
M5	41	73	regulation of cartilage development	0.006342
			actin filament organization	0.006492
			actomyosin structure organization	0.007398
M6	8	2	epithelial cell differentiation	0.018536
			epithelium development	0.02915

Figure 9 Functional interpretations of the clustered signature using network analysis.
The significantly upregulated proteins in CD44-knockdown cells were subjected to the HumanBase for module enrichment analysis. The color of the node's border indicates each module. The colors of the inner nodes correlate with the $\log_2$ fold-changes of quantified proteins, and the sizes of nodes correspond to the p-values of each protein. (A) MDA-MB-231 cells. (B) Hs 578T cells.



	Cluster genes	Cluster terms	Top terms	q-value
M1	100	195	chromosome segregation	3.75E-31
			nuclear chromosome segregation	3.11E-24
			mitotic nuclear division	5.15E-24
M2	61	141	nuclear division	2.82E-08
			organelle fission	8.27E-08
			mitotic sister chromatid segregation	1.16E-07
M3	37	176	epithelial cell migration	9.26E-05
			tissue migration	9.26E-05
			ameboidal-type cell migration	0.000152
M4	140	173	translational initiation	0.000198
			chromatin assembly or disassembly	0.000241
			mRNA metabolic process	0.000368
M5	48	34	mRNA processing	0.005429
			mRNA splicing, via spliceosome	0.015944
			cellular response to dsRNA	0.020697



	Cluster genes	Cluster terms	Top terms	q-value
M1	71	20	chromosome segregation	1.78E-23
			mitotic nuclear division	7.29E-21
			nuclear division	1.11E-20
M2	66	18	spindle assembly	1.57E-08
			nuclear chromosome segregation	1.60E-08
			spindle organization	3.02E-08
M3	23	34	ribosomal small subunit biogenesis	1.54E-05
			translation	0.000120
			peptide biosynthetic process	0.000141
M4	63	113	mRNA metabolic process	0.000141
			mRNA processing	0.000493
			regulation of mRNA metabolic process	0.000672
M5	5	15	regulation of wound healing	0.000930
			regulation of response to wounding	0.000970
			cell migration	0.001218
M6	12	2	cell cycle phase transition	0.022669
			mRNA metabolic process	0.032380

Figure 10 Functional interpretations of the clustered signature using network analysis.
The significantly downregulated proteins in CD44-knockdown cells were subjected to the HumanBase for module enrichment analysis. The color of the node's border indicates each module. The colors of the inner nodes correlate with the log2 fold-changes of quantified proteins, and the sizes of nodes correspond to p-values of each protein. (A) MDA-MB-231 cells. (B) Hs 578T cells.

5. CD44 Regulates Cell Growth and Migration

To confirm the impact of CD44 knockdown in claudin-low breast cancer cells, cell viability, proliferation, and migration assays were applied. First, cell viability during the log phase of growth following 48 h was determined by MTS assays. In both cell lines, a significant decrease in cell viability was observed when compared to that in the negative control sample (Figure 11B). Cells treated with CD44 siRNA exhibited inhibition of cell viability in MDA-MB231 (p-value = 0.0177) and Hs 578T (p-value = 0.0015) cells, respectively. Next, we examined the effect of CD44 knockdown on cell proliferation using BrdU-based cell proliferation assays. As shown in Figure 11C, the proliferation rate was significantly reduced by CD44 knockdown in MDA-MB-231 (p-value = 0.0038) and Hs 578T (p-value = 0.0005) cells, respectively, in agreement with the results of the MTS assays. These results showed that a reduction in CD44 expression inhibits cell proliferation.

The migration ability of the two breast cancer cell lines was also evaluated using the transwell migration assay. The assay revealed that the knockdown of CD44 significantly suppressed the migration of MDA-MB-231 and Hs 578T cells, as compared to that in control cells (Figure 11D). In MDA-MB-231 cells, the knockdown of CD44 decreased the migration by approximately 2.8-fold (p-value = 0.0003). Similarly, with Hs 578T cells, this diminished migration by approximately 3.3-fold (p-value = 0.0012). These results confirm our in vitro observations that CD44 promotes proliferation and migration in claudin-low breast cancer cell lines.

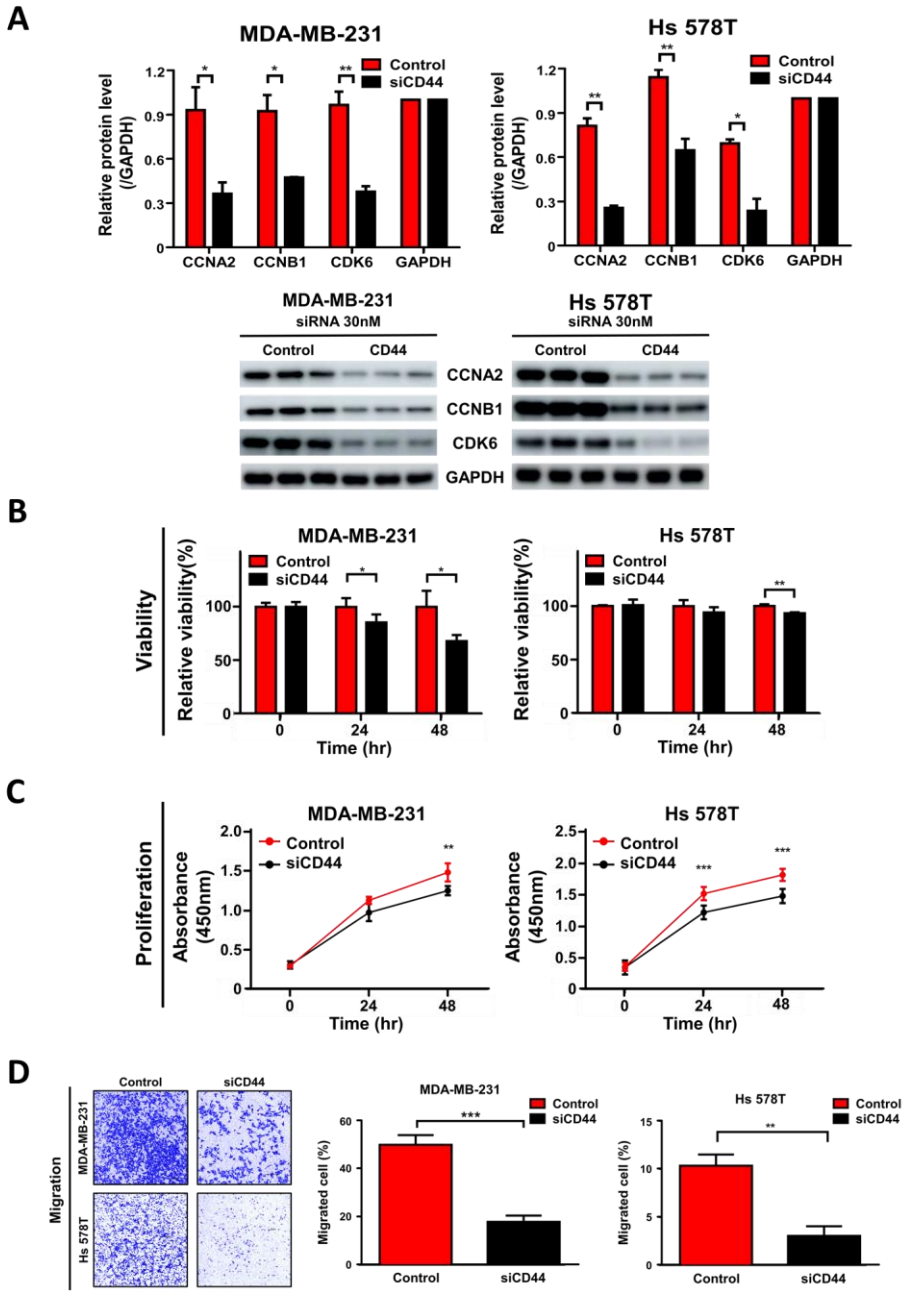


Figure 11 Effects of CD44 knockdown on two claudin-low breast cancer cell lines.
(A) Western blot analysis of cell cycle-related significantly expressed (B) Results of cell viability assays. (C) Results of cell proliferation assays. (D) Results of cell migration assays.

Discussion

Our proteomic and in vitro data support the notion that CD44 knockdown might cause changes in the genetic process leading to a decrease in cell proliferation. Moreover, downregulated CD44 expression levels could enhance cell adhesion and metabolic change, which suppresses cancer cell migration. Interestingly, it is widely accepted that many cell adhesion molecules function as tumor suppressors [67]. Cell adhesion molecules allow cells to communicate with one another or to the extracellular environment by mediating cell–cell or cell–extracellular matrix (ECM) interactions [68]. The resulting loss of cell–cell or cell–ECM adhesion promotes cell growth as well as tumor dissemination. Therefore, the cell adhesion molecules that function as tumor suppressors are also involved in limiting tumor cell migration [69]. In addition, mitochondria are important regulators of the metabolic plasticity of cancer cells, and mitochondrial metabolic reprogramming has a critical role in cancer growth, stemness, and therapy resistance [70]. In particular, the reprogramming of mitochondria affects both the phenotype of cancer cells and resistance to chemotherapy [71]. However, the role of mitochondrial metabolism in breast cancer cell migration is controversial. Some evidence reported a direct correlation between OXPHOS and cancer cell metastasis [72]. Interestingly, Lunetti et al. suggested that luminal-like and basal-like breast cancer cells display a distinct bioenergetic and metabolic phenotype, and consequently exhibit altered dependency on specific metabolic pathways [73]. They also confirmed that the basal-like and highly metastatic MDA-MB-231 cell line exhibits a higher glycolytic flux and lower mitochondrial respiratory rate compared to the luminal-like MCF-7 cell line. The higher

glycolytic rate and the reduction in OXPHOS metabolism observed in MDA-MB-231 cells established their metastatic and invasive properties. In our study, we observed that CD44 knockdown leads to metabolic shift from glycolysis to OXPHOS in MDA-MB-231 and Hs 578T, suggesting that the metabolic change could contribute to the decrease of breast cancer cell invasive capability.

For functional validation of proliferation, BrdU cell proliferation assay detects BrdU incorporation into cellular DNA during cell proliferation. The BrdU-labeled DNA has to be denatured to be detected by the BrdU antibody. The magnitude of the absorbance for the developed color is proportional to the quantity of BrdU incorporated into DNA, thus indicating cells that were actively replicating their DNA. The result of assay showed that CD44 knockdown significantly inhibited proliferation of MDA-MB-231 and Hs 578T cells, suggesting the CD44 silenced cells transit the cell cycle more slowly. In addition, viability assay was substituted with the results of MTS assay. Because MTS assay protocol is based on the reduction of the MTS tetrazolium compound by viable cells into a formazan product, it is more suitable for viability assay than proliferation assay. In both cell lines, a significant decrease in viability was observed when compared to that in the negative control sample.

The PPI network map in each module using StringDB showed that CD44 interacts with MET (hepatocyte growth factor receptor) in both cell lines (Figure 12). MET is a receptor tyrosine kinase that activates oncogenic signals from the ECM into the cytoplasm by binding ligands [74]. Many studies have elucidated that HGF/MET signaling contributes to the migratory and invasive phenotype of breast cancer [75]. Previous studies have proven that CD44 regulates MET and activates downstream signaling pathways in colon [76, 77], prostate [77], and pancreatic

cancer cells [76]. Moreover, specific CD44v isoforms were previously shown to interact directly with MET in only HGF-dependent mode [76, 78]. Interestingly, Orian-Rousseau et al. [76] suggested that this interaction between CD44 and MET is not found in cells expressing only CD44s, even though the cells expressed at least the same amount of MET. Actually, we observed no direct interaction between CD44s and MET in MDA-MB-231 and Hs 578T cells, which exclusively express CD44s (data not shown). Therefore, further research is needed to decipher which of these mediates the CD44–MET association in claudin-low breast cancer cells. Nevertheless, we detected a significant decrease in MET expression induced by CD44 knockdown (Figure 4), suggesting that CD44 induces cell migration by regulating MET with direct or indirect interactions occurring in claudin-low breast cancer cells. Taken together, this network-based functional clustering analysis predicted that CD44 probably contributes to the regulation of cell proliferation and migration.

In this study, we present a valuable TMT-based proteomic resource representing the response to CD44 regulation in claudin-low breast cancer cell lines. Importantly, it is clear that CD44 contributes to the complex molecular mechanisms of claudin-low breast cancer progression, which involves intricate networks. The results could lead to a better understanding of the biological functions of CD44 at the molecular level. These findings provide the framework for future proteomic investigations and also suggest that CD44 is a potential therapeutic target for treatment of the claudin-low breast cancer (Figure 13).

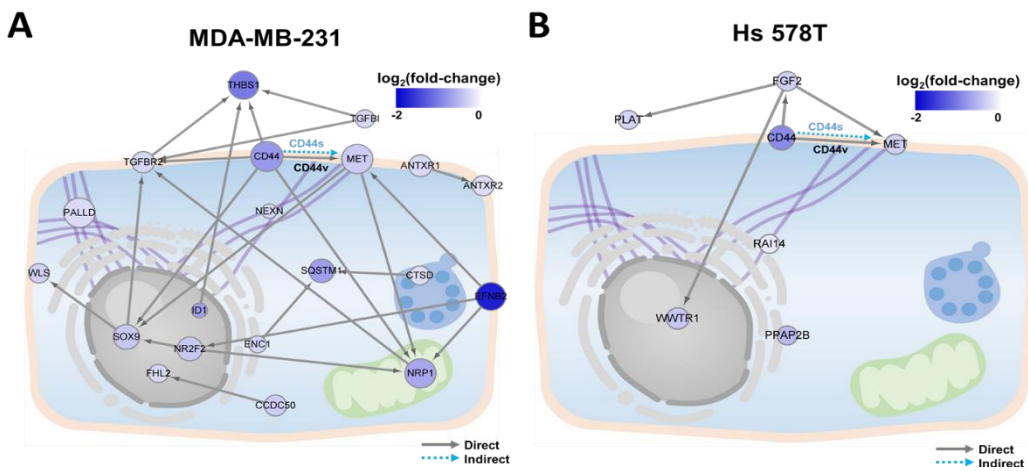


Figure 12 Protein–protein interaction subnetwork
from the HumanBase cell migration-specific module in (A) MDA-MB-231 and (b) Hs 578T cells. The node colors indicate expression levels of the significantly downregulated proteins in CD44-knockdown condition. Edges drawn between nodes correlate with connectivity.

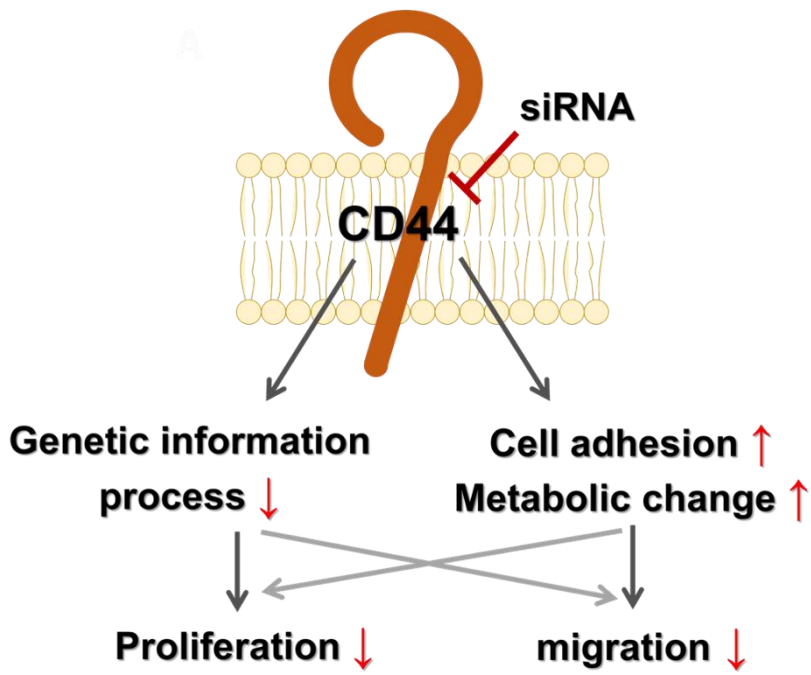


Figure 13 Schematic overview of proposed biological functions of CD44 in claudin-low breast cancer cell lines.

Chapter II

In-depth proteome profiling of breast cancer formalin-fixed paraffin-embedded tissue for distant metastasis

Abstract

In recent, improvements in the early detection of breast cancer makes for a decline in the mortality rate. However, the major-cancer related deaths from breast cancer are not due to the primary tumor, yet the result of distant metastasis to other organs. Importantly, 10-15 % of breast cancer patients develop distant metastasis within 2 years (early distant metastasis) of diagnosis with a poor 5-yr survival rate. Moreover, half of the distant metastasis patients experience late-onset occurring more than 5-yrs after diagnosis. Until now, the underlying mechanisms of distant-metastasis in breast cancer are currently not well understood. In this study, we aimed to develop a novel prognostic score for breast cancer distant metastasis patients and provide insights into the molecular mechanism leading to distant metastasis.

We present a study of formalin-fixed paraffin-embedded (FFPE) tissue specimens using a novel in-depth quantitative proteomic strategy. In this study, we analyzed 9 early-, 9 late-, and 10 no- distant metastasis breast primary tumor FFPE tissues. To achieve an in-depth proteome in the minimum of FFPE slides per sample, we performed well-defined proteomic strategies including isolation of cancerous region via examination by pathologists, protein extraction, filter-aided sample preparation, high-pH peptide fractionation based on stage-tip, and high-resolution quadruple Orbitrap LC-MS/MS. Label-free quantification and data analysis were performed using Maxquant and Perseus software.

We identified a total of 9,455 protein groups using FFPE slides at 1% of the peptide and protein FDR level. For quantitative analysis, proteins quantified in at least 70% of samples in each group were filtered, resulting in expression profiles for 7671 proteins. The proteome revealed a separation of early-, late-, no- distant metastasis

based on proteomic profiles and four clusters of proteins displayed. Also, we focused on proteins significantly expressed in early- and late- distant metastasis compared to no- metastasis. The pair-wise comparison showed that 832 proteins and 548 proteins were significantly expressed in set 1 (early- versus no- distant metastasis) and set 2 (late- versus non-metastasis), respectively. Functional annotation analyses were performed to establish signaling pathways with significantly expressed proteins. In the case of early-distant metastasis PI3K-Akt-mTOR pathway and PDGFR-beta pathway were enriched. For late-distant metastasis, Ras pathway and TGF-beta pathway were associated.

Proteomic profiling allowed us to define ten patterns of protein level changes, four of which correlate to the status of distant metastasis. Weighted gene correlation network analysis revealed highly significant proteins are associated with distant metastasis. Signature proteins were validated by survival analysis based on external datasets including METABRIC.

Our study presents the largest proteome profile of early and late distant metastasis of breast cancer using FFPE tissue specimens. Our analyses reveal a stronger association between proteomic profiles and distant metastasis and identify unique protein-based classification. Also, our platform can easily be implemented in other types of cancer to analyze large numbers of pathologically relevant proteins in clinical specimens including FFPE tissues.

Introduction

Breast Cancer is a frequent diagnosed cancer and the leading cause of female cancer-related death [79]. The survival rates of breast cancer are very high; however, metastatic relapse is a major cause of death. Distant metastasis is the dissemination of tumor cells from a primary tumor to distant organs with sequential events. Metastatic relapse can occur months to decades after initial diagnosis. The introduction of advances in screening and treatments such as chemotherapy and radiation has reduced metastatic recurrence and improved survival of patients with relapse [80]. Nevertheless, 20-30% of early breast cancer patients still die because of distant metastasis [80]. Therefore, a definition for a patient's risk of distant metastasis and a better understanding for the biological process is needed.

Formalin fixation paraffin embedding (FFPE) has been used for common method to preserve tissue and diagnose pathogenesis in clinical field. FFPE tissues can be stored at room temperature for decades and intrinsically linked to clinical information. Hence, FFPE is a valuable source for clinical research including DNA, RNA and protein measurements. In a previous MS-based proteomics study, lysine methylation (+14 Da) and slight increase in methylene (+12Da) and methylol (+30Da) were detected compare to fresh frozen tissues [81]. However, the modifications due to formalin fixation are minor, 2-6% of the overall peptide lists. Efficient proteomics techniques [82-84] have been developed for the reversal of crosslinks and protein extraction from FFPE tissues.

System-level analysis, co-expression network analysis has been used as essential method for identifying key molecular processes or potential key drivers that play

significant roles in disease. The weighted co-expression network analysis (WGCNA) assesses co-expression of proteins and relating co-expression modules into clinically meaningful modules. In several previous studies[85, 86], proteome analysis using the WGCNA algorithm elucidate the connection of proteins modules and clinical phenotypes.

Here, we present the first distant metastatic breast cancer proteome using FFPE tissues that can predict distant metastasis interval after initial diagnosis. We have assembled a cohort of 28 breast cancer patients of varying distant metastasis intervals and performed MS-based proteomic analysis to clarify molecular features associated with distant metastasis. Furthermore, we compare early and late distant metastasis so that we can find proteins that can act as key metastasis drivers. Moreover, we compare our proteomic data to published mRNA data of breast primary tumor. We found that some candidates tend to have a high correlation of protein and mRNA levels. Altogether, these results highlight the potential distant metastasis protein indicator in breast cancer and identify biological processes contributing to distant metastasis progression.

Materials and Methods

1. Patient selection for proteomic discovery analysis.

Formalin-fixed paraffin-embedded (FFPE) tissues used in this study were collected from the Seoul National University Hospital biorepository operated by the department of pathology. The study was approved by the Institutional Review Board of Seoul National University Hospital (Approval No. 1612-011-811).

2. Protein Extraction and digestion.

FFPE tissues were sectioned 5 or10-μm using a microtome. The 5-μm slide of each sample was stained with hematoxylin and eosin and mounted under a glass coverslip. Using the H&E slides, only tumor tissues were selectively displayed through visual analysis by pathologists.

The unstained 10-μm slides generated from identical FFPE block were prepared without a coverslip. The FFPE slides were deparaffination xylene for 5min and 3min. Followed by rehydration with 100% ethanol, 50% ethanol for 1.5min and LC-MS grade water for 1min. The pre-displayed cancerous cellular area was scrapped using a scalpel and transferred to a microcentrifuge tube. The extracted tissues were lysed in 4% SDS, 2 mM TCEP, and 0.1 M Tris-HCl, pH 7.5 buffer followed by direct sonication (10%, 5 cycles, 5 s, 2 s). Lysates were heated for 2hr at 95 °C. After 10min centrifugation (15,000rpm, 21℃), the supernatant was transferred to a new tube and the concentration of protein was measured using a reducing agent-compatible BCA assay (Thermo Fisher Scientific, Waltham, MA). Then, each sample containing 200 μg of total protein was precipitated with cold acetone overnight.

The pellet was resuspended in denaturation buffer containing 2% sodium dodecyl sulfate (SDS), 10 mM tris(2-carboxyethyl) phosphine (TCEP), 50 mM

chloroacetamide (CAA), and 0.1 M tris-HCl, pH 8.5 and heated for 15 min at 95 °C. 300μL of UA buffer (8 M UREA in 0.1 M Tris-Cl, pH 8.5) were added to the sample and the mixtures were loaded onto a 30-kDa spin filter (Merck Millipore, Darmstadt, Germany). By 15min centrifugation (14,000rcf, 21℃), the buffer was exchanged 3 times with UA buffer and 40mM Ammonium bicarbonate, respectively. The proteins were digested by multienzyme. The first protein digestion was performed overnight at 37℃ by adding trypsin/LysC (Promega, Madison, WI; protein to protease ratio = 100:1). On the following day, the peptides were collected by centrifugation and remaining proteins in the filter were digested for 3h at 37℃ with sequencing grade-modified trypsin (protein to protease ratio = 200:1). After second digestion, resulting peptides were collected together and the peptide concentrations were measured by fluorescence spectrometry. For 96 well plate-based tryptophan assay, the excitation wavelength was set to 295nm and the emission wavelength was set to 350nm.

3. Peptide desalting and peptide fractionation based on C18 StageTips.

The 20ug of eluted peptides were acidified with trifluoroacetic acid (TFA). Peptides were then desalted using StageTip-C18 as follows. Desalted peptides were dried in a vacuum. Dried peptides were resolved with loading solution (15mM ammonium hydroxide and 2% acetonitrile) and separated with StageTip based microcolumn prepared as described in the protocol (DOI: 10.1039/C9AY01269A). Different 20 elution buffers were used to fractionation and pooled into 6 fractions. The fractionated peptides were lyophilized for LC-MS analysis.

4. Mass spectrometry and Database Search

Mass spectrometry-based proteomic analysis was performed using a Q Exactive Plus Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific Inc.), coupled to an Ultimate 3000 RSLC system (Dionex, Sunnyvale, CA, USA) via a

nano electrospray source. Peptide samples were separated on a two-column system, consisting of a trap column and an analytic column (75 μm × 50 cm) with a 120 min gradient from 7% to 32% acetonitrile at 300 nl/min and analyzed by mass spectrometry. Survey scans (350 to 1650 m/z) were acquired with a resolution of 70,000 at m/z 200. MS/MS spectra were acquired at an HCD-normalized collision energy of 30 with a resolution of 17,500, at m/z 200. The maximum ion injection times for the full scan and MS/MS scan were 20 and 100 ms, respectively.

Raw MS/MS files were processed with MaxQuant (version 1.6.1.0) using the Andromeda search engine against the Human Uniprot protein sequence database (December_2014, 88 657 entries). Primary searches were performed using the MS/MS ion tolerance and set to 20 ppm. Cysteine carbamidomethylation N-acetylation of protein and oxidation of methionine were set as fixed and variable modifications, respectively. Enzyme specificity was set to full tryptic digestion. Peptides with a minimum length of six amino acids and up to two missed cleavages were considered. The required false discovery rate (FDR) was set to 1% at the peptide, protein, and modification level. We enabled the 'Match between Runs' option on the MaxQuant platform to maximize the number of quantification events across samples.

5. Statistical Analysis

For statistical analysis of proteomic data, Perseus software (version 1.6.14.0) was used. Based on the logarithmic (log_2) intensity of the intensity-based absolute quantification (iBAQ) data, proteins quantified in at least 70% of samples in each group were filtered and missing values were imputed based on normal distribution separated for each column (width=0.3 and downshift=1.8) to represent low expressed proteins. Using quantile normalization, data were normalized. Pairwise

significantly altered proteins were identified using Student's t-test, followed by permutation-based FDR correction. Venn diagrams were plotted using Biovenn web tool[87].

6. Survival Analysis

Survival analysis was performed using the R environment. We collected the two public breast cancer datasets containing distant metastasis-related information with transcriptomic data and clinical data. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) data were downloaded through the *cgdsr* R package (version 1.3.0). We used the distant metastasis-free survival to assess the prognostic value of gene expression, divided into binary (high- and low- expression) groups using Kaplan-Meier survival curves through the *survminer* R package (version 0.4.9). For differences between survival curves, we applied the criteria of log-rank test p-value less than 0.05. Also, we used Cox proportional hazards regression model with distant metastasis-free survival to estimate the hazard ratio (HR).

7. Gene Ontology (GO) Analysis

To identify the function of proteins significantly altered by distant metastasis status, we performed gene set over-representation analysis using ConsensusPathDB. Gene sets were defined by all GO terms and pathway-based sets in the DB.

8. Weighted Gene Correlation Network Analysis (WGCNA)

A weighted protein co-expression network was derived from protein abundance values through the *WGCNA* R package (version 1.70-3) [88]. The WGCNA was used step-by-step network construction and module detection with the following parameters: soft threshold power 7, deepSplit 2, minModuleSize = 160. The remaining parameters used a default values. MEs were correlated with different

biological traits indicating distant metastasis status. Multiple comparisons were accounted for by FDR correction across modules, and the *p* values and correlation efficient for the model are reported.

9. Selection of Candidate Prognostic Biomarkers

Protein marker selection of quantitative proteomic data was conducted using the *geNetClassifier* R package (version 1.26.0). Based on multi-class support vector machine (SVM) classification, proteins with a posterior probability greater than 0.95 showed significant difference across the groups. Each protein can only be on the ranking of one group. Ranked proteins were returned with parameters (exprsMeanDiff (Mean difference of expression values), discriminant Power, discrPwClass) calculated for gene selection.

Results

1. Clinical Characteristics of The Study Subjects

To identify the proteomic alterations that reflect distant metastasis status in primary breast cancer, we collected primary breast FFPE tissues from 28 individuals. Depending on the interval of the primary breast cancer diagnosis to the occurrence of distant metastasis, it was divided into the non-metastasis (NM) group that did not occur for more than five years, the early-metastasis (EM) group that occurred within two years, and the late-metastasis (LM) group that occurred after five years. Clinical characteristics are outlined in Table 1. All the patients were diagnosed with infiltering duct carcinoma and similar tumor gross appearance with widely applied tumor-node-metastasis (TNM) stage IIA (AJCC 8^{th}) system. There were no differences in demographic characteristics including T, N stage, nuclear grade, pathological prognostic group, HER2, ER, PR status and molecular subtype across the group. The Chi-square test is used to compare the clinical characteristics among the groups.

2. Workflow of distant metastasis prognosis marker candidate discovery

We identified proteins that expression patterns correlated with distant metastasis status, as shown in Figure 1. To identify altered protein when distant metastasis occurred early and late after diagnosed primary tumor, the following analysis were conducted. (i) A list of common proteins representing significant difference was selected by supervised analysis. (ii) With unsupervised powerful systems biology methods, WGCNA, we found co-expressed protein modules correlated with metastasis status and filtered candidate proteins by key proteins. (i) and (ii) suggested late metastasis and no metastasis markers. To find early metastasis specific markers,

(iii) machine learning-based feature selection was conducted. Top-ranked early metastasis specific proteins were filtered by (ii) methods. Consequently, (iv) resulted protein marker candidates conducted survival analysis with open source mRNA expression data.

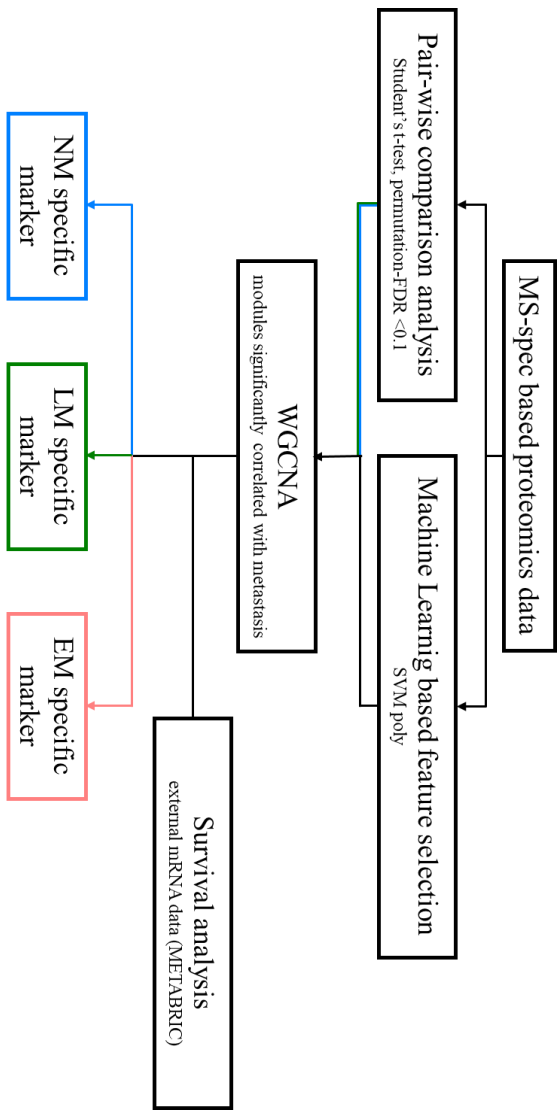


Figure 1 Workflow of discovery distant metastasis related candidate protein markers

Feature	average			
	NM	LM	EM	*p*-value
Age	46.0	46.1	47.6	0.946
Feature	**Count (percentage%)**			
	NM	**LM**	**EM**	**test statistics**
T stage				0.338
T1c	1 (3.6%)	3 (10.7%)	1 (3.6%)	
T2	9 (32.1%)	6 (21.4%)	8 (28.6%)	
N stage				0.335
N0	10 (35.7%)	8 (28.6%)	9 (32.1%)	
pN1mi	0 (0%)	1 (3.6%)	0 (0%)	
Nuclear grade				0.188
1/3	0 (0%)	1 (3.6%)	0 (0%)	
2/3	5 (17.9%)	4 (14.3%)	1 (3.6%)	
3/3	5 (17.9%)	4 (14.3%)	8 (28.6%)	
Stage group (AJCC 8th)				0.543
IA	1 (3.6%)	2 (7.1%)	1 (3.6%)	
IB	0 (0%)	1 (3.6%)	0 (0%)	
IIA	9 (32.1%)	6 (21.4%)	8 (28.6%)	
Pathological prognostic group				0.746
IA	5 (17.9%)	4 (14.3%)	2 (7.1%)	
IB	2 (7.1%)	2 (7.1%)	2 (7.1%)	
IIA	3 (10.7%)	3 (10.7%)	5 (17.9%)	
HER2 status				0.136
Negative	4 (14.3%)	7 (25.0%)	7 (25.0%)	
Positive	6 (21.4%)	2 (7.1%)	2 (7.1%)	
ER status				0.619
Negative	4 (14.3%)	3 (10.7%)	5 (17.9%)	
Positive	6 (21.4%)	6 (21.4%)	4 (14.3%)	
PR status				0.470
Negative	4 (14.3%)	4 (14.3%)	6 (21.4%)	
Positive	6 (21.4%)	5 (17.9%)	3 (10.7%)	
molecular subtype				0.215
HER2	1 (3.6%)	0 (0%)	0 (0%)	
LUMINAL	8 (28.6%)	8 (28.6%)	5 (17.9%)	
TNBC	1 (3.6%)	1 (3.6%)	4 (14.3%)	
Distant Metastasis organ				
Bone	0 (0%)	3 (10.7%)	3 (10.7%)	
Chest wall	0 (0%)	2 (7.1%)	0 (0%)	
Liver	0 (0%)	2 (7.1%)	2 (7.1%)	
Lung	0 (0%)	2 (7.1%)	3 (10.7%)	
Pleura	0 (0%)	0 (0%)	1 (3.6%)	
Skin	0 (0%)	0 (0%)	1(3.6%)	

Table 1 Comparison of Clinical Characteristics

NM, no metastasis. LM, late metastasis. EM, early metastasis.
TNBC, triple negative breast cancer.

3. Global Proteomic Analysis of Primary Breast Cancer FFPE Tissues

To identify proteomic alterations that can affect distant metastasis, we performed LC-MS/MS based proteomic analysis on 10 NMs, 9 EM and 9 LM FFPE tissues. Briefly, the samples were deparaffinized with xylene, rehydrated in graded ethanol and the proteins were extracted from the tumorous region. The lysates were digested and the same amounts of peptides were desalted. To improve sensitivity the peptides were separated into 6 fractions using Stage-Tip based high-pH fractionation and analyzed on an orbitrap mass spectrometer. In this study, six fractionations for each sample were analyzed using 168 LC-MS/MS runs (Figure 2A).

At the 1% peptide and FDR level, total 9,455 proteins were identified (Figure 2B). We used the intensity based absolute quantification (iBAQ) value calculated from MaxQuant for the analysis. Across three groups, 23 (0.3%), 26 (0.3%), 16 (0.2%) unique proteins were quantified in NMs, EMs, and LMs, respectively, and 8247 (97.8%) proteins were common in all three groups (Figure 2C). The cumulative numbers of quantified proteins increased steadily in all three groups (Figure 1D). Also, the protein expression showed a globally homogenous pattern among the groups. The number of proteins quantified at least one sample in EM was 8373 (99% of total quantified proteins) and at least three samples in EM was 8148 (96.4%). The number of proteins quantified at least one sample in LM was 8317 (98.4% of total quantified proteins) and at least three samples in LM was 8148 (94.9%). The number of proteins quantified at least one sample in NM was 8356 (98.8% of total quantified proteins) and at least three samples in NM was 8107 (95.9%).

In subsequent analyses, we included 7,671 proteins with fewer than 30% missing values in at least one group to improve on the absence information. Using the ESTIMATE tool, tumor purity was estimated to investigate the effects of stromal cell

contamination. All samples showed consistent and high tumor purity (0.701-0.830, Figure 3A). There is no statistically significant difference among the three groups. Principle component analysis (PCA) showed no clear separation among the groups of samples, presenting low biological variation across the samples (Figure 3B). Also, Figure 2C showed that the value proteome Pearson's correlation was high, and the average correlation was 0.844 (0.780 – 0.893).

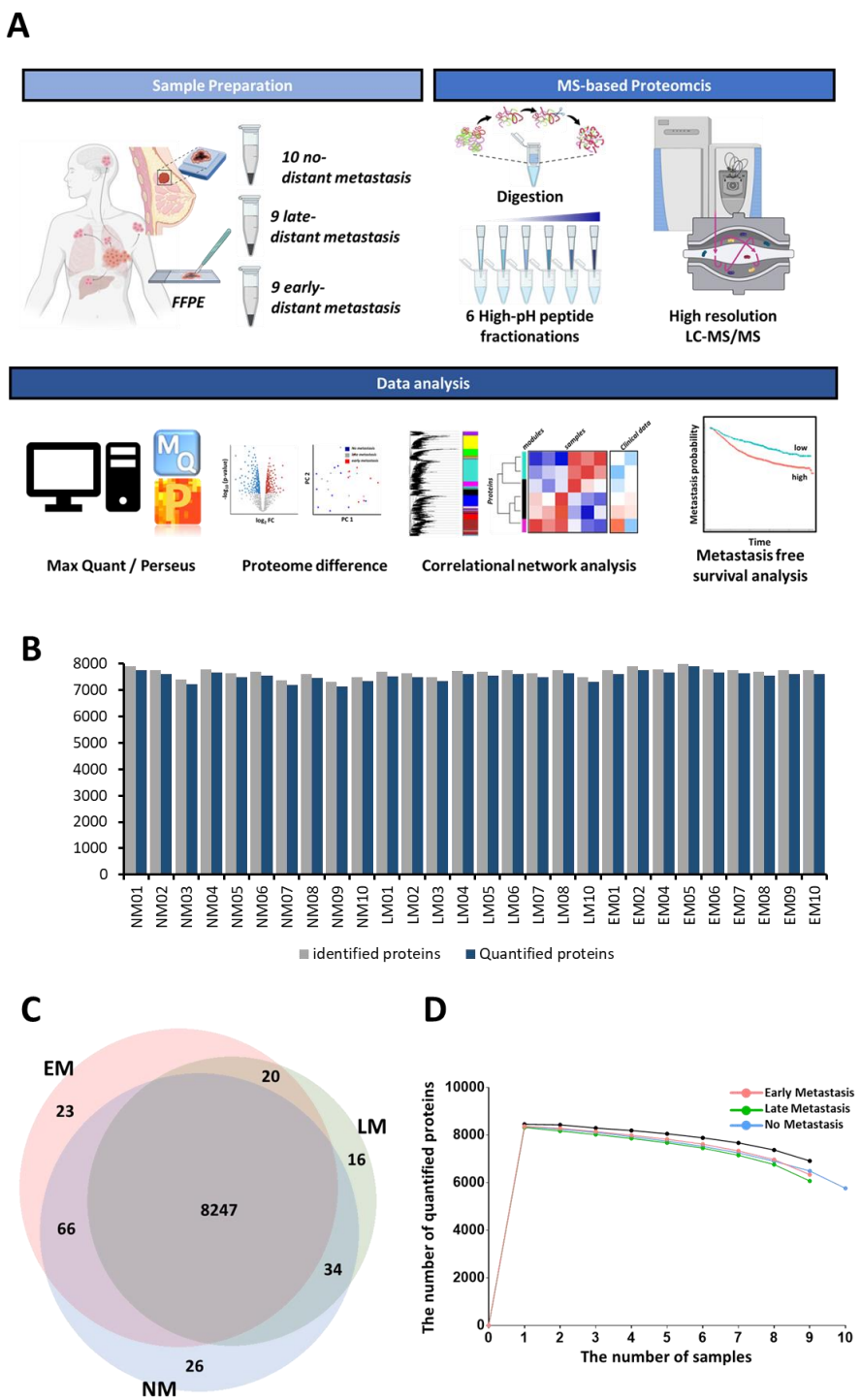


Figure 2 Deep proteome profiling (A) Overall workflow of FFPE primary breast cancer tissue proteomic analysis. (B) Bar graph of the number of identified proteins (gray) and quantified proteins (blue). (C) Venn diagram of quantified proteins per group. (D) the number of quantified proteins overlapping for each sample in the same group.

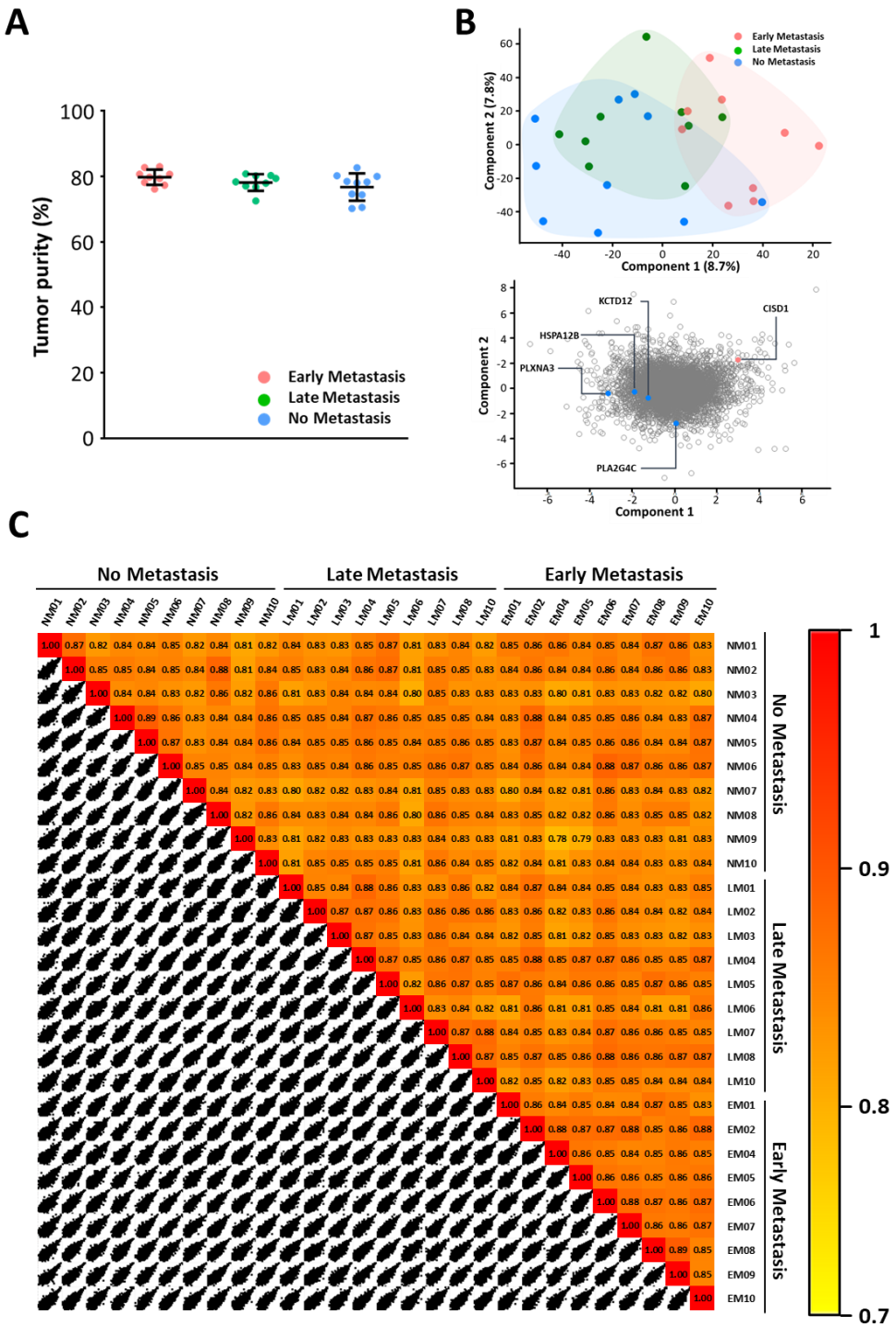


Figure 3 Biological variations among the samples
(A)Distribution of tumor purity scores. (B) Principal component analysis of protein quantified in each sample. (C) Reproducibility of proteomic analysis. The heatmap represents a correlation matrix of protein expression values across the all 28 samples.

4. Significantly Altered Proteins with Distant Metastasis Status

Prior to system-level analysis, we performed the statistical analysis to identify proteins and biological processes that can regulate distant metastasis in breast cancer. Differential expression was assessed using statistical t-test analysis (permutation-based FDR < 0.1; Figure 4), which identified proteins with dysregulated abundance levels according to the distant metastasis interval. There were 22 proteins with significantly increased abundance and 28 proteins were decreased with significantly in EM compared with NM. When LM was compared with NM, 4 proteins were significantly increased and 8 proteins were significantly decreased. In EM, 5 proteins expression were elevated; however, 2 proteins expression level was diminished in LM. Among these DEPs, KUTLG and VWA5A were decreased in EM, previously described breast cancer progression marker. With these altered proteins, GO enrichment analysis was performed. Differently expressed proteins among three groups were overrepresentation of “membrane proteins”, “envelop” and “secretory vesicle”.

To identify significant proteins that induced distant metastasis, we found proteins that demonstrated significant different expression patterns in all groups. As shown in Figure 5, the common DEPs were CISD1 and PLXNA3 in EM vs NM and LM vs NM, KCTD12 and HSPA12B in EM vs NM and EM vs LM, and PLA2G4C in LM vs NM and EM vs LM.

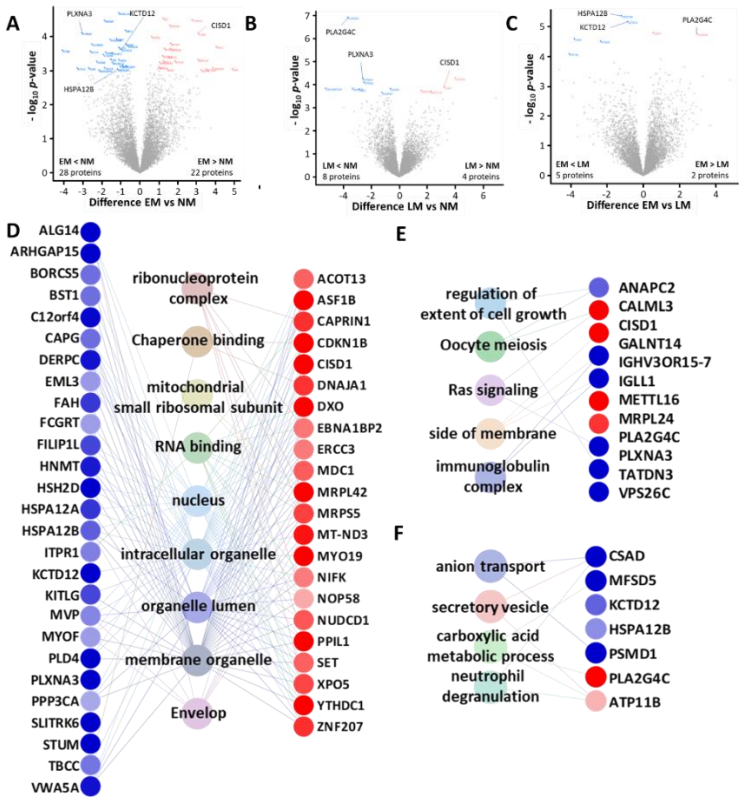


Figure 4 Supervised pair-wise comparison analysis.

Volcano plot represented significantly altered proteins in (A) EM vs NM, (B) LM vs NM, (C) LM vs EM. GO enrichment analysis of significant proteins in (D) EM vs NM, (E) LM vs NM, (F) LM vs EM.

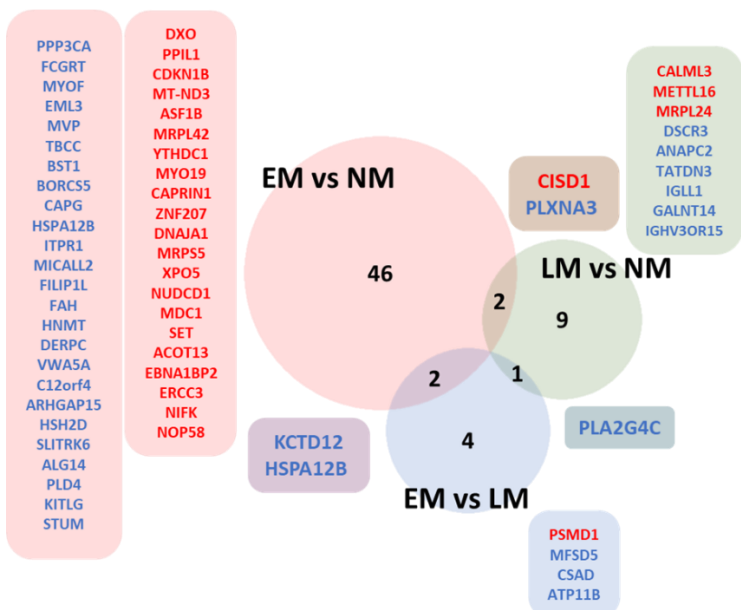


Figure 5 Venn diagram of significantly expressed proteins.

The red proteins were up regulated and blue proteins indicated downregulated in metastasis group.

5. Correlation of Distant Metastasis Status and Protein Networks

We subsequently performed a network analysis of the discovery using the weighted gene co-expression network analysis (WGCNA). This analysis assembles the dataset into protein modules with a similar expression pattern, a system biology approach, across the samples. The soft threshold value was 7, as it was the smallest threshold that resulted in a scale-free R2 fit of 0.9. The resulting network constructed 10 protein co-expression modules. The size of these modules was from largest (turquoise, n = 1714) to smallest (purple, n=286) (Figure 6A). We then calculated the correlation between each module and distant metastasis status (Figure 6B). Overall, there were 4 modules that showed correlations significantly ($p < 0.05$). The M2 (turquoise) was negatively correlated ($p = 0.00237$) with EM. The M3 (brown) module shown positive correlation ($p = 0.0002$) with the EM and negative correlation ($p = 0.0257$) with LM. The M5 (blue) modules presented positive correlation with EM ($p = 0.008$) and negative correlation with NM ($p = 0.005$). The M9 (black) modules were positively correlated ($p = 0.0001$) with NM and negatively correlated ($p = 0.0109$) with LM.

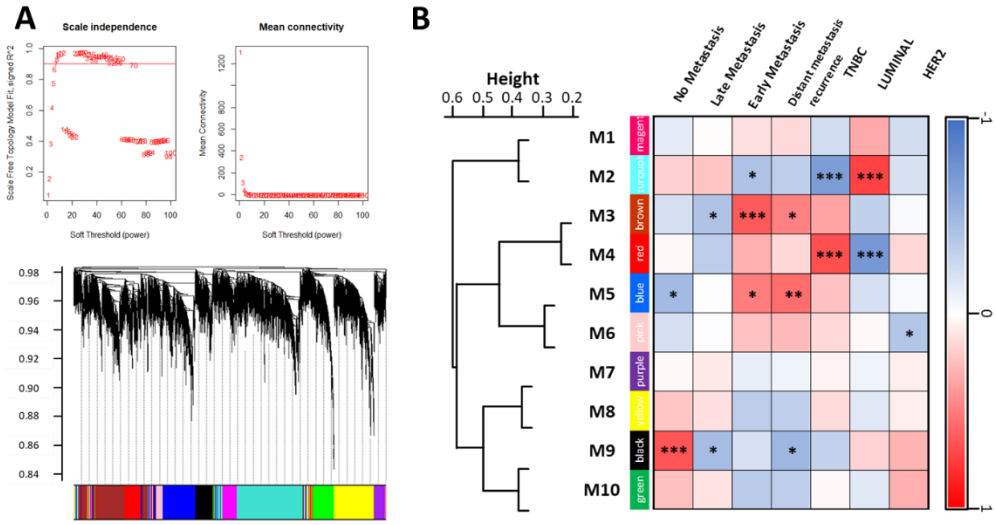


Figure 6 Construction of WGCNA identification and modules

(A) Scale independence of network topology for different soft-thresholding powers. Mean connectivity of network topology for different soft-thresholding powers. Clustering and module screening based on protein expression pattern. The top was gene dendrogram and the bottom was genes' modules with different colors. (B) Dendrogram of consensus module eigenproteins obtained by WGCNA on the consensus correlation. Relationships of consensus module eigneproteins and clinical traits. Intensity and the direction of correlations are indicated on the right side of the heatmap (red, positively correlated; blue, negatively correlated). $p* < 0.05$, $p** < 0.005$, $p*** < 0.05$

To generate lists of protein markers, we introduced a highly stringent cut-off combining co-expression analytical approaches and ANOVA statics. Proteins included in four significant modules and proteins with significant differences in three groups were compared. The number of common proteins was 216, 217, 188 and 105 in M2, M3, M5, and M9, respectively. With these proteins, we defined our proteomic analyzed samples using hierarchical clustering analysis. Interestingly, all groups (NM, LM, and EM) were demonstrated clear separation. Proteins were also shown 4 clusters and the expression patterns were identical with WGCNA color modules (Figure 7A). It was confirmed the previous selected distant metastasis-associated proteins, the common 5 proteins were overlapped with these proteins that were classified distant metastasis status well. As a result, the three proteins (CISD1, PLXNA3 and PLA2G4C) were identified (Figure 7B). To further confirm the

relationship between these proteins and distant metastasis, we conducted Kaplan-Meier analysis with distant metastasis-free survival (DMFS) analysis of the METABRIC data (Figure 7C). The metastasis free prognostic value of 3 proteins was evaluated by log-rank test. Kaplan-Meier (KM) curves were generated for the high expression and low expression groups based on the best cut-off. In mRNA level of the primary tumor, the Kaplan-Meier survival analysis showed that high expression of CISD1 ($p = 0.017$), and PLXNA3 ($p < 0.001$) were significantly associated with distant metastasis status. In contrast, low expression of PLA2G4C ($p < 0.0001$) was significantly associated with distant metastasis status. In addition, the univariate Cox hazard regression models indicate that high expression of CISD1 (HR: 1.23, 95% CI: 1.08 - 1.41, $p = 0.002$), and PLXNA3 (HR: 1.46, 95% CI: 1.28 – 1.68, $p < 0.001$) as well as low expression of PLA2G4C (HR: 0.78, 95% CI: 0.69 – 0.88, $p < 0.001$) were significantly relative with distant metastasis in breast cancer. As a result, we suggested CISD1 and PLXNA3 as no metastasis specific protein marker candidates and PLA2G4C as a late metastasis protein marker candidate.

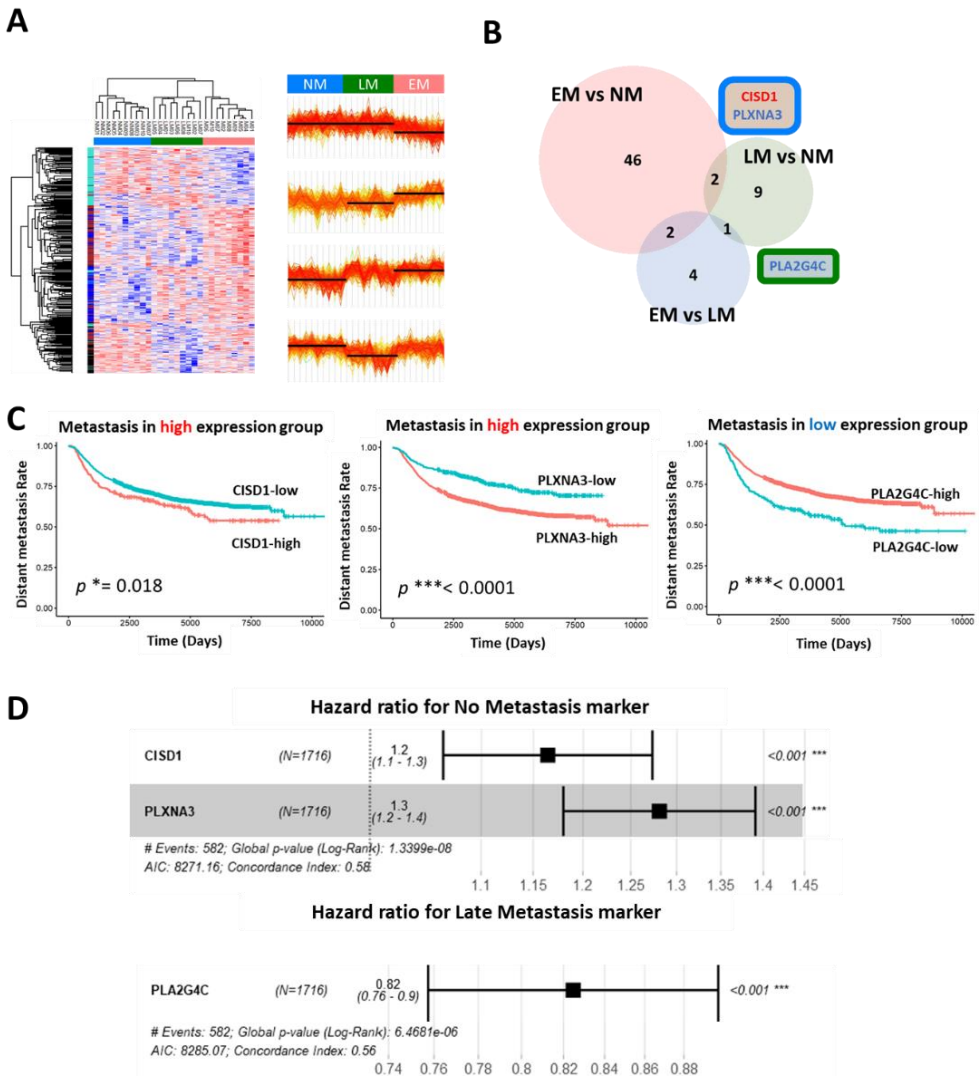


Figure 7 Selection for no metastasis and late metastasis specific protein marker candidates
(A) Non-supervised hierarchical clustering of z-scored normalized protein intensities. Proteins were enriched in significantly correlated with distant metastasis status and significant in ANOVA analysis (p-value < 0.05). (B) Venn diagram for suggesting metastasis specific protein marker. (C) Survival analysis. Kaplan-Meier curves of candidate proteins. (D) Forest plots of the Cox proportional regression adjusted HRs and the corresponding *p*-values of all patients.

6. Selection of candidate proteins through support vector machine analysis

In the above analysis, we could not select early metastasis specific candidate proteins. Therefore, we analyzed our protein data using the *geNetClassifier* package. Proteins were ranked with the greatest discrimination power. In total, 121 proteins exceeded the posterior

probability cut off value of 0.95 and were used in training the SVM. The top 10 ranked proteins that can discriminate early metastasis with the lowest error rate were in Figure 8A. As the strategy devised in the above analysis for finding no metastasis and late metastasis specific candidate proteins, we overlapped top-ranked 10 proteins and significant WGCNA module proteins. Interestingly, there is 7 common proteins, MYOZ1, PGR, LGI2, RASGRP1 were in M2, UBE2T was in M3, and IGKV2D-40 and CIAPIN1 were in M5 (Figure 8B).

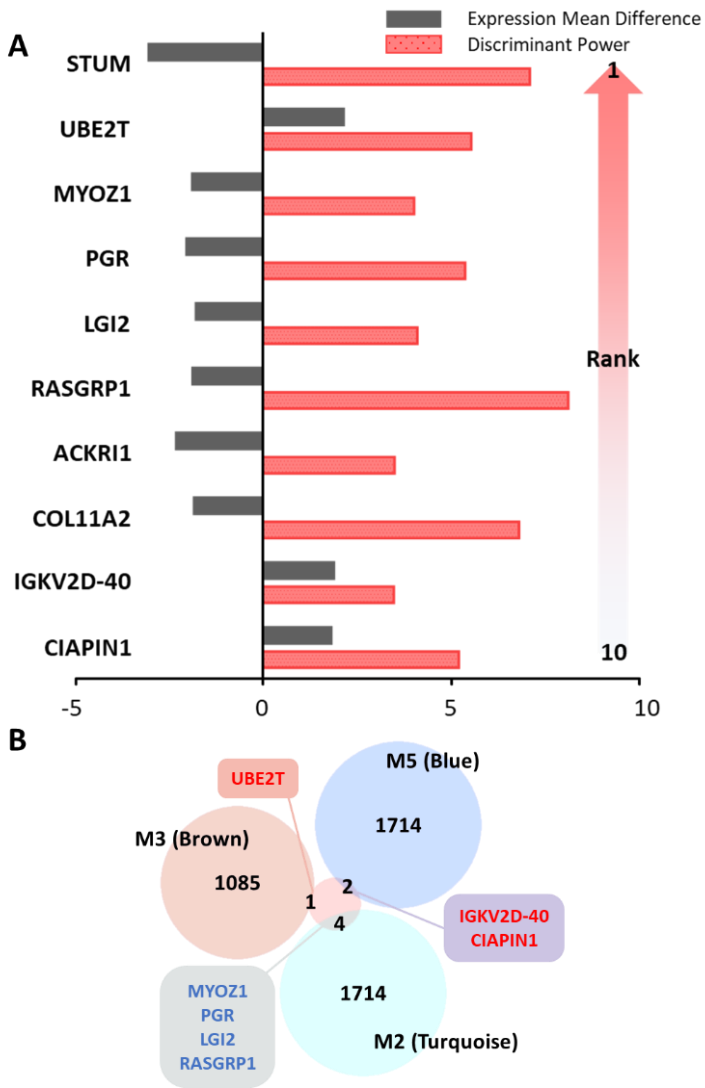


Figure 8 Selection of early distant metastasis prognostic candidate markers
(A) Discriminant Power and Expression Mean Difference of TOP ranked 10 proteins for classifying early metastasis. (B) Venn diagram of early metastasis specific candidate proteins and WGCNA module proteins.

To advance confirmation of the relationship between these proteins and distant metastasis, we also performed Kaplan-Meier analysis with DMFS analysis of the METABRIC data (Figure 9). Because there was no mRNA data of IGKV2D-40, we could not evaluate the potential as a candidate protein. KM curves were generated for the high expression and low expression groups based on the best cut-off. In the mRNA level of the primary tumor, the survival analysis revealed that high expression of UBE2T ($p < 0.0001$), and CIAPIN1 ($p < 0.001$) were significantly associated with distant metastasis status. In contrast, low expression of MYOZ1 ($p = 0.0051$), PGR ($p < 0.0001$), and RASGRP1 ($p = 0.013$) were significantly associated with distant metastasis status. In addition, the univariate Cox hazard regression models indicate that high expression of UBE2T (HR: 1.4, 95% CI: 1.3 - 1.5, $p < 0.001$), and CIAPIN1 (HR: 1.7, 95% CI: 1.4 – 2.1, $p < 0.001$) as well as low expression of PGR (HR: 0.74, 95% CI: 0.68 – 0.82, $p < 0.001$), and RASGRP1(HR: 0.83, 95% CI: 0.71-0.97, $p < 0.017$) were significantly relative with distant metastasis in breast cancer. Consequently, we suggested UBE2T, CIAPIN1, PGR and RASGRP1 as protein marker candidates that could classify early distant metastasis.

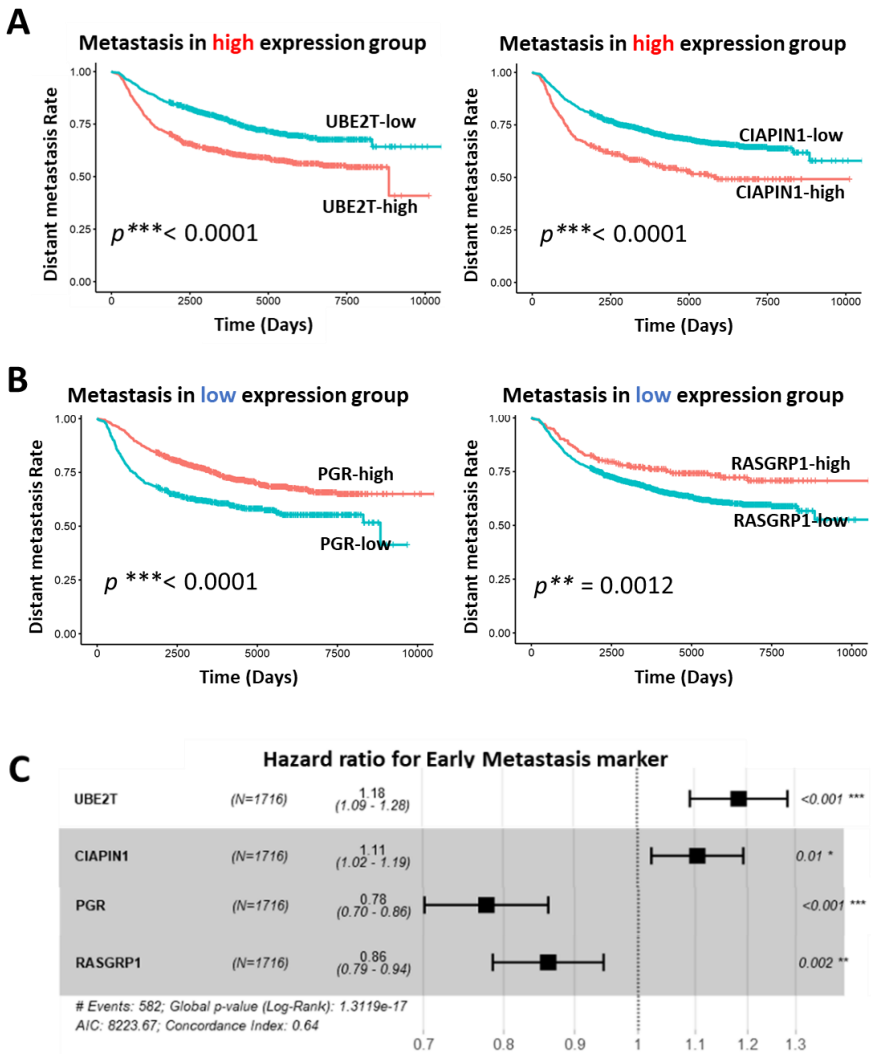


Figure 9 Survival analysis. Kaplan-Meier curves of candidate proteins.
(A) Survival analysis of proteins that distant metastasis occurred when the expression value is high.
(B) Survival analysis of proteins that distant metastasis occurred when the expression value is low.
(C) Forest plots of the Cox proportional regression adjusted HRs and the corresponding *p*-values of all patients.

7. Distant metastasis signature composed of CISD1, PLXNA3, PLA2G4C, PGR, RASGRP1, UBE2T, and CIAPIN1

Beyond the clinical significance of seven proteins in our dataset, we also validated its significance on external datasets. Evaluation of 1149 samples from the METABRIC data, we categorized samples that did not occur distant metastasis until 10 years as NM, occur distant metastasis after 5 years and occur distant metastasis within 2 years as EM from first breast primary tumor diagnosis. Student's t-test was conducted for pairwise comparison. An ANOVA test was performed to determine whether several groups were significantly different from each other. The results revealed that seven candidate proteins classified samples well into two groups, no metastasis versus metastasis, and four groups NM, LM, M and EM (Figure 10).

Eventually, we found seven distant metastasis-related genes with multivariate Cox regression analysis to build a predictive signature in the METABRIC dataset. The signature's concordance index was 0.65, and the Log rank *p*-value was 5.39083-23 (Figure 11A). Using the risk score formula from multivariate Cox regression, we calculated the risk score of each patient. We divided into two groups, high risk and low risk group according to the best cutoff point of the risk scores (1.008453). We illustrated seven distant metastasis related-molecular signature expression values in the formula related to the risk score using the heatmap (Figure 11B). The patients with lower risk group were owned better metastasis time (Figure 11C). For additional comparison with other breast cancer risk factors (molecular subtype, ER-status and grade), Kaplan-Meier survival analysis was conducted.

We observed a difference in outcome for tumors classified with distant metastasis related-molecular signature and other risk factors. distant metastasis related-molecular signature was associated with the worst outcome.

Collectively, these results indicate that differentially regulated protein networks exist in clinically relevant sample groups depending on distant metastasis status and that these protein networks impact both cancer biology as well as the abundance of potential biomarkers. While the evaluation of new biomarkers might generally be restricted to nucleic acid or protein analysis, a protein-wide evaluation of these targets is of considerable relevance. Furthermore, to better understand the molecular mechanisms relative to distant metastasis, integrated studies will be followed by functional assays that should be the method of choice to shed light on the fine regulation of key cancer protein products.

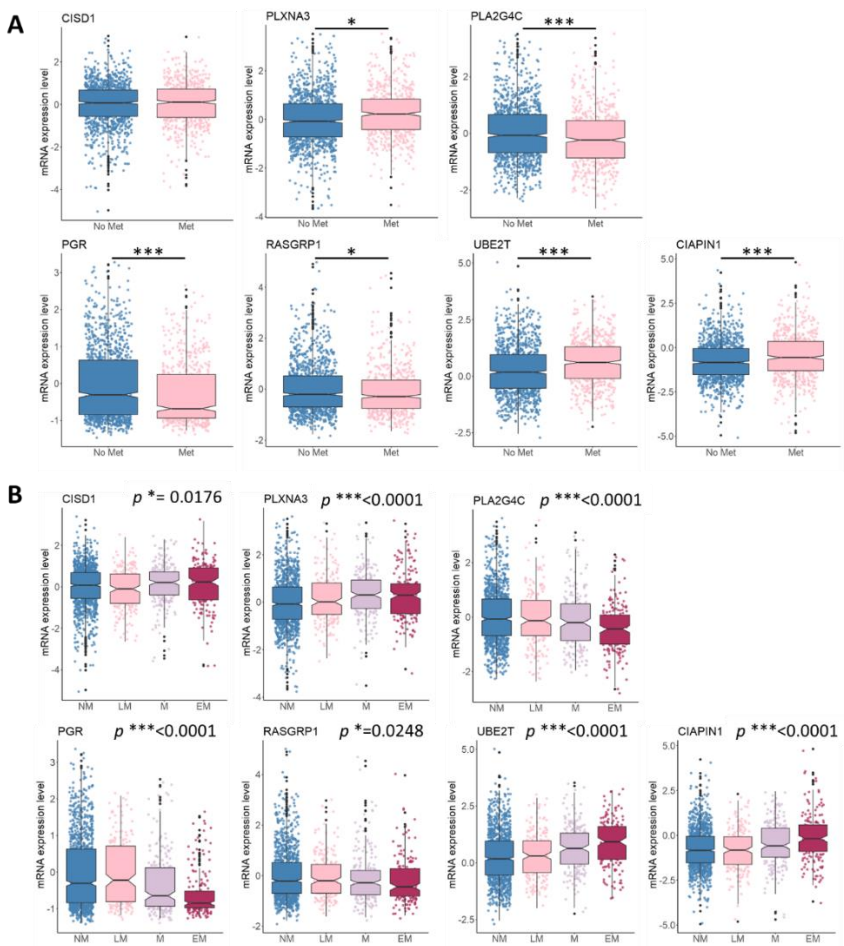


Figure 10 Pair-wise comparison analysis of distant metastasis relative signatures.
Differences of mRNA expression of METABRIC data. (*p*<0.05, p**<0.005, p***<0.0005)*

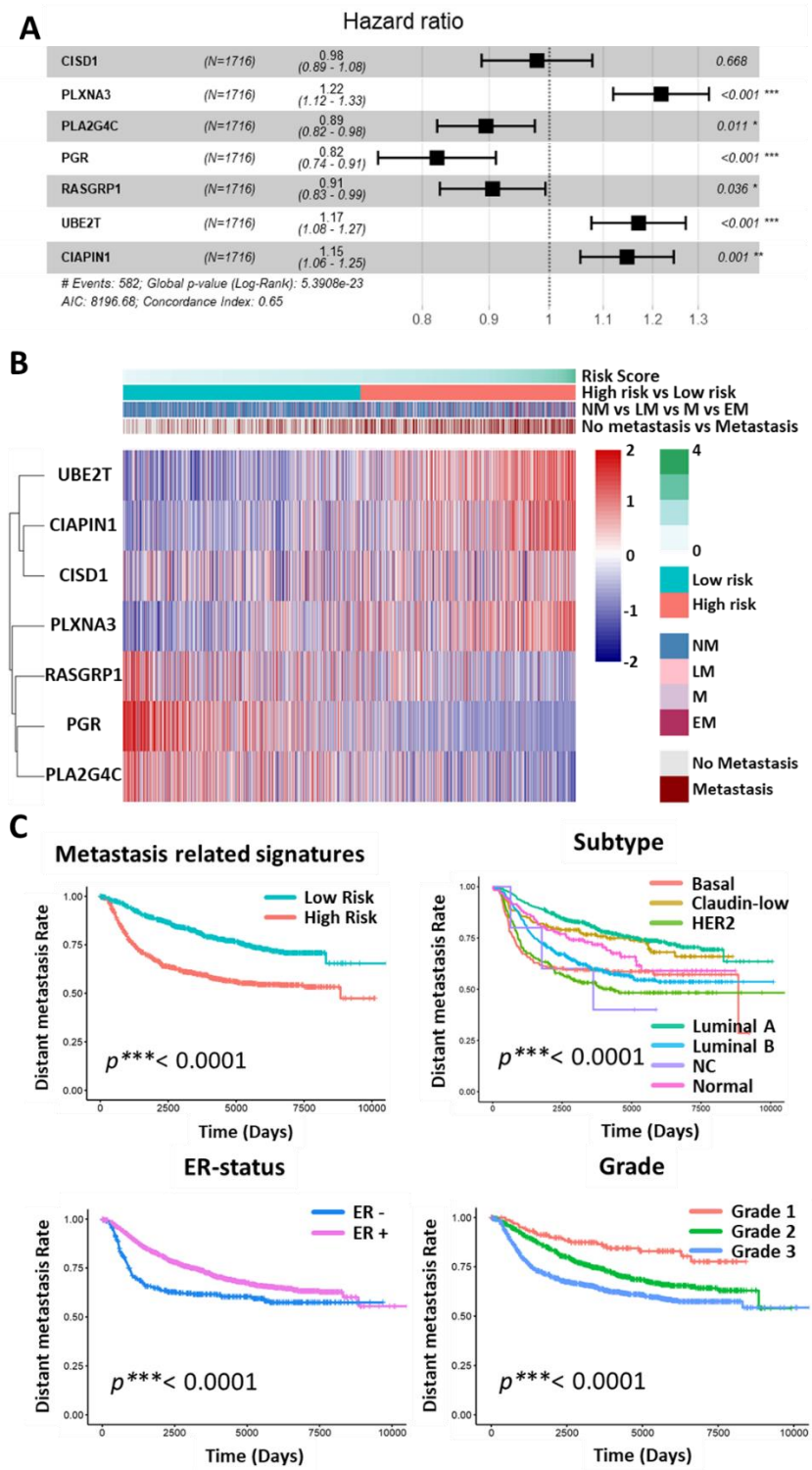


Figure 11 Construction of a distant metastasis-related molecular signatures of breast cancer. (A) Hazard ratio and *p*-value in multivariate Cox regression. (B) Heatmap of the expression profiles of seven distant-metastasis prognosis candidates. (C) Kaplan-Meier survival curves plotted to estimate the distant metastasis probabilities for the risk score-based groups, molecular subtypes, ER-status and grades.

8. The metastasis effect of CISD1 and PLXNA3 depletion on breast cancer.

For the discovery of basal protein levels of no metastasis prognostic candidates, PLXNA3 and CISD1 in breast cancer cell line, we performed a tandem mass tag (TMT) proteomic analysis and immunoblotting. Combining two results (Figure 12 A), PLXNA3 was highly expressed in MDA-MB-157 and Hs578T cell lines. In addition, CISD1 showed overexpressed in T-47D and BT474 cell lines. Therefore, to identify CISD1 and PLXNA3 associated metastatic effect in breast cancer, siRNA-mediated knockdown approaches were applied to breast cancer cell lines. We performed invasion and migration assay. After transfection with negative control and target siRNA (40nM), we confirmed that the siRNA targeting PLXNA3 could reduce the PLXNA3 levels in MDA-MB-157 and Hs 578T cells, and the CISD1 siRNA reduced the CISD1 levels in BT474 and T47D cells (Figure 12B)**.** In MDA-MB-157 and Hs578T, the migration and invasion ability increased with PLXNA3 depleted condition. However, migrative and invasive capacity decreased in CISD1 downregulated BT474 and T47D cells (Figure 12C)**.** Taken together, PLXNA3 and CISD1 can be critical in distant metastasis in breast cancer.

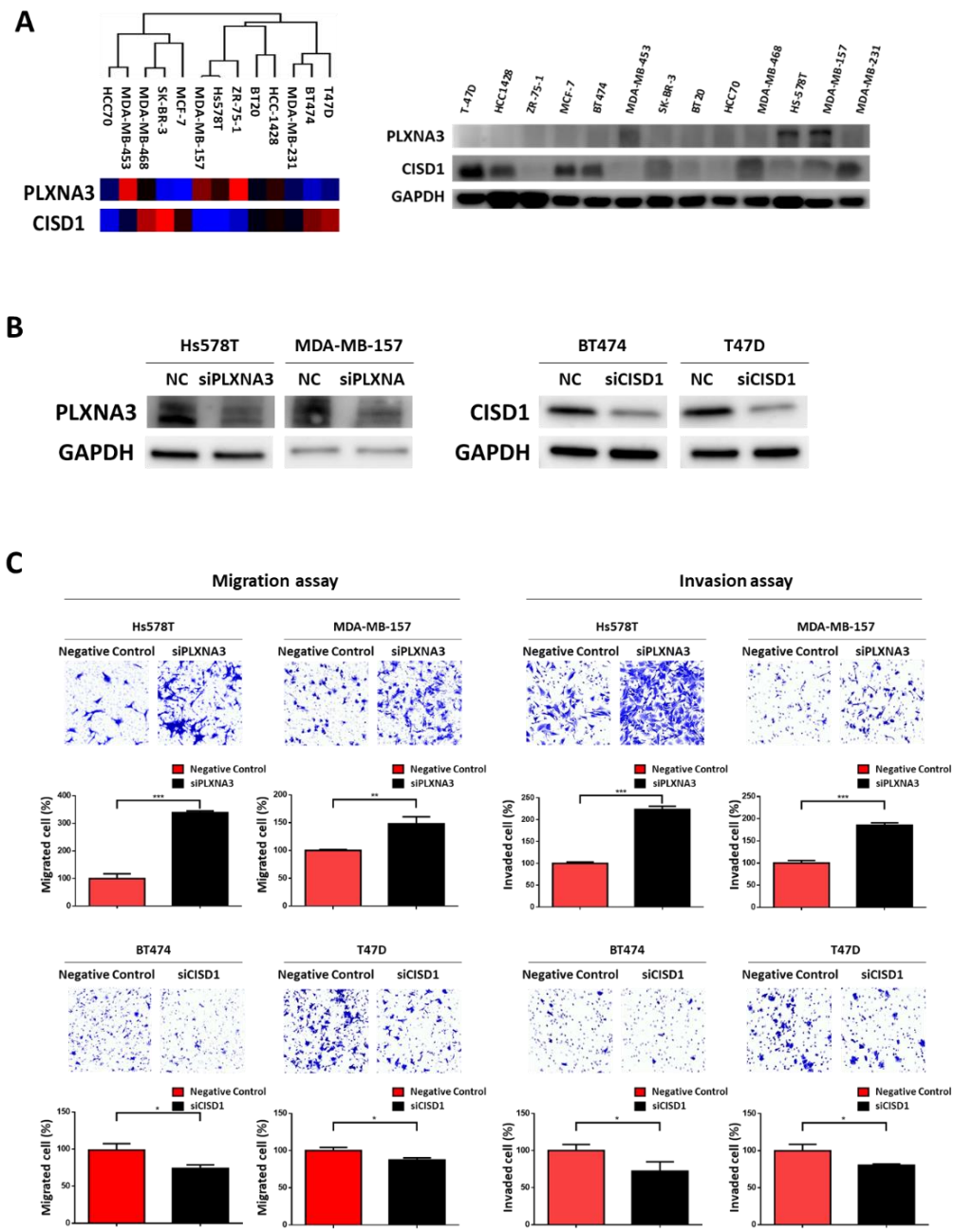


Figure 12 Functional validation of invasive role of PLXNA3 and CISD1 using siRNAs. (A) Basal PLXNA3 and CISD1 protein expression level in breast cancer cell lines. Heatmap on the left represents the proteomics data. Immunoblotting results on the right presented protein expression level with antibodies against PLXNA3, CISD1 and GAPDH (as a loading control). (B) siRNA transfection decreases the levels of PLXNA3 and CISD1 proteins in two different breast cancer cell lines, analyzed by western blotting. (C) Results of cell migration and invasion assay.

Discussion

We present a proteomic study of distant metastasis, analyzing 28 primary breast FFPE tumor samples. The patients were classified into three groups according to distant metastasis status and interval. Using our approaches, we were able to identify a number of the novel (CIAPIN1, CISD1, PLA2G4C, PLXNA3 and RASGRP1) and existing (PGR and UBE2T) [89] proteins.

Breast cancer metastasis is a complex process that requires alteration of gene and protein expression allowing tumor cells to escape from the primary tumor site. Various risks and molecular characteristics of distant metastatic breast cancer have not been established, and most of the prediction and diagnosis of distant metastasis of breast cancer using molecular biomarkers remain unexamined.

Most archived tissues in pathology collections exist as FFPE samples, representing a rich resource for clinical research. Over the past decade, MS-based proteomics has been used to analyze proteins from FFPE samples [90-95]. However, FFPE samples may harbor greater variation in protein quality than fresh frozen (FF) samples due to formalin-induced chemical modifications [96]. Ostasiewicz et al. demonstrated a comparison of FFPE and FF mouse liver tissues and found similar protein patterns. However, this was not confirmed in human tissues. Recently, Piehowski et al. [97] analyzed 60 ovarian cancer FFPE samples with storage from 7 to 32 years using TMT-labeling method and label-free proteomics approach and reported no significant proteome expression difference in terms of age and storage time. Although the practicality, robustness and reproducibility of FFPE protein mechanics in terms of sample preparation and LC-MS analysis have not been strictly established, it is beneficial to investigate the clinical value of FFPE samples. We developed a

platform for in-depth profiling of breast FFPE tumor tissues. In this study, we identified total 9,455 proteins. Using this platform, Our data had greater depth than other proteomic studies on FFPE tissues [3]. Our in-depth data on the breast FFPE tissue proteome will be a valuable resource for breast cancer distant metastasis-related research.

The first, pair wise comparison analysis and svm based feature selection was conducted for supervised analysis. For comparison analysis, we applied a multiple testing correction. There were 50 proteins, 12 proteins, and 7 proteins with significantly altered abundance in EM compared with NM, LM compared with NM, and EM compared with LM, respectively. This result suggested that heterogeneity between EM and NM would be the largest. However, it was difficult to confirm the heterogeneity through comparison analysis, we conducted the svm-based feature selection to find specific markers in EM. WGCNA provides powerful module preservation statistics which can be used to quantify similarity to another condition and widely used in the screening of biomarkers that predict disease progression. Also, module preservation statistics allow one to study differences between the modular structure of networks[98]. Therefore, we conducted WGCNA for mining protein module information as unsupervised analysis. Common features extracted from supervised analysis and unsupervised analysis were selected for further survival analysis.

For further survival analysis, we used METABRIC datasets. Previous work on the METABRIC dataset used only gene expression data to figure out the effective genes for each subtype, without applying integration to benefit from all data sources. Yet, we used mRNA data and clinical data (distant metastasis status and distant metastasis-free survival). Surprisingly, mRNA data of all seven protein marker

candidates revealed significant associations with distant metastasis free survival. PLXNA3, PGR, and RASGRP1 were found in the M2 (turquoise) module. PLXNA3 (Plexin-A3) is a semaphorin receptor (SEMA3A and SEMA3F). PLXNA3 might be involved in cytoskeletal remodeling and as well as apoptosis. PGR is a progesterone receptor known as prognostic marker and a predictive marker in breast cancer. The expression of PGR is included in both the 21-gene recurrence score (Oncotype DX, Genomic Health Inc., Redwood City, CA, USA) and the 50-gene signature classifying BC into the molecular intrinsic subtypes [99] (PAM-50). The loss of PGR occurs before or during transcription [100]. At the genetic level, PgR loss might be explained by a copy number loss of the PGR gene, which was reported to occur in 27–52% of cases of BC [101]. Also, in metastatic BC, the loss of PgR expression on CTCs may occur, even if still present in both primary tumors and metastases [102]. RASGRP1 is an activator of Ras and other related small GTPases by the virtue of functioning as guanine nucleotide exchange factors (GEFs) [103]. It functions as diacylglycerol (DAG)-regulated nucleotide exchange factor specifically activating Ras through the exchange of bound GDP for GTP. It activates the ERK/MAP kinase cascade and regulates T-cells and B-cells development, homeostasis and differentiation. Alternatively, spliced transcript variants encoding different isoforms have been identified. Altered expression of the different isoforms of this protein may be a cause of susceptibility to systemic lupus erythematosus (SLE). Recently, increasing evidence has shown that RasGRP1 plays an important role in many human diseases, especially inflammatory diseases and cancers [104]. Furthermore, RASGRP1 expression is relevant to tumor cells metastasize from breast to the lymph nodes and brain in breast cancer.

UBE2T was found in the M3 (brown) module. UBE2T (Ubiquitin-conjugating

enzyme E2T) related to this gene include chromatin binding and ubiquitin-protein transferase activity. Among its related pathways are Metabolism of proteins and Gastric Cancer Network. Also, UBE2T relate to breast, cervical [105], prostate [106] and other cancer [107] distant metastasis. UBE2T is included in the 50-gene signature classifying BC into the molecular intrinsic subtypes (PAM-50). In breast cancer cell lines. UBE2T promotes proliferation, invasion and glycolysis [108]. UBE2T can mediate the transfer of ubiquitin from ubiquitin-activating enzyme E1 to a substrate protein or E3 ligase. However, the intricate association between UBE2T and the underlying biological functions related to distant metastasis in breast cancer still remain unknown.

CISD1 and CIAPIN1 were in the M5. CISD1 (mitoneet) regulates mitochondrial homeostasis [109], especially iron homeostasis. While a single genetic mutation, amplification or deletion is insufficient to cause metastasis, ROS through Fenton reactions can stimulate widespread modifications to DNA, proteins and lipids which promotes a more aggressive tumor phenotype. With altered iron homeostasis, iron in the mitochondria accumulates then the cellular microenvironment became acidic. This condition causes breaks down the extracellular matrix, and promotes migration and invasion. CIAPIN1 is cytokine-induced apoptosis inhibitor 1 in both nucleus and cytoplasm. In a previous study, CIAPIN1 accelerates vascular remodeling that could induce angiogenesis. Cytosolic [4Fe-4S] cluster biogenesis in eukaryotes is a complex process requiring several enzymes and has not been fully resolved. The process is thought to begin with the NEET proteins on the outer mitochondrial membrane followed by [2Fe-2S] cluster transfer via CIAPIN1 to the CIA assembly factors for MMS19-mediated insertion into target apo-proteins. MitoNEET (CISD1) and NAF-1 (CISD2) are [2Fe-2S] proteins located on the outer membrane of the

mitochondria that aid in the export and completion of extramitochondrial Fe–S proteins[110]. Both CISD1 and CISD2 are presumed to transfer their [2Fe-2S] cluster to anamorsin (CIAPIN1); however, their function has not been elucidated definitively [111]. The CIAPIN1/NDOR1 complex directly interacts with mitoNEET (CISD1) to reduce the [2Fe-2S] cluster [112].

Beyond the clinical significance of seven proteins in our dataset, we also validated its significance on external datasets. Evaluation of 1,904 primary breast cancer samples from the METABRIC dataset showed that mRNA levels of candidate proteins were also significantly associated with distant metastasis-free survival. The gene signatures discovered from expression profiles have been used in survival analysis [113-115]. Through the univariate regression model, survival-related genes are selected and the resulting *p*-value and hazard ratio were used for gene selection. Based on univariate Cox regression analysis, we select 7 proteins as a distant metastasis-related prognostic signature. METABRIC data were divided into low-risk and high-risk groups according to the risk score and results of survival analyses revealed that a higher risk score was a poor prognostic indicator of distant metastasis-free survival.

In summary, our study reveals the biological effects of candidate proteins in the distant metastasis of BC. We discovered novel protein biomarker candidate and preciously suggested prognosis markers that have the potential to distinguish distant metastatic breast cancer. We expect that our protein candidates can be used to diagnose and predict distant metastatic breast cancer. Given those breast cancer distant metastasis risk related signatures, I will additionally validate in vitro and human specimens.

General Discussion

Cancer is the second major cause of mortality worldwide, presenting a major challenge to healthcare systems. In 2018, an estimated 1.7 million new incidences of cancer were diagnosed in the United States alone, resulting in 600,000 new deaths [116]. Clinical practices are now being improved by research into early detection strategies, risk group classification, and treatment efficacies. Using a systems biology approach targeted at biomarker discovery, most of this research has defined tumors at the molecular level [117].

The heterogeneity of breast cancer and underlying biology mechanisms deeply challenges the drive for prognosis and treatment decision. A widely used staging system, the Tumor-Node-Metastasis (TNM) classification of tumors, is used for diagnosis reflecting the primary tumor size, regional nodal extent, and absence or presence of metastasis, but the prediction of prognosis is not accurate [118]. Also, based on estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) protein expression, tumor status is determined by immunohistochemistry (IHC) [119]. However, since the diversity and plasticity make difficult to diagnose, predict prognosis and establish a treatment, clinical biomarkers of breast cancer must be developed.

Other fields of clinical research aim to uncover molecular variations between cancer patients and healthy controls, or between different stages of cancer as it progresses. These include genomics and transcriptomics, which have discovered lots of new cancer-initiating genes. While these omics datasets have shown the potential to compare and contrast different clinical cancer groups, one problem is that these alterations do not always correlate to our understanding of disease

biology. Proteins, on the other hand, are the biomolecules that carry out the majority of biological functions, making them predictors of disease progression [120]. Furthermore, most cancer treatments, including the rapidly emerging field of immunotherapies, use proteins as effective targets. Clinical proteomics, or the large-scale study of proteins, including their expression, functions, and structure, and applying the findings to enhance clinical outcomes [117], is a developing field in molecular clinical research.

Proteomic techniques, in particular, have advanced significantly beyond the total numbers of proteins quantified and lists of proteins and peptides detected in a specimen. We can now identify and quantify low abundance proteins, characterize and quantify their post-translational modifications, employ antibody and antibody-free approaches to validate protein identity, and designate prospective therapeutic protein targets [121].

In this study, we demonstrated two quantitative proteomic studies to identify tumor progression of breast cancer. In chapter I, we performed a TMT-based proteomic resource representing the response to CD44 regulation in aggressive, claudin-low breast cancer cell lines. In chapter II, we presented a quantitative proteomic landscape of distant metastatic breast cancer. Using FFPE tissues from patients, we provided insights into the biology of aggressive distant metastatic breast cancer by large-scale proteomics and bioinformatics.

The cellular proteomic results identified CD44 as a cancer regulator in claudin-low breast cancer by affecting the expression levels of many proteins in chapter I. We applied a multiple testing correction to our expression data to remove the potential problem, false-positive results. As a result, 4908 and 855 proteins were statistically significant in MDA-MB-231 and Hs 578T, respectively. The number

of differentially expressed proteins were quite different. Several previous studies including gene expression profiling [44] and global proteome profiling [122] showed that MDA-MB231 and Hs578T are closely related, despite of differences of genetic backgrounds, such as origin and mutation. That means the two cell lines are quite similar in molecular portraits at the basal expression level. However, despite these similarities, differences in physiological features of two cell lines including progenitor cell properties (CD44+/CD24-) and heterogenous responses to stimulus [123] or drug were also observed in many studies. Despite the same analysis platform and the same statistical criteria, a large difference in the numbers of DEPs was observed in this study. To pinpoint the roles of CD44 in two cell lines exactly, we focused on commonly modulated proteins and biological processes between two cell lines as much as possible. The commonly altered biological processes were investigated. For the purpose of the study, comparing the differences protein expression levels in each cell line would provide valuable information. As a result, the effect of CD44-knockdown, specific to Claudin-low breast cancer, could regulate the metabolic change contributed to the decrease of breast cancer cell invasive capability. These findings provide the framework for future proteomic investigations and also suggest that CD44 is a potential therapeutic target for the treatment of the claudin-low breast cancer.

The clinical proteomic results identified seven proteins, CIAPIN1, CISD1, PGR, PLA2G4C, PLXNA3, RASGRP1 and UBE2T, as distant metastasis prognosis markers using FFPE tissues in chapter II. To discover marker candidates, supervised analysis, pairwise differential analysis, and svm-based feature selection were performed first. The pattern analysis takes advantage of the cohort's matched nature and focuses on changes in protein expression levels through states [124].

The expression-based network construction provides powerful module preservation statistics that may be used to quantify similarity to another condition and are commonly employed in the screening of disease progression biomarkers. For further validation, a common feature identified from the supervised and unsupervised analysis was used. Kaplan–Meier analysis using Cox proportional hazard model showed that indeed patients with high or low candidate mRNA levels in primary tumor could predict prognosis. Some of the seven identified distant metastasis-related proteins have already been reported to play essential roles in tumor progression across malignancies [89]. Of note, one of the strengths of our work was the combination of proteomics data mRNA data that validated distant metastasis signatures that we identified. In conclusion, our research reveals the biological effects of candidate proteins in distant BC metastasis. We identified a novel protein biomarker candidate and beneficial properties markers that have the potential to recognize distant metastatic breast cancer from localized breast cancer. Our protein candidates should be able to diagnose and indicate distant metastatic breast cancer.

In my study, tissue global proteomic studies can be further extended to the investigation of post-translational modifications (PTMs). Yang et al. [125] studied the global and glycoproteome of non-small cell lung carcinoma subtypes. In total 18 patient samples consisted of three squamous cell carcinoma (SqCC) tumor samples with matched benign tissues, six adenocarcinoma (ADC) samples with five matched benign samples and one normal healthy tissue. The digested samples were iTRAQ labelled and enriched for N-glycopeptides followed by reversed-phase LC fractionation prior to shotgun MS analysis. Different protein and glycoprotein signatures were found in ADC and SqCC samples, with pathways

distinguishing between tumour types. In addition, Lehmann et al. [126] performed a comprehensive analysis of mutation, copy number, transcriptomic, epigenetic, proteomic, and phospho-proteomic patterns, which describe the genomic landscape of TNBC subtypes. However, the current conventional approach for global proteomic analysis is often carried out separately from PTM analysis [127]. Using breast cancer cell lines and tumor samples, I plan to develop an integrated approach for multiplex analysis of global, phospho-, and ubiquitin-proteomics.

Finally, without enough rigor and reproducibility in both human and animal model pre-clinical research, new biomarker candidates revealed in my investigation must be confirmed, verified, or advanced to actual patient use [128, 129]. Pre-clinical and clinical investigations must contain processes to minimize variability, ensuring the inclusion of appropriate control subjects with significant statistical power for discovery, authentication of reagents and materials, followed by verification and validation of the biomarkers [121]. In addition, comparable to IHC, MS-based proteomic assays will become clinically acceptable companion diagnostic techniques for breast cancer diagnosis, prognosis, and therapy decisions. Multiple reaction monitoring (MRM)-based mass spectrometry biomarker validation provides antibody-free confirmation of protein/peptide identity. MRM analyses are becoming more widely used and may be easily implemented in clinical laboratories [121]. In a conclusion, as part of my research, I'll establish MS-based proteomic assays of new biomarker candidates for the prediction of breast cancer metastasis.

Reference

1. Luond, F., S. Tiede, and G. Christofori, *Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression.* Br J Cancer, 2021. **125**(2): p. 164-175.
2. Kosok, M., et al., *Comprehensive Proteomic Characterization Reveals Subclass-Specific Molecular Aberrations within Triple-negative Breast Cancer.* iScience, 2020. **23**(2): p. 100868.
3. Tyanova, S., et al., *Proteomic maps of breast cancer subtypes.* Nat Commun, 2016. **7**: p. 10259.
4. Shenoy, A., et al., *Proteomic patterns associated with response to breast cancer neoadjuvant treatment.* Mol Syst Biol, 2020. **16**(9): p. e9443.
5. Jun, S., et al., *ABHD12 Knockdown Suppresses Breast Cancer Cell Proliferation, Migration and Invasion.* Anticancer Res, 2020. **40**(5): p. 2601-2611.
6. Yang, J., et al., *Lipocalin 2 promotes breast cancer progression.* Proc Natl Acad Sci U S A, 2009. **106**(10): p. 3913-8.
7. Vinik, Y., et al., *Proteomic analysis of circulating extracellular vesicles identifies potential markers of breast cancer progression, recurrence, and response.* Sci Adv, 2020. **6**(40).
8. Mertins, P., et al., *Proteogenomics connects somatic mutations to signalling in breast cancer.* Nature, 2016. **534**(7605): p. 55-62.
9. Pozniak, Y., et al., *System-wide Clinical Proteomics of Breast Cancer Reveals Global Remodeling of Tissue Homeostasis.* Cell Syst, 2016. **2**(3): p. 172-84.
10. Nakshatri, H., E.F. Srour, and S. Badve, *Breast cancer stem cells and intrinsic subtypes: controversies rage on.* Curr Stem Cell Res Ther, 2009. **4**(1): p. 50-60.
11. Radler, P.D., et al., *Highly metastatic claudin-low mammary cancers can originate from luminal epithelial cells.* Nat Commun, 2021. **12**(1): p. 3742.
12. Colzani, E., et al., *Time-dependent risk of developing distant metastasis in breast cancer patients according to treatment, age and tumour characteristics.* Br J Cancer, 2014. **110**(5): p. 1378-84.

13. Louderbough, J.M. and J.A. Schroeder, *Understanding the dual nature of CD44 in breast cancer progression.* Mol Cancer Res, 2011. **9**(12): p. 1573-86.
14. Jordan, A.R., et al., *The Role of CD44 in Disease Pathophysiology and Targeted Treatment.* Front Immunol, 2015. **6**: p. 182.
15. Olsson, E., et al., *CD44 isoforms are heterogeneously expressed in breast cancer and correlate with tumor subtypes and cancer stem cell markers.* BMC Cancer, 2011. **11**: p. 418.
16. Ponta, H., L. Sherman, and P.A. Herrlich, *CD44: from adhesion molecules to signalling regulators.* Nat Rev Mol Cell Biol, 2003. **4**(1): p. 33-45.
17. Zoller, M., *CD44: can a cancer-initiating cell profit from an abundantly expressed molecule?* Nat Rev Cancer, 2011. **11**(4): p. 254-67.
18. Misra, S., et al., *Hyaluronan-CD44 interactions as potential targets for cancer therapy.* FEBS J, 2011. **278**(9): p. 1429-43.
19. Turley, E.A., P.W. Noble, and L.Y. Bourguignon, *Signaling properties of hyaluronan receptors.* J Biol Chem, 2002. **277**(7): p. 4589-92.
20. Xu, H., et al., *CD44 as a tumor biomarker and therapeutic target.* Exp Hematol Oncol, 2020. **9**(1): p. 36.
21. Chen, C., et al., *The biology and role of CD44 in cancer progression: therapeutic implications.* J Hematol Oncol, 2018. **11**(1): p. 64.
22. Todaro, M., et al., *CD44v6 is a marker of constitutive and reprogrammed cancer stem cells driving colon cancer metastasis.* Cell Stem Cell, 2014. **14**(3): p. 342-56.
23. Schmitt, F., et al., *Cancer stem cell markers in breast neoplasias: their relevance and distribution in distinct molecular subtypes.* Virchows Arch, 2012. **460**(6): p. 545-53.
24. Herrera-Gayol, A. and S. Jothy, *CD44 modulates Hs578T human breast cancer cell adhesion, migration, and invasiveness.* Exp Mol Pathol, 1999. **66**(1): p. 99-108.
25. Prat, A., et al., *Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.* Breast Cancer Res, 2010. **12**(5): p. R68.
26. Gerhard, R., et al., *Immunohistochemical features of claudin-low intrinsic*

subtype in metaplastic breast carcinomas. Breast, 2012. **21**(3): p. 354-60.

27. Creighton, C.J., et al., *Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features.* Proc Natl Acad Sci U S A, 2009. **106**(33): p. 13820-5.
28. Dan, T., et al., *CD44 is prognostic for overall survival in the NCI randomized trial on breast conservation with 25 year follow-up.* Breast Cancer Res Treat, 2014. **143**(1): p. 11-8.
29. Bellerby, R., et al., *Overexpression of Specific CD44 Isoforms Is Associated with Aggressive Cell Features in Acquired Endocrine Resistance.* Front Oncol, 2016. **6**: p. 145.
30. Zhang, L. and J.E. Elias, *Relative Protein Quantification Using Tandem Mass Tag Mass Spectrometry.* Methods Mol Biol, 2017. **1550**: p. 185-198.
31. Choudhary, C. and M. Mann, *Decoding signalling networks by mass spectrometry-based proteomics.* Nat Rev Mol Cell Biol, 2010. **11**(6): p. 427-39.
32. Aebersold, R. and M. Mann, *Mass-spectrometric exploration of proteome structure and function.* Nature, 2016. **537**(7620): p. 347-55.
33. Deshmukh, A.S., et al., *Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors.* Mol Cell Proteomics, 2015. **14**(4): p. 841-53.
34. Cox, J. and M. Mann, *1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data.* BMC Bioinformatics, 2012. **13 Suppl 16**: p. S12.
35. Lee, H., et al., *Quantitative Proteomic Analysis Identifies AHNAK (Neuroblast Differentiation-associated Protein AHNAK) as a Novel Candidate Biomarker for Bladder Urothelial Carcinoma Diagnosis by Liquid-based Cytology.* Mol Cell Proteomics, 2018. **17**(9): p. 1788-1802.
36. Wisniewski, J.R. and F.Z. Gaugaz, *Fast and sensitive total protein and Peptide assays for proteomic analysis.* Anal Chem, 2015. **87**(8): p. 4110-6.
37. Kim, J.Y., et al., *Reconstruction of pathway modification induced by nicotinamide using multi-omic network analyses in triple negative breast cancer.* Sci Rep, 2017. **7**(1): p. 3466.
38. Tyanova, S. and J. Cox, *Perseus: A Bioinformatics Platform for Integrative*

Analysis of Proteomics Data in Cancer Research. Methods Mol Biol, 2018. **1711**: p. 133-148.

39. Liebermeister, W., et al., *Visual account of protein investment in cellular functions.* Proc Natl Acad Sci U S A, 2014. **111**(23): p. 8488-93.
40. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PLoS One, 2011. **6**(7): p. e21800.
41. Greene, C.S., et al., *Understanding multicellular function and disease with human tissue-specific networks.* Nat Genet, 2015. **47**(6): p. 569-76.
42. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-D613.
43. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
44. Charafe-Jauffret, E., et al., *Gene expression profiling of breast cell lines identifies potential new basal markers.* Oncogene, 2006. **25**(15): p. 2273-84.
45. Fougner, C., et al., *Re-definition of claudin-low as a breast cancer phenotype.* Nat Commun, 2020. **11**(1): p. 1787.
46. Lehmann, B.D., et al., *Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.* J Clin Invest, 2011. **121**(7): p. 2750-67.
47. Lawrence, R.T., et al., *The proteomic landscape of triple-negative breast cancer.* Cell Rep, 2015. **11**(4): p. 630-44.
48. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.
49. Warburg, O., F. Wind, and E. Negelein, *The Metabolism of Tumors in the Body.* J Gen Physiol, 1927. **8**(6): p. 519-30.
50. DeBerardinis, R.J. and N.S. Chandel, *Fundamentals of cancer metabolism.* Sci Adv, 2016. **2**(5): p. e1600200.
51. Tamada, M., et al., *Modulation of glucose metabolism by CD44 contributes to antioxidant status and drug resistance in cancer cells.* Cancer Res, 2012. **72**(6): p. 1438-48.
52. Lanning, N.J., et al., *Metabolic profiling of triple-negative breast cancer*

cells reveals metabolic vulnerabilities. Cancer Metab, 2017. **5**: p. 6.

53. Hershey, B.J., et al., *Lipid Droplets Define a Sub-Population of Breast Cancer Stem Cells.* J Clin Med, 2019. **9**(1).
54. Toole, B.P., *Hyaluronan-CD44 Interactions in Cancer: Paradoxes and Possibilities.* Clin Cancer Res, 2009. **15**(24): p. 7462-7468.
55. Yuen, H.F., et al., *Impact of oncogenic driver mutations on feedback between the PI3K and MEK pathways in cancer cells.* Biosci Rep, 2012. **32**(4): p. 413-22.
56. Tiwari, A., et al., *Blocking Y-Box Binding Protein-1 through Simultaneous Targeting of PI3K and MAPK in Triple Negative Breast Cancers.* Cancers (Basel), 2020. **12**(10).
57. Eckert, L.B., et al., *Involvement of Ras activation in human breast cancer cell signaling, invasion, and anoikis.* Cancer Res, 2004. **64**(13): p. 4585-92.
58. Munoz-Maldonado, C., Y. Zimmer, and M. Medova, *A Comparative Analysis of Individual RAS Mutations in Cancer Biology.* Front Oncol, 2019. **9**: p. 1088.
59. Senbanjo, L.T. and M.A. Chellaiah, *CD44: A Multifunctional Cell Surface Adhesion Receptor Is a Regulator of Progression and Metastasis of Cancer Cells.* Front Cell Dev Biol, 2017. **5**: p. 18.
60. Nam, K., et al., *CD44 regulates cell proliferation, migration, and invasion via modulation of c-Src transcription in human breast cancer cells.* Cell Signal, 2015. **27**(9): p. 1882-94.
61. Yen, T.Y., et al., *Glycoproteins in Claudin-Low Breast Cancer Cell Lines Have a Unique Expression Profile.* J Proteome Res, 2017. **16**(4): p. 1391-1400.
62. Daemen, A., et al., *Modeling precision treatment of breast cancer.* Genome Biol, 2013. **14**(10): p. R110.
63. Prat, A., et al., *Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes.* Breast Cancer Res Treat, 2013. **142**(2): p. 237-55.
64. Muller, S., et al., *CD44 regulates epigenetic plasticity by mediating iron endocytosis.* Nat Chem, 2020. **12**(10): p. 929-938.
65. Krishnan, A., et al., *Genome-wide prediction and functional*

characterization of the genetic basis of autism spectrum disorder. Nat Neurosci, 2016. **19**(11): p. 1454-1462.

66. Whitfield, M.L., et al., *Common markers of proliferation.* Nat Rev Cancer, 2006. **6**(2): p. 99-106.
67. Okegawa, T., et al., *Cell adhesion proteins as tumor suppressors.* J Urol, 2002. **167**(4): p. 1836-43.
68. Nair, K.S., R. Naidoo, and R. Chetty, *Expression of cell adhesion molecules in oesophageal carcinoma and its prognostic value.* J Clin Pathol, 2005. **58**(4): p. 343-51.
69. Moh, M.C. and S. Shen, *The roles of cell adhesion molecules in tumor suppression and cell migration: a new paradox.* Cell Adh Migr, 2009. **3**(4): p. 334-6.
70. Jia, D., et al., *Elucidating the Metabolic Plasticity of Cancer: Mitochondrial Reprogramming and Hybrid Metabolic States.* Cells, 2018. **7**(3).
71. Lyons, A., et al., *Insulin-like growth factor 1 signaling is essential for mitochondrial biogenesis and mitophagy in cancer cells.* J Biol Chem, 2017. **292**(41): p. 16983-16998.
72. LeBleu, V.S., et al., *PGC-1alpha mediates mitochondrial biogenesis and oxidative phosphorylation in cancer cells to promote metastasis.* Nat Cell Biol, 2014. **16**(10): p. 992-1003, 1-15.
73. Lunetti, P., et al., *Metabolic reprogramming in breast cancer results in distinct mitochondrial bioenergetics between luminal and basal subtypes.* FEBS J, 2019. **286**(4): p. 688-709.
74. Gherardi, E., et al., *Targeting MET in cancer: rationale and progress.* Nat Rev Cancer, 2012. **12**(2): p. 89-103.
75. Ho-Yen, C.M., J.L. Jones, and S. Kermorgant, *The clinical and functional significance of c-Met in breast cancer: a review.* Breast Cancer Res, 2015. **17**: p. 52.
76. Orian-Rousseau, V., et al., *CD44 is required for two consecutive steps in HGF/c-Met signaling.* Genes Dev, 2002. **16**(23): p. 3074-86.
77. Cecchi, F., et al., *Targeted disruption of heparan sulfate interaction with hepatocyte and vascular endothelial growth factors blocks normal and oncogenic signaling.* Cancer Cell, 2012. **22**(2): p. 250-62.

78. Joosten, S.P.J., et al., *Hepatocyte growth factor/MET and CD44 in colorectal cancer: partners in tumorigenesis and therapy resistance.* Biochim Biophys Acta Rev Cancer, 2020. **1874**(2): p. 188437.
79. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2020.* CA Cancer J Clin, 2020. **70**(1): p. 7-30.
80. Early Breast Cancer Trialists' Collaborative, G., *Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials.* Lancet, 2005. **365**(9472): p. 1687-717.
81. Zhang, Y., et al., *Unrestricted modification search reveals lysine methylation as major modification induced by tissue formalin fixation and paraffin embedding.* Proteomics, 2015. **15**(15): p. 2568-79.
82. Jin, M.S., et al., *Integrated Multi-Omic Analyses Support Distinguishing Secretory Carcinoma of the Breast from Basal-Like Triple-Negative Breast Cancer.* Proteomics Clin Appl, 2018. **12**(5): p. e1700125.
83. Jang, H.N., et al., *Mass Spectrometry-Based Proteomic Discovery of Prognostic Biomarkers in Adrenal Cortical Carcinoma.* Cancers (Basel), 2021. **13**(15).
84. Buczak, K., et al., *Spatially resolved analysis of FFPE tissue proteomes by quantitative mass spectrometry.* Nat Protoc, 2020. **15**(9): p. 2956-2979.
85. Yanovich-Arad, G., et al., *Proteogenomics of glioblastoma associates molecular patterns with survival.* Cell Rep, 2021. **34**(9): p. 108787.
86. Umoh, M.E., et al., *A proteomic network approach across the ALS-FTD disease spectrum resolves clinical phenotypes and genetic vulnerability in human brain.* EMBO Mol Med, 2018. **10**(1): p. 48-62.
87. Hulsen, T., J. de Vlieg, and W. Alkema, *BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams.* BMC Genomics, 2008. **9**: p. 488.
88. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
89. Prat, A., et al., *Molecular features and survival outcomes of the intrinsic subtypes within HER2-positive breast cancer.* J Natl Cancer Inst, 2014. **106**(8).

90. Broeckx, V., et al., *Comparison of multiple protein extraction buffers for GeLC-MS/MS proteomic analysis of liver and colon formalin-fixed, paraffin-embedded tissues.* Mol Biosyst, 2016. **12**(2): p. 553-65.
91. Fowler, C.B., T.J. O'Leary, and J.T. Mason, *Improving the Proteomic Analysis of Archival Tissue by Using Pressure-Assisted Protein Extraction: A Mechanistic Approach.* J Proteomics Bioinform, 2014. **7**(6): p. 151-157.
92. Fu, Z., et al., *Improved protein extraction and protein identification from archival formalin-fixed paraffin-embedded human aortas.* Proteomics Clin Appl, 2013. **7**(3-4): p. 217-24.
93. Giusti, L. and A. Lucacchini, *Proteomic studies of formalin-fixed paraffin-embedded tissues.* Expert Rev Proteomics, 2013. **10**(2): p. 165-77.
94. Gustafsson, O.J., G. Arentz, and P. Hoffmann, *Proteomic developments in the analysis of formalin-fixed tissue.* Biochim Biophys Acta, 2015. **1854**(6): p. 559-80.
95. Jain, M.R., et al., *Proteomic identification of immunoproteasome accumulation in formalin-fixed rodent spinal cords with experimental autoimmune encephalomyelitis.* J Proteome Res, 2012. **11**(3): p. 1791-803.
96. Gaffney, E.F., et al., *Factors that drive the increasing use of FFPE tissue in basic and translational cancer research.* Biotech Histochem, 2018. **93**(5): p. 373-386.
97. Piehowski, P.D., et al., *Residual tissue repositories as a resource for population-based cancer proteomic studies.* Clin Proteomics, 2018. **15**: p. 26.
98. Langfelder, P., et al., *Is my network module preserved and reproducible?* PLoS Comput Biol, 2011. **7**(1): p. e1001057.
99. Johansson, H.J., et al., *Breast cancer quantitative proteome and proteogenomic landscape.* Nat Commun, 2019. **10**(1): p. 1600.
100. Liu, X.Y., et al., *Genomic Landscape and Endocrine-Resistant Subgroup in Estrogen Receptor-Positive, Progesterone Receptor-Negative, and HER2-Negative Breast Cancer.* Theranostics, 2018. **8**(22): p. 6386-6399.
101. Mohammed, H., et al., *Progesterone receptor modulates ERalpha action in breast cancer.* Nature, 2015. **523**(7560): p. 313-7.
102. Madaras, L., et al., *BRCA Mutation-Related and Claudin-Low Breast*

Cancer: Blood Relatives or Stepsisters. Pathobiology, 2016. **83**(1): p. 1-12.

103. Ksionda, O., A. Limnander, and J.P. Roose, *RasGRP Ras guanine nucleotide exchange factors in cancer.* Front Biol (Beijing), 2013. **8**(5): p. 508-532.
104. Salzer, E., et al., *RASGRP1 deficiency causes immunodeficiency with impaired cytoskeletal dynamics.* Nat Immunol, 2016. **17**(12): p. 1352-1360.
105. Liang, J., et al., *The ubiquitin-conjugating enzyme E2-EPF is overexpressed in cervical cancer and associates with tumor growth.* Oncol Rep, 2012. **28**(4): p. 1519-25.
106. Wen, M., et al., *Elevated expression of UBE2T exhibits oncogenic properties in human prostate cancer.* Oncotarget, 2015. **6**(28): p. 25226-39.
107. Yu, H., et al., *Ubiquitin-Conjugating Enzyme E2T is an Independent Prognostic Factor and Promotes Gastric Cancer Progression.* Tumour Biol, 2016. **37**(9): p. 11723-11732.
108. Qiao, L., C. Dong, and B. Ma, *UBE2T promotes proliferation, invasion and glycolysis of breast cancer cells by regualting the PI3K/AKT signaling pathway.* J Recept Signal Transduct Res, 2021: p. 1-9.
109. Bai, F., et al., *The Fe-S cluster-containing NEET proteins mitoNEET and NAF-1 as chemotherapeutic targets in breast cancer.* Proc Natl Acad Sci U S A, 2015. **112**(12): p. 3698-703.
110. Mittler, R., et al., *NEET Proteins: A New Link Between Iron Metabolism, Reactive Oxygen Species, and Cancer.* Antioxid Redox Signal, 2019. **30**(8): p. 1083-1095.
111. Lipper, C.H., et al., *Cancer-Related NEET Proteins Transfer 2Fe-2S Clusters to Anamorsin, a Protein Required for Cytosolic Iron-Sulfur Cluster Biogenesis.* PLoS One, 2015. **10**(10): p. e0139699.
112. Camponeschi, F., S. Ciofi-Baffoni, and L. Banci, *Anamorsin/Ndor1 Complex Reduces [2Fe-2S]-MitoNEET via a Transient Protein-Protein Interaction.* J Am Chem Soc, 2017. **139**(28): p. 9479-9482.
113. Pucci, F., et al., *PF4 Promotes Platelet Production and Lung Cancer Growth.* Cell Rep, 2016. **17**(7): p. 1764-1772.
114. Song, H., et al., *High expression of FOXR2 in breast cancer correlates with poor prognosis.* Tumour Biol, 2016. **37**(5): p. 5991-7.
115. Zhao, Y.F., et al., *FOXD1 promotes breast cancer proliferation and*

chemotherapeutic drug resistance by targeting p27. Biochem Biophys Res Commun, 2015. **456**(1): p. 232-7.

116. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018.* CA Cancer J Clin, 2018. **68**(1): p. 7-30.
117. Macklin, A., S. Khan, and T. Kislinger, *Recent advances in mass spectrometry based clinical proteomics: applications to cancer research.* Clin Proteomics, 2020. **17**: p. 17.
118. Weiss, A., et al., *Validation Study of the American Joint Committee on Cancer Eighth Edition Prognostic Stage Compared With the Anatomic Stage in Breast Cancer.* JAMA Oncol, 2018. **4**(2): p. 203-209.
119. Aitken, S.J., et al., *Quantitative analysis of changes in ER, PR and HER2 expression in primary breast cancer and paired nodal metastases.* Ann Oncol, 2010. **21**(6): p. 1254-1261.
120. Yaffe, M.B., *Why geneticists stole cancer research even though cancer is primarily a signaling disease.* Sci Signal, 2019. **12**(565).
121. Mueller, C., et al., *Protein biomarkers for subtyping breast cancer and implications for future research.* Expert Rev Proteomics, 2018. **15**(2): p. 131-152.
122. Lawrence, R.T., et al., *The Proteomic Landscape of Triple-Negative Breast Cancer.* Cell Rep, 2015. **11**(6): p. 990.
123. Ye, I.C., et al., *Molecular Portrait of Hypoxia in Breast Cancer: A Prognostic Signature and Novel HIF-Regulated Genes.* Mol Cancer Res, 2018. **16**(12): p. 1889-1901.
124. Jansen, J.P. and H. Naci, *Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers.* BMC Med, 2013. **11**: p. 159.
125. Yang, S., et al., *Protein signatures of molecular pathways in non-small cell lung carcinoma (NSCLC): comparison of glycoproteomics and global proteomics.* Clin Proteomics, 2017. **14**: p. 31.
126. Lehmann, B.D., et al., *Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes.* Nat Commun, 2021. **12**(1): p. 6276.
127. Zhou, Y., et al., *An Integrated Workflow for Global, Glyco-, and Phospho-*

proteomic Analysis of Tumor Tissues. Anal Chem, 2020. **92**(2): p. 1842-1849.

128. Letai, A., *Functional precision cancer medicine-moving beyond pure genomics.* Nat Med, 2017. **23**(9): p. 1028-1035.

129. Freedman, L.P., G. Venugopalan, and R. Wisman, *Reproducibility2020: Progress and priorities.* F1000Res, 2017. **6**: p. 604.

Abstract in Korean

국문 초록

정량 단백체학 및 생물정보학을 이용한 공격적인 유방암 바이오 마커의 발굴

서울대학교

의학과 병리학 전공

김 혜 윤

서론: 질량분석기 기반 단백체학은 대규모 분자생물학과 세포생물학을 단백질 수준에서 다루는 기술이다. 대량 단백질의 동정 및 정량으로 단백체학 분석기법은 단백질의 서열, 발현, 전사 후 변형 및 단백질-단백질 상호작용 등을 해석할 수 있도록 한다. 세포주부터 제한된 양의 임상 시료인 체액, 신선한 냉동 조직, 파라핀 포매 (FFPE) 조직 등으로부터 단백질을 추출한다. 높은 처리량과 감도를 가진 차세대 고속 질량분석기 기반 분석으로 수천 개의 단백질을 동시에 정량 하여 대량의 데이터를 생산한다. 생물정보학 분석 기법을 활용하여 질병의 상태, 예후, 치료에 따른 효과에 따른 단백질 발현 수준의 차이를 감지할 수 있고, 더 나아가 질병의 생물학적 메커니즘을 제시할 수

있다.

방법: 1장에서, 가장 공격적인 삼중 음성 (TNBC) 유방암 하위 유형인 클라우딘 낮은 (Claudin-low) 하위 유형에서 암 줄기세포 마커인 CD44의 역할을 규명하였다. 유전자 조작 기법을 통해 CD44 발현을 조절한 세포주를 구축하였다. CD44의 발현을 감소시켰을 때, 단백질 발현 양상의 변화를 분석하여 분자생물학적 역할을 입증하였다.

2장에서, 유방암 환자 중 타장기로의 원격 전이 고위험군 환자에 대한 예후 예측 바이오 마커를 발굴하기 위하여 동일 병기 28명의 환자 (조기 원격전이: 9명, 지연 원격전이: 9명, 비원격전이: 10명) FFPE 종양 조직을 수집하였다. 제한적인 양의 시료를 분석하기 위한 단백체 분석법을 확립하였다. 원격전이 예후예측을 위한 바이오 마커 후보군을 발굴하였고, 전사체 외부 데이터에서 회귀 모델을 개발하여 검증하였다.

결과: 1장에서, Cluain-low 하위유형 유방암 세포주 MDA-MB-231에서 7396개, Hs578T 에서 6567개의 단백질을 동정하였다. 통계적으로 유의한 발현의 차이를 나타낸 MDA-MB-231의 4908개 단백질, Hs578T의 855개 단백질을 생물정보학 분석 (gene ontology, 단백질-단백질 상호작용 네트워크 분석) 하여 세포 증식, 대사과정, 유전자의 발현 조절을 통한 암화 과정을 제시하였다. 생물학적 메커니즘의 확인을 위해 기능 연구를 수행하였고, CD44가 대량의 단백질의 발현을 조절하여 세포 증식과 이동을 조절하는 것을 검증하였다.

2장에서, 유방암 FFPE 슬라이드에서 종양 부분만을 선별하여 분리하여 질량 분석하여 9455개의 단백질을 동정하였다. 원 발암 진단 후 원격전이가 일어난 기간에 따라 발현의 유의한 차이가 있는 단백질 중 비교 분석, 상관관계 네트워크 분석, 머신 러닝 기반 특성 추출, 생존분석을 통해 7개의 최종 바이오 마커 후보군을 발굴하였다. 7개의 마커 후보군으로 외부데이터를 활용하여 Cox 비례 위험 회귀 모델을 구축하여 원격전이 예측할 수 있음을 확인하였다.

결론: 1장에서 2가지 Cluain-low 하위유형 유방암 세포주에서 암 줄기세포 마커인 CD44 발현을 감소시킨 단백체 발현 비교 데이터를 생성하였다. 이를 분석하여 CD44가 암세포의 유전적 발현, 대사, 부착을 유기적으로 조절하여 핵심 암화 과정인 세포 증식, 이동에 영향을 주는 것을 확인하였다. 이를 통해 공격적인 삼중 음성 유방암의 핵심 조절인자인 CD44의 생물학적 기전을 분자적 수준에서 이해할 수 있도록 도왔으며, 더 나아가 Cluain-low 하위유형 유방암의 잠재적 치료의 표적 물질이 될 수 있음을 확인하였다.

2장에서 유방암 환자의 파라핀 포매 종양 조직 단백체 분석을 통해 심층적인 단백체 데이터를 생성하였고, 원격전이 예측을 위한 잠재적 바이오 마커를 발굴하였다. 이러한 원격전이 예후 예측 바이오 마커의 개발과 생물 정보학 분석을 통한 분자 생물학적 기전의 규명은 정밀의학 실현의 핵심 근거자료로 활용할 수 있을 것이며, 유방암 환자의 효과적인 치료 계획 수립에 도움을 줄 것으로 기대한다.

주 요 어: 유방암; 단백체학; 질량 분석학; 바이오 마커; CD44; 암화 과정; 파라핀 포매; 원격 전이; 예후 예측

학 번: 2017-20079

* 본 내용은 학술지(Journal of Proteome Research)에 게재된 논문, Quantitative Proteomics Reveals Knockdown of CD44 Promotes Proliferation and Migration in Claudin-Low MDA-MB-231 and Hs 578T Breast Cancer Cell Lines (Kim, H., Woo, J., Dan, K., Lee, K. M., Jin, M. S., Park, I. A., Ryu, H. S., & Han, D.). (2021년 1월 게재)을 바탕으로 작성하였음.