이학박사 학위논문

Biomarker Development of Distant Metastatic Breast Cancer and Mood Disorders using Quantitative Proteomics and Bioinformatics

정량 단백체학 및 생물정보학을 이용한 원격 전이성 유방암 및 정서질환의 바이오마커 개발

2022년 08월

서울대학교 대학원

의과학과 의과학전공

신 동 윤

A thesis of the Degree of Doctor of Philosophy

정량 단백체학 및 생물정보학을 이용한 원격 전이성 유방암 및 정서질환의 바이오마커 개발

Biomarker Development of Distant Metastatic Breast Cancer and Mood Disorders using Quantitative Proteomics and Bioinformatics

August 2022

Major in Biomedical Sciences

Department of Biomedical Sciences

Seoul National University

Graduate School

Dongyoon Shin

정량 단백체학 및 생물정보학을 이용한 원격 전이성 유방암 및 정서질환의 바이오마커 개발

지도 교수 김 영 수

이 논문을 이학박사 학위논문으로 제출함

2022년 4월

서울대학교 대학원

의과학과 의과학전공

신 동 윤

신동윤의 이학박사 학위논문을 인준함

2022년 7월

위 원 장 이 용 석 (인)

부위원장 김 영 수 (인)

위 원 한 범 (인)

위 원 김 종 서 (인)

위 원 김 경 곤 (인)

Biomarker Development of Distant Metastatic Breast Cancer and Mood Disorders using Quantitative Proteomics and Bioinformatics

by

Dongyoon Shin

A thesis submitted to the Department of Biomedical Sciences in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Biomedical Sciences at Seoul National University Graduate School

July 2022

Approved by Thesis Committee:

Professor ______________________Chairman

Professor ______________________Vice chairman

Professor ______________________

Professor ______________________

Professor ______________________

ABSTRACT

Biomarker Development of Distant Metastatic Breast Cancer and Mood Disorders using Quantitative Proteomics and Bioinformatics

Dongyoon Shin

Major in Biomedical Sciences

Department of Biomedical Sciences

Seoul National University

Graduate School

Introduction: Liquid chromatography (LC)-mass spectrometry (MS)-based proteomic approaches have been applied to discover and develop biomarkers that are associated with specific diseases and disorders. Untargeted proteomics based on LC-high resolution MS has enabled simultaneous identification and quantification of thousands of proteins and hundreds of differentially expressed proteins (DEPs) in small amounts of samples. Targeted proteomics including LC-multiple reaction monitoring (MRM)-MS has been used to quantify interesting proteins with high sensitivity, accuracy, and reproducibility. Numerous clinical proteomics studies employ pathological and clinical specimens collected from clinical cohorts such as formalin-fixed paraffin-embedded (FFPE) tissues, blood, and other body fluids,

. For clinical proteomic analysis, LC-MS-based approaches are powerful technologies in discovery and development of biomarkers with their high throughput and high sensitivity. In addition, proteomic studies based on LC-MS will contribute to understanding of biological and molecular features of specific diseases and disorders.

Methods: In chapter I, an integrated untargeted proteomic approach that combined filter-aided sample preparation (FASP), tandem mass tag labeling (TMT), high pH fractionation, and LC-high resolution-MS was applied to acquire in-depth proteomic profiling data from FFPE tissues of distant metastatic breast cancer patients collected from a clinical cohort. Statistical analyses were performed to determine DEPs and discover candidate biomarkers for predicting distant metastatic breast cancer. Bioinformatics analyses were performed to examine molecular characteristics of distant metastatic breast cancer. In addition, in vitro assays were performed to validate distant metastatic potential of candidate biomarkers. In chapter II, targeted proteomic approach based on LC-MRM-MS was applied to quantify protein targets associated with major depressive disorder (MDD) and bipolar disorder (BD) in plasma samples collected from a clinical cohort. Batch-effect correction of LC-MRM-MS data was performed to reduce technical variations. Subsequently, univariate analysis was performed to determine proteomic candidate features, and machine learning approaches were performed to develop a potential diagnostic model for discriminating MDD and BD. In addition, network analysis was performed to examine biological associations between proteins included in the model and mood disorders.

Results: In chapter I, a total of 9,441 and 8,746 proteins were identified from FFPE-TMT pooled samples set and FFPE-TMT individual samples set comparing distant metastasis and non-distant metastasis groups, respectively. In addition, 7,823 proteins were identified from the TMT-labeled breast cancer cell lines set comparing low invasive and high invasive cell lines. Two proteins (LTF and TUBB2A) were determined as candidate biomarkers. As a result, TUBB2A, which maintained consistent expression patterns between different quantitation platforms, was selected as a novel biomarker candidate. TUBB2A showed potential of distant metastatic activities. In addition, distinct alterations of proteome and molecular functions of distant metastatic breast cancer between breast cancer subtypes were demonstrated. In chapter II, 210 protein targets corresponding to 671 peptides pertinent to MDD and BD were stably and reproducibly quantified by LC-MRM-MS in individual plasma samples. In the training set, nine plasma protein biomarkers were developed and a generalizable model comprised of the nine proteins was constructed. The model demonstrated good performance (AUC > 0.8) in discriminating MDD from BD in the training (AUC = 0.84) and test sets (AUC = 0.81) and in distinguishing MDD from BD without current hypomanic/manic/mixed symptoms (AUC > 0.83). Subsequently, the model demonstrated excellent performance for drug-free MDD vs BD (AUC > 0.96) and good performance for MDD vs HC (AUC > 0.87) and BD vs HC (AUC > 0.86). Furthermore, the nine proteins were associated with neuro, oxidative and nitrosative stress, and immunity and inflammation-related biological functions.

Conclusions: In chapter I, I constructed the largest FFPE tissue proteome of distant metastatic breast cancer proteome using. The depth of our dataset allowed us to discover a novel biomarker candidate as well as the proteomic characteristics of distant metastatic breast cancer. Distinct molecular features of various breast cancer subtypes were also established. Thus, our proteomic data can serve as a valuable resource for research on distant metastatic breast cancer. In chapter II, the viability of discriminating MDD and BD patients using a targeted proteomic approach was proposed. Our results suggest that the nine plasma proteins and their combined model has the potential to discriminate between MDD and BD patients and help diagnostic decision-making. Through both studies, the potential of LC-MS-based proteomics in the discovery and development of biomarkers was demonstrated.

Keywords: Untargeted and targeted proteomics; Liquid chromatography-high resolution mass spectrometry; Liquid chromatography-multiple reaction monitoring -mass spectrometry; Biomarker; Distant metastatic breast cancer; Mood disorder

Student number: 2017-25626

* This work is published in Clinical Proteomics. Identification of TUBB2A by quantitative proteomic analysis as a novel biomarker for the prediction of distant metastatic breast cancer (Shin, D., Park, J., Han, D., Moon, J. H., Ryu, H. S., & Kim, Y.) Published 24 May 2020/Clinical Proteomics 10.1186/s12014-020-09280-z.

* This work is published in Journal of Proteome Research. Quantitative Proteomic

Approach for Discriminating Major Depressive Disorder and Bipolar Disorder by Multiple Reaction Monitoring-Mass Spectrometry (Shin, D., Rhee, S. J., Lee, J., Yeo, I., Do, M., Joo, E. J., ... & Kim, Y.) Published 7 May 2021/Journal of Proteome Research 10.1021/acs.jproteome.1c00058.

CONTENTS

LIST OF TABLES

Chapter I

Chapter II

LIST OF FIGURES

GENERAL INTRODUCTION

Chapter I

Chapter II

LIST OF ABBREVIATIONS

ACN: acetonitrile

TUBB2A: tubulin beta 2A chain

LTF: lactotransferrin

HSPA9: stress-70 protein: mitochondrial

PSMB4: proteasome subunit beta type-4

CTNNA1: catenin alpha-1

XPO5: exportin-5

PAFAH1B3: platelet-activating factor acetylhydrolase IB subunit gamma

TMSB10: thymosin beta-10

TMSB4X: thymosin beta-4

RAC1: ras-related protein-Rac1

PFKM: ATP-dependent 6-phosphofructokinase

PKM: pyruvate kinase PKM

FN1: fibronectin

GP6: glycoprotein 6

CV: coefficient of variation

FDR: false discovery rate

DEP: differentially expressed protein

FFPE: formalin-fixed paraffin-embedded

MS: mass spectrometry

TMT: tandem mass tag

RT-PCR: real-time polymerase chain reaction

Dis-meta: distant metastasis

Non dis-meta: non-distant metastasis

GO: gene ontology

HCD: higher-energy collisional dissociation

AGC: automatic gain control

NCE: normalized collision energy

DDA: data-dependent acquisition

Maximum IT: maximum ion injection time

HPRP: high-pH reversed-phase

RP: reverse-phase

FFPE: formalin-fixed paraffin-embedded

TMT: tandem mass tag

BC: breast cancer.

MDD: major depressive disorder

BD: bipolar disorder

MRM: multiple reaction monitoring

MS: mass spectrometry

HC: healthy control

LASSO: least absolute shrinkage and selection operator

CSF: cerebrospinal fluid

C1QB: complement C1q subcomponent subunit B

DSG3: desmoglein-3

FBLN1: fibulin-1

FCGBP: IgG Fc-binding protein

FHR3: complement factor H-related protein 3

GPX3: glutathione peroxidase 3

IGHM: immunoglobulin heavy constant mu

ITIH2: inter-alpha-trypsin inhibitor heavy chain H2

PLF4: platelet factor 4

DTT: dithiothreitol

IAA: iodoacetamide

SIS: stable isotope-labeled internal standard

LC: liquid chromatography

QQQ: a triple-quadrupole

ESI: electrospray ionization

AuDIT: automated detection of inaccurate and imprecise transitions

PAR: peak area ratio

L/H: light/heavy

BMI: body mass index

BPRS: Brief Psychiatric Rating Scale

MADRS: Montgomery-Asberg Depression Rating Scale

YMRS: Young Mania Rating Scale

HAM-A: Hamilton Anxiety Rating Scale

AP: antipsychotics

MS: mood stabilizer

AD: antidepressants

BZD/HNT: benzodiazepines/hypnotics

AUROC: area under the receiver operating characteristics

AIC: Akaike’s information criterion

AICc: the bias-corrected version of AIC

w: Akaike weight

ANOVA: analysis of variance

ANCOVA: a blend of analysis of variance and regression

IPA: Ingenuity Pathway Analysis

SZ: schizophrenia

DEPs: differentially expressed proteins

MALDI: matrix-assisted laser desorption and ionization

TOF: time-of-flight

HSD: honestly significant difference

NOS: not otherwise specified

DSM-5: manual of mental disorders 5th version

MINI: mini-international neuropsychiatric interview

NSAIDs: non-steroidal anti-inflammatory drugs

ECT: electroconvulsive therapy

TMS: transcranial magnetic stimulation

tDCS: transcranial direct current stimulation

DBS: deep brain stimulation

EDTA: ethylenediaminetetraacetic acid

C3: complement C3

C4BPA: c4b-binding protein alpha chain

CFI: complement factor I

B2RAN2: cDNA: FLJ95014: highly similar to homo sapiens vanin 1 (VNN1): mRNA

ENG: endoglin

RAB7A: ras-related protein Rab-7a

ROCK2: rho-associated protein kinase 2

XPO7: exportin-7

GENERAL INTRODUCTION

Biomarkers are substances of organisms that discriminate diseased individuals from other normal individuals, which are gradually modified or present at abnormal amounts in specific diseases, disorders, and other health conditions. Thus, biomarkers for diseases and disorders are significant for discriminating types of diseases and disorders or predicting the progression of diseases and disorders, or might contribute to the effect of a particular treatment on clinical outcomes. Even though there are various biomarker's types, protein biomarkers are considered the most extensively influenced in disease, response, and recovery. Thus, protein biomarkers development has been needed because proteins directly affect the physiological status of the diseased cells. Although a lot of protein biomarkers have been reported to represent high accuracy, sensitivity, and specificity, there remain many diseases and disorders, which lack protein biomarkers. Because these diseases and disorders are regarded as highly devastating (e.g. distant metastatic breast cancer) or intractable (e.g. mood disorders), the requirement for efficient biomarkers is expanding in clinics.

Several protein assays such as immunoassays for the discovery of disease and disorder-specific biomarkers have been developed. Recently, among them, Liquid chromatography (LC)-mass spectrometry (MS)-based proteomic approaches have become the preferred methods, proving their analytical accuracy, sensitivity, reproducibility, precision, and stability during high-throughput analysis. A qualitative proteomic approach is a method of identifying proteins in untargeted proteomics to determine the proteome composition in a biological or clinical sample. Simultaneously, the identified proteins can be quantified by mass spectrometer signals. This approach has been applied to generating proteome map or discovering protein biomarkers associated with specific diseases and disorders. In this dissertation, an LC-high resolution MS-based untargeted proteomic approach combining proteomic sample preparation methods was used to discover protein biomarker candidates for specific diseases and disorders. Proteins were extracted from clinical samples such as formalin-fixed paraffin-embedded (FFPE) tissues. Peptides were extracted by enzymatic digestion based on filter-aided sample preparation (FASP). The peptide mixture was labeled using tandem mass tag (TMT) labeling method, and the labeled peptide mixture was desalted and fractionated in condition of high pH. The fractionated peptide samples were injected into high-resolution MS and were scanned through data dependent acquisition (DDA) mode. Subsequently, statistical and bioinformatics analyses were performed in the collected proteomic profiling data for biomarker discovery. Overall scheme of LC-MS-based untargeted proteomics was presented in Figure 1.

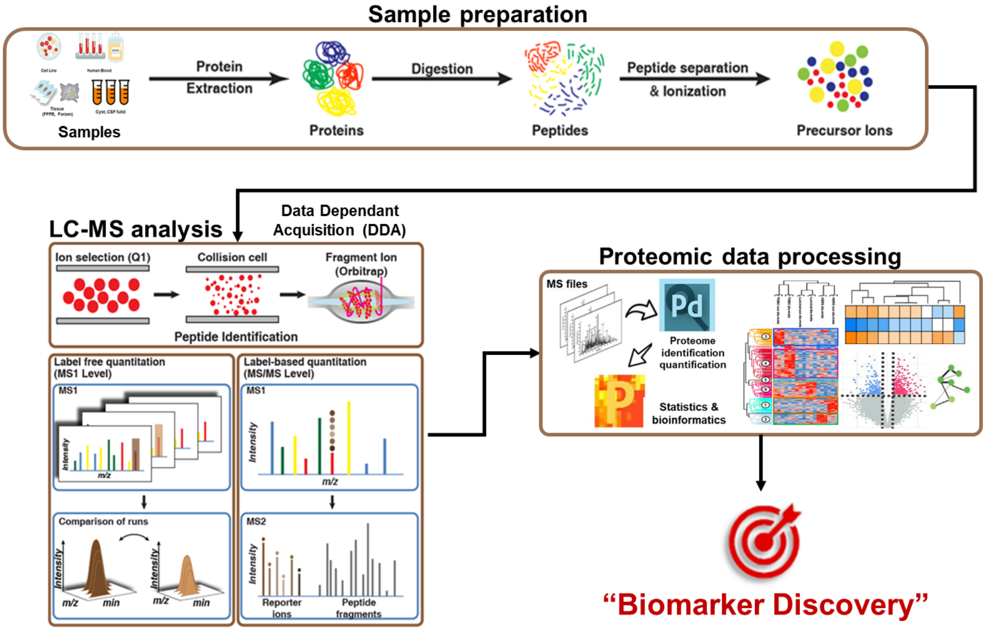


Figure 1. Overall workflow of LC-MS-based untargeted proteomics. Graphical representation of the workflow for our LC-MS-based untargeted proteomics.

A quantitative proteomic approach is a method for multiplexed quantification of protein targets of interest with high accuracy and reproducibility in a targeted proteomics. This approach has been applied to clinical research for reliable and stable quantification of existing biomarkers or for discovering protein biomarker candidates in clinical samples of a large cohort. In this dissertation, LC-MS-based targeted proteomic approach was used to quantify protein biomarker candidates of interest in clinical samples such as plasma. The protein biomarker candidates of interest were determined by various databases and references. High abundant proteins such as albumin were depleted and low abundant proteins were concentrated. Peptides were extracted by enzymatic in-solution digestion using commercial detergent such as Rapigest-SF. The stable-isotope-labeled (SIL) peptides corresponding to the protein biomarker candidate of interest were spiked into peptides mixture as internal standards. The mixture of endogenous and (SIL) peptides was desalted and injected into triple quadrupole mass spectrometer. The injected peptides were scanned in multiple reaction monitoring (MRM) mode. Subsequently, statistical and machine learning analyses were performed in the collected targeted proteomic data for biomarker development. Overall scheme of LC-MS-based targeted proteomics was presented in Figure 2.

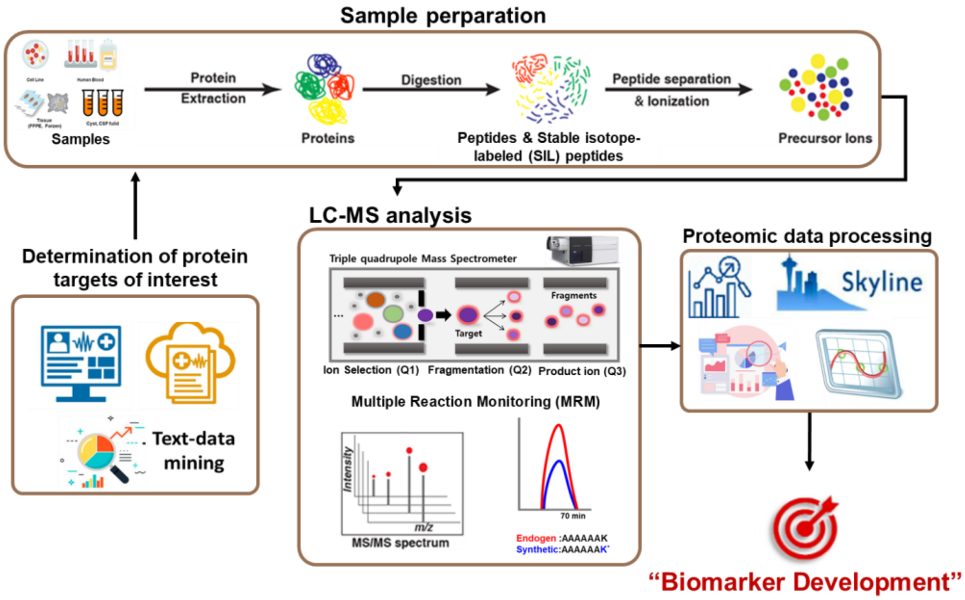


Figure 2. Overall scheme of LC-MS-based targeted proteomics. Graphical representation of the workflow for our LC-MS-based targeted proteomics.

Proteomic analyses of clinical specimens allow screening of specific alteration of proteins (e.g. differentially expressed in protein abundance) under a disease or disorder, making them suitable approaches for biomarker discovery and development. For clinical proteomic analysis, a proper proteomic approach based on LC-MS should be determined in accordance with the type and the number of clinical specimens.

In this dissertation, the application of LC-MS-based proteomic techniques to clinical specimens for biomarker discovery and development was described. The protein biomarkers of the diseases and disorders dealt with in this study have not been well examined because of technical limitations or analytical difficulties. Therefore, new-elaborate analytical procedures appropriate for the type of clinical sample to discover and develop protein biomarkers was established. In addition, novel protein biomarkers for diseases and disorders were discovered and developed through these established procedures.

In chapter I, distant metastatic breast cancer corresponds to stage 4 breast cancer that has spread to other areas of the body such as brain, bone, lung and liver. It is estimated that approximately 40,000 women die each year from invasive breast cancer (stage1-4 breast cancer) in the US. Approximately 90% of the deaths result from stage 4 breast cancer. Among breast cancers, stage 4 breast cancer that is not curable showed the lowest survival rate, representing a poor prognosis. Prediction of this breast cancer can reduce patient's burden when considering the curability of the disease and quality of life for distant metastatic breast cancer patients. Thus, molecular biomarkers that can predict distant metastatic breast cancer is of critical interest. Through LC-MS-based untargeted proteomics, the proteome of the FFPE tissues of distant metastatic breast cancer and breast cancer without distant metastasis were investigated. Overall scheme of chapter 1 was presented in Figure 3.

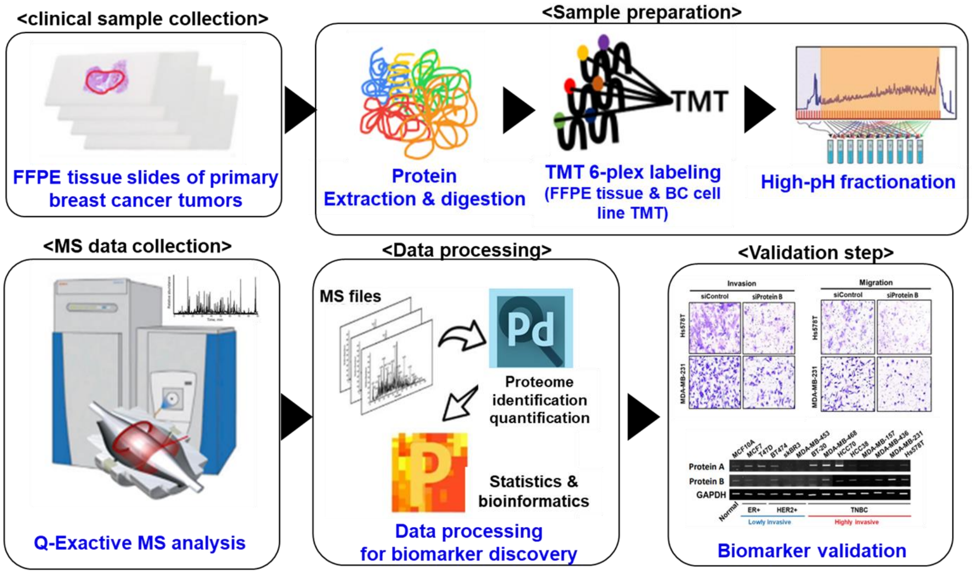


Figure 3. Overall scheme of the study of chapter 1. Graphical representation of the workflow of study of chapter 1

The proteins that compose FFPE tissues of the distant metastatic breast cancers have not been well researched. To examine candidate biomarkers, qualitative proteomic analysis employing LC-high resolution-MS was performed using the FFPE tissues collected from 18 breast cancer patients with distant metastasis and 18 breast cancer patient without distant metastasis. Through statistical analyses and step-by-step criteria for the determination of candidate biomarkers, differentially expressed proteins (DEPs) including candidate biomarkers were determined. One final selected protein was selected, and its expression level and distant metastatic potential were validated by several in vitro assays. This protein was proposed as a novel candidate biomarker for the prediction of distant metastatic breast cancer. Furthermore, bioinformatics analyses were performed regarding gene ontology, disease and functions, and canonical pathways using the DEPs to examine molecular characteristics of distant metastatic breast cancer. Through these analyses, biological functions of a novel candidate biomarker and distinct biological functions of distant metastatic breast cancer between molecular subtypes were examined. Therefore, this proteomic study will help predict distant metastatic breast cancer and understand molecular features of distant metastatic breast cancer.

In chapter II, major depressive disorder (MDD) and bipolar disorder (BD) are common mood disorders. Although diagnosis of MDD and BD relies on subjective behavioral observations and symptoms, the complexity and commonality between MDD and BD complicate the diagnosis. Misdiagnosis of both disorders results in the erroneous prescription of medication, which aggravates the symptoms of the disorders. Thus, there is an unmet need for molecular biomarkers that can discriminate MDD and BD. Through LC-MS-based targeted proteomics, plasma protein biomarkers were developed and a proteomic-based diagnostic model comprising the biomarkers for discriminating MDD and BD was established. Overall scheme of chapter2 was presented in Figure 4.

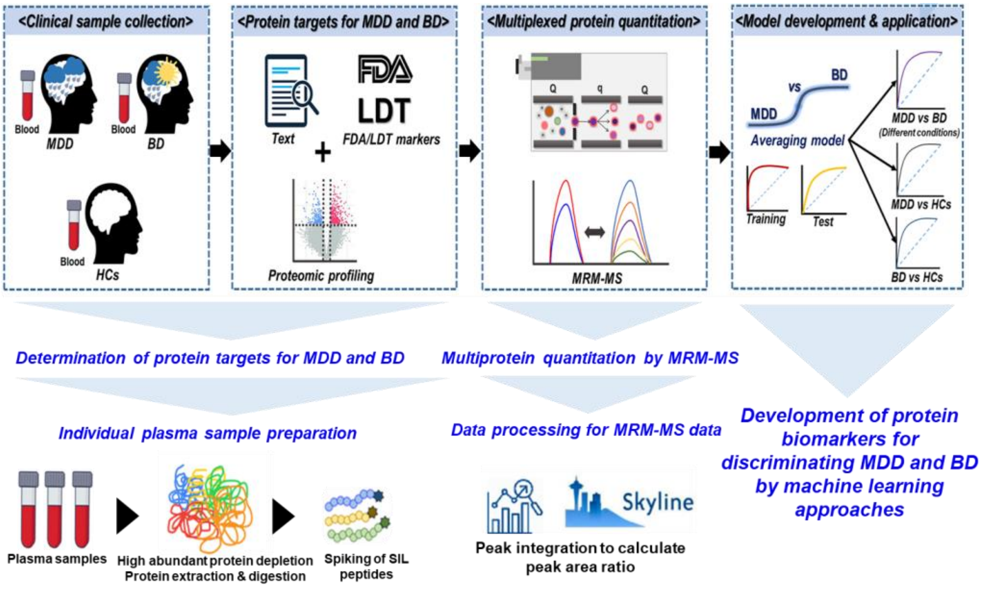

<Clinical sample collection>
Blood
MDD
Blood
BD
Blood
HCs
<Protein targets for MDD and BD>
Text
FDA
LDT
FDA/LDT markers
Proteomic profiling
<Multiplexed protein quantitation>
Q
q
Q
MRM-MS
<Model development & application>
MDD
vs
BD
Averaging model
Training
Test
MDD vs BD
(Different conditions)
MDD vs HCs
BD vs HCs
Determination of protein targets for MDD and BD
Multiprotein quantitation by MRM-MS
Individual plasma sample preparation
Data processing for MRM-MS data
Development of protein biomarkers for discriminating MDD and BD by machine learning approaches
Plasma samples
High abundant protein depletion
Protein extraction & digestion
Spiking of SIL peptides
Skyline
Peak integration to calculate peak area ratio

Figure 4. Overall scheme of the study of chapter 2. Graphical representation of the workflow of study of chapter 2

The blood protein biomarkers for the diagnosis of MDD and BD have not been well studied in a large number of clinical samples. To develop candidate biomarkers, important protein targets associated with MDD and BD were stably quantified with high sensitivity and reproducibility using LC-multiple reaction monitoring (MRM)-MS in a total of 270 individual plasma samples consisting of 90 MDD, 90 BD, and 90 healthy control (HC). Subsequently, 9-plasma protein diagnostic model was developed by machine learning approaches, resulting in good discriminatory and diagnostic performances. Furthermore, biological interactions between the 9 proteins, MDD and BD were investigated. This study proposes the potential of the developed plasma protein biomarkers and the proteomic-based diagnostic model for distinguishing MDD and BD.

CHAPTER I

Identification of TUBB2A by Quantitative Proteomic Analysis as a Novel Biomarker for the Prediction of Distant Metastatic Breast Cancer

INTRODUCTION

Breast cancer is one of the most prevalent and lethal cancers in women worldwide [1]. In particular, its annual incidence—currently 17 million cases—is increasing at an alarming rate [2, 3]. There are approximately 232,000 new cases of invasive breast cancer each year in the US, and approximately 40,000 women die each year from the disease; furthermore, roughly 90% of these deaths are caused by the most malignant form of breast cancer: distant metastatic breast cancer [2, 4]. Distant metastatic breast cancer, which preferentially metastasizes to distal organs, such as the bone, liver, lung, and brain, has a poor prognosis [5, 6]. In addition, this type of breast cancer causes various complications at the affected sites, such as pericardial effusion, pleural effusion, bone fracture, hypercalcemia, and red blood cell anemia, which worsens survival outcomes [7-9].

Distant metastatic breast cancer is assessed, based on various factors, such as tumor size, lymphovascular invasion, histological grade, nodal involvement, and hormone receptor status—all of which are independent risk factors for distant metastatic breast cancer [10-13]. Among these factors, breast cancer molecular subtypes are associated with various patterns of distant metastatic spread and related to differences in survival outcomes [10, 14]. For instance, the most widely known molecular subtypes, such as the luminal A, luminal B, HER2, and basal-like (triple-negative) groups, have site-specific, cumulative metastatic incidence rates, demonstrating substantial differences in the distant metastatic behavior of and overall survival between breast cancer subtypes [10].

Although various risks and molecular characteristics of distant metastatic breast cancer have been established, the prediction and diagnosis of distant metastasis in breast cancer with molecular biomarkers remain largely unexamined [4-6, 10-13]. Thus, characterizing the molecular signatures that are associated with distant metastasis using omics-based approaches, such as genomics, transcriptomics, and proteomics, might identify previously overlooked biomarker candidates.

Many genomic or transcriptomic studies have examined the molecular characteristics of distant metastatic breast cancer—for instance, genes that are associated with lung, brain, and bone metastasis from breast tumor [15-18, 20, 21]. In addition, genetic signatures that predict distant metastasis in breast cancer have been established through genomic profiling [19]. However, given the relatively low correlation between gene expression and protein expression, it is difficult to assume that the tendencies in genomic data will translate fully to proteomic data without verification [22-23]. Similarly, considering that transcriptomic and proteomic data have a moderate correlation, the molecular characteristics of the transcriptome could not perfectly represent those of the proteome [24-26]. In the case of breast cancer, recent large dataset-based proteomic approaches have reported an intermediate correlation between the breast tumor proteome and the corresponding transcript levels [27-28]. Furthermore, a recent report has described a low correlation between proteomes and transcriptomes in human breast cancer tissues, suggesting that a proteomic approach to human BC tissues could complement a transcriptomic method

[29].

Although proteomic studies have been performed for various diseases, including breast cancer, none has investigated the overall characteristics of distant metastatic breast cancer [29-37, 44]. Proteomic research is expected to provide greater insight into the pathogenesis of distant metastatic breast cancer, generating novel information about the molecular features of distant metastasis—for example, by discovering novel protein biomarkers for the prediction or diagnosis of distant metastatic breast cancer. Thus, an in-depth proteomic analysis is important for yielding valuable resources in distant metastatic breast cancer—data that have not been found in genomic and transcriptomic analyses.

Recent advances in mass spectrometry (MS)-based proteomics have accelerated the development of high-throughput techniques for proteomic quantification [38, 39]. In addition, a tandem mass tag (TMT)-based strategy has facilitated relative protein quantification by comparing the reporter ion intensities that are obtained by MS/MS. Because this approach can quantify thousands of proteins precisely with high sensitivity, TMT-based techniques have been used widely to generate substantial datasets [40-43]. With a 6-plex TMT quantification technique, in combination with high-resolution MS, I constructed an in-depth proteomic map of distant metastatic breast cancer.

In this study, I hypothesized that in-depth proteomic data would supply important proteins to profile the molecular signatures of distant metastatic breast cancer. Using our proteomic techniques, I identified by far the largest number of proteins from FFPE distant and nondistant metastatic breast cancer tissues. Furthermore, I determined important protein targets to validate distant metastatic potential of breast cancer. The function of these targets was determined using several approaches, including RT-PCR and invasion/migration assays.

Through our criteria to narrow down the important proteins, I discovered a novel protein biomarker candidate differentially expressed in distant metastatic breast cancer. Furthermore, I examined the distinct biological functions of distant metastatic breast cancer between molecular subtypes. In summary, I have proposed the first protein biomarker candidate that potentially be able to distinguish distant metastasis, derived from primary breast tumors using FFPE tissue samples. I performed the initial examination of its molecular features at the protein level, providing insights into the pathogenesis of distant metastatic breast cancer.

MATERIALS AND METHODS

1. Materials and reagents

Sodium dodecyl sulfate (SDS) and Trizma base were purchased from USB (Cleveland, OH), and sequencing-grade modified trypsin was purchased from Promega Corporation (Madison, WI). Dithiothreitol (DTT) and urea were obtained from AMRESCO (Solon, OH). POROS20 R2 beads were purchased from Applied Biosystems (Foster City, CA). High-purity (>97%) mass spectrometry (MS)-grade ovalbumin was obtained from Protea (Morgantown, WV), and HLB OASIS columns were purchased from Waters (Milford, MA). Tandem mass tag (TMT) 6-plex isobaric reagents; a bicinchoninic acid (BCA) assay kit; LC/MS-grade solvents, such as acetone, acetonitrile (ACN), and water; and reducing agents, such as tris (2-carboxyethyl) phosphine (TCEP), were purchased from Thermo Fisher Scientific (Waltham, MA). All other reagents, if not noted otherwise, were obtained from Sigma-Aldrich (St. Louis, MO).

2. Sample selection

All clinical samples were collected from the Department of Pathology, Seoul National University Hospital (Seoul, South Korea). The distant metastasis group (dis-meta) was defined as patients who developed distant metastasis with or without lymph node metastasis. The nondistant metastasis group (nondis-meta) comprised patients who were not diagnosed as having distant metastasis with or without lymph node metastasis. All clinical specimens were collected from 18 patients with dis-meta and 18 patients with nondis-meta. The 18 patients in each group were divided into 3 breast cancer molecular subtypes (HER2, TNBC, and luminal). Tissue samples for distant and nondistant metastatic breast cancer were derived from the primary breast tumor. Clinical information on the patient samples is detailed in Table 1. All patients consented to participate in the study per institutional review board guidelines (IRB No.1612-011-811).

Table 1. Clinical information on patients. Clinical information on all 36 patients is listed.

The pooled sample set	Sample ID	Case	Sex	Age	Molecular subtype	Date of sample collection (year)	Observation period (years)	Distant metastasis-free period (years)	Date of distant metastasis (year)	Lymph node metastasis	Organ of distant metastasis	Tumor content (%)	Tumor-infiltrating lymphocytes (%)	Surgical therapy	Chemo therapy	Radiation therapy
	S-1D	Dis-meta	F	56	HER2	2007	12	x	2009	no	bone	75	1	mastectomy	FAC #6	no
	S-2D	Dis-meta	F	35	TNBC	2007	12	x	2008	yes	lung	65	3	quadrantectomy	AC #4, Taxol #4	no
	S-3D	Dis-meta	F	37	Luminal	2008	11	x	2012	no	bone	60	1	mastectomy	FAC #6	no
	S-4D	Dis-meta	F	39	TNBC	2008	11	x	2010	yes	lung	70	1	mastectomy	AC #4, Taxol #4	yes
	S-5D	Dis-meta	F	56	TNBC	2009	10	x	2011	no	lung	80	2	quadrantectomy	FAC #6	yes
	S-6D	Dis-meta	F	41	HER2	2009	10	x	2010	yes	bone	80	2	mastectomy	no	yes
	S-7D	Dis-meta	F	43	Luminal	2009	10	x	2016	yes	bone	70	1	quadrantectomy	AC#4	yes
	S-8D	Dis-meta	F	31	HER2	2010	9	x	2011	no	bone	85	3	quadrantectomy	TCH#6,	yes
	S-9D	Dis-meta	F	48	Luminal	2009	10	x	2011	yes	brain	65	1	mastectomy	AC #4, genexol #4	yes
	S-1ND	Non-dis-meta	F	55	HER2	2006	13	13	x	no	no	75	3	mastectomy	no	no
	S-2ND	Non-dis-meta	F	50	Luminal	2008	11	11	x	yes	no	55	1	mastectomy	AC #4, docetaxel #2	no
	S-3ND	Non-dis-meta	F	59	HER2	2008	11	11	x	yes	no	80	1	mastectomy	AC #4	no
	S-4ND	Non-dis-meta	F	39	Luminal	2008	11	11	x	yes	no	60	5	mastectomy	AC #4, genexol #4	yes
	S-5ND	Non-dis-meta	F	50	TNBC	2007	12	12	x	yes	no	70	1	mastectomy	AC #4, Taxol #4	yes
	S-6ND	Non-dis-meta	F	43	HER2	2006	13	13	x	yes	no	80	1	quadrantectomy	FAC #6	yes
	S-7ND	Non-dis-meta	F	31	TNBC	2008	11	11	x	yes	no	65	1	mastectomy	AC #4, Taxol #4	yes
	S-8ND	Non-dis-meta	F	40	TNBC	2007	12	12	x	yes	no	60	5	mastectomy	AC #4, Taxol #4	yes
	S-9ND	Non-dis-meta	F	45	Luminal	2008	11	11	x	yes	no	80	1	mastectomy	AC #4, Taxol #4	yes

The individual sample set	Sample ID	Case	Sex	Age	Molecular subtype	Date of sample collection (year)	Observation period (years)	Distant metastasis-free period (years)	Date of distant metastasis (year)	Lymph node metastasis	Organ of distant metastasis	Tumor content (%)	Tumor-infiltrating lymphocytes (%)	Surgical therapy	Chemo therapy	Radiation therapy
	S-10D	Dis-meta	F	41	HER2	2006	13	x	2008	yes	bone	90	3	quadrantectomy	AC #4, Taxol #4	yes
	S-11D	Dis-meta	F	67	HER2	2007	12	x	2009	yes	lung	85	1	mastectomy	AC #4, Taxol #4	no
	S-12D	Dis-meta	F	52	Luminal	2007	12	x	2010	yes	lung	87	2	mastectomy	AC #4, Taxol #4	yes
	S-13D	Dis-meta	F	32	Luminal	2008	11	x	2011	yes	bone, brain	55	3	quadrantectomy	AC #4, Taxol #4	yes
	S-14D	Dis-meta	F	40	TNBC	2008	11	x	2011	no	bone, lung	90	1	quadrantectomy	FAC #6	yes
	S-15D	Dis-meta	F	41	TNBC	2009	10	x	2014	no	lung	87	2	mastectomy	FAC #6	yes

S-16D	Dis-meta	F	48	Luminal	2009	10	x	2016	yes	brain	80	3	mastectomy	AC #4, Paclitaxel #4	yes
S-17D	Dis-meta	F	62	TNBC	2009	10	x	2011	no	brain	50	10	quadrantectomy	FAC #6	yes
S-18D	Dis-meta	F	44	HER2	2010	9	x	2013	yes	bone	65	3	mastectomy	TCH #6	yes
S-10ND	Non-dis-meta	F	54	HER2	2009	10	10	x	no	no	90	3	mastectomy	FAC #6	no
S-11ND	Non-dis-meta	F	49	Luminal	2008	11	11	x	yes	no	45	2	quadrantectomy	AC #4, Taxol #4	yes
S-12ND	Non-dis-meta	F	56	Luminal	2008	11	11	x	yes	no	75	2	mastectomy	AC #4, Taxol #4	yes
S-13ND	Non-dis-meta	F	46	TNBC	2008	11	11	x	no	no	85	2	quadrantectomy	FAC #6	yes
S-14ND	Non-dis-meta	F	43	Luminal	2007	12	12	x	yes	no	80	2	mastectomy	AC #4, Taxol #4	no
S-15ND	Non-dis-meta	F	40	HER2	2006	13	13	x	yes	no	60	5	mastectomy	AC #4, Taxol #4	yes
S-16ND	Non-dis-meta	F	63	TNBC	2007	12	12	x	no	no	90	1	quadrantectomy	FAC #6	yes
S-17ND	Non-dis-meta	F	38	HER2	2007	12	12	x	yes	no	80	2	mastectomy	AC #4	yes
S-18ND	Non-dis-meta	F	38	TNBC	2007	12	12	x	yes	no	25	7	quadrantectomy	AC #4, Paclitaxel #4	yes

AC: Adriamycin Cyclophosphamide
FAC: Fluorouracil Adriamycin Cyclophosphamide
TCH: Docetaxel Carboplatin Trastuzumab
#: The number of attempts of chemotherapy

3. Sample preparation of FFPE tissues for proteomic analysis

FFPE sections (10 µm) were incubated twice in xylene (Sigma-Aldrich, St. Louis, MO)—once each for 5 and 2 minutes—and then twice in 100% (v/v) ethanol for 90 seconds. The sections were then hydrated in 75% (v/v) ethanol for 90 seconds and distilled water for 90 seconds [33, 44]. Next, the tissues were scraped off the glass slides into microfuge tubes, after which protein extraction buffer (4% SDS; 0.3M Tris, pH 8.5; 2 mM TCEP) was added. Following sonication, the samples were incubated at 100°C for 2.5 hours. Protein concentrations were measured using a bicinchoninic acid (BCA) reducing agent-compatible kit (Thermo Fisher Scientific, Waltham, MA).

Protein digestion was performed using a combination of acetone precipitation and filter-aided sample preparation (FASP) [45, 46]. Before the digestion step, 250 µg of extracted protein was precipitated with cold acetone at a buffer: acetone ratio of 1:5 and incubated at -20°C for 18 hours. Next, the pellet was washed with 500 µl cold acetone, centrifuged at 15,000 rpm for 15 min, and air-dried for 1.5 hours. The proteins that had precipitated were dissolved in 35 µl denaturation buffer (4% SDS and 100 mM DTT in 0.3 M TEAB pH 8.5).

After being heated at 100°C for 35 min, the denatured proteins were loaded onto 30 kDa spin filters (Merck Millipore, Darmstadt, Germany). The buffer was exchanged 3 times with UREA solution (8 M UREA in 0.1 M TEAB, pH 8.5). After SDS was removed, cysteine residues were treated with alkylation buffer (50 mM IAA, 8 M UREA in 0.1 M TEAB, pH 8.5) for 1 hour at room temperature in the dark.

UREA buffer was exchanged with TEAB buffer (40 mM TEAB, pH 8.5). The proteins were digested with trypsin (enzyme-to-substrate ratio [w/w] of 1:50) and 4% ACN at 37°C for 18 hours. The digested peptides were eluted by centrifugation, and their concentrations were measured, based on the fluorescence emission of tryptophan at 350 nm, using an excitation wavelength of 295 nm [47]. The external standard sample, ovalbumin, was digested in the same manner.

4. 6-Plex Tandem Mass Tag (TMT) Labeling

Because the number of samples exceeded that of the TMT channels, 2 independent TMT 6-plex labeling experiments—using a pooled sample set and individual sample set—were performed. Each TMT experiment consisted of 18 samples that were divided into 2 groups (dis-meta and non dis-meta). For the pooled sample set, equal amounts of 3 samples with identical molecular subtypes in each group were pooled, generating 6 pooled samples. Next, they were labeled with TMT 6-plex: 126-non dis-meta (HER2), 127-non dis-meta (TNBC), 128-non dis-meta (Luminal), 129-dis-meta (HER2), 130-dis-meta (TNBC), and 131-dis-meta (Luminal). At this step, several technical replicates of the sample sets were prepared. For the individual sample set, 18 individual patients were positioned in 3 TMT 6-plex sets: 126-non dis-meta (HER2), 127-non dis-meta (TNBC), 128-non dis-meta (Luminal), 129-dis-meta (HER2), 130-dis-meta (TNBC), and 131-dis-meta (Luminal). The detailed experimental workflow is described in Figure 1.

Prior to the TMT labeling step, 45 μg of each peptide sample was mixed with an equivalent volume of ovalbumin. Then, 40 mM TEAB buffer was added to each sample to equalize the volume. Next, TMT reagents were reconstituted in 110 μl anhydrous ACN. Each sample was labeled using 25 μl of the reconstituted TMT reagent. Then, 45 μl ACN was added in varying volumes to a final concentration of 30% and incubated at room temperature (25°C) for 1.25 hours. Hydroxylamine was added in various volumes to a concentration of 0.3% (v/v) to quench the reaction. TMT-labeled samples for each set were pooled at a ratio of 1:1. The pooled sample was lyophilized and desalted.

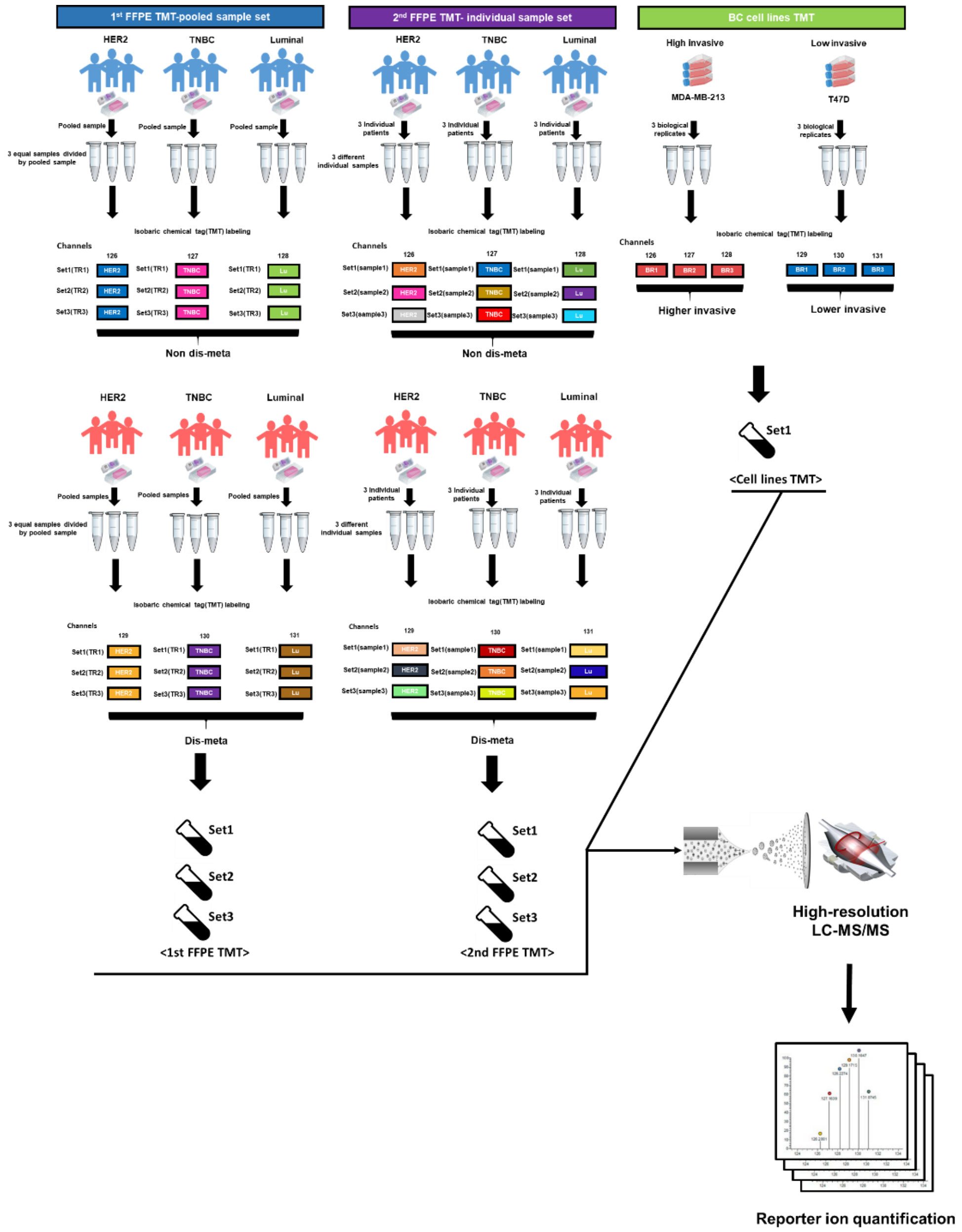


Figure 1. Detailed experimental workflow of TMT-based proteomic study. Graphical representation of the workflow for our TMT experiments. Three sample sets were analyzed using our TMT-based proteomic techniques.

5. Desalting and High-pH Reversed-Phase (HPRP) Peptide Fractionation

The TMT-labeled samples were desalted on an HLB OASIS column per the manufacturer's instructions. High-pH reversed-phase (HPRP) peptide fractionation was performed on an Agilent 1260 bioinert HPLC instrument (Agilent, Santa Clara, CA) with an Agilent 300 Extended-C18 column (4.6 mm I.D x 15 cm long, 5-μm C18 particle). TMT-labeled peptide samples were prefractionated at a flow rate of 1 mL/min for 60 min on a linear gradient, which ranged from 5% to 40% ACN with 15 mM ammonium hydroxide. The sample was separated into 96 fractions, which were then assembled into 12 fractions. The 12 fractions were lyophilized and stored at -80°C before MS analysis.

6. Sample preparation of breast cancer cells for proteomic analysis

MDA-MB-231 breast cancer cells were cultured in DMEM, and T47D cells were cultured in RPMI, containing 10% FBS and 1% penicillin and streptomycin. The cells were seeded in 75-cm2 culture plates. After a 24-hour incubation at 37°C with 5% CO2, the cells were scraped using a cell scraper and washed 3 times with 1x PBS. The scraped cell pellets were centrifuged and washed again 3 times with 1x PBS. The pellets were then transferred to microfuge tubes and mixed with protein extraction buffer (4% SDS; 0.3 M Tris, pH 7.5; 2 mM TCEP). Following sonication, the samples were incubated at 100°C for 30 minutes. After protein extraction, the subsequent experimental procedures, such as protein digestion, TMT labeling, desalting, and peptide fractionation, were performed in the same manner as the FFPE tissues.

7. Reversed-Phase (RP)-nano LC-ESI-MS/MS Analysis

The prefractionated peptides were analyzed on an LC-MS system with an Easy-nLC 1000 (Thermo Fisher Scientific, Waltham, MA) that was equipped with a nanoelectrospray ion source (Thermo Fisher Scientific, Waltham, MA) and coupled to a Q-Exactive mass spectrometer (Thermo Fisher Scientific, Waltham, MA), as described in our previous studies [45, 46]. The peptide samples were separated on a 2-column system, comprising a trap column (Thermo Fisher Scientific, 75 μm I.D. x 2 cm long, 3-μm Acclaim PepMap100 C18 beads) and an analytical column (Thermo Fisher Scientific, 75 μm I.D. x 50 cm long, 3-μm ReproSil-Pur C18-AQ beads). Lyophilized peptide samples were dissolved in Solvent A (0.1% formic acid water and 2% ACN) prior to injection.

The peptides were separated on a 180-min linear gradient, ranging from 6% to 26% Solvent B (100% ACN and 0.1% formic acid) for all peptide samples. The spray voltage was set to 2.2 kV in positive ion mode, and the heated capillary temperature was set to 320°C. Mass spectra were collected in data-dependent acquisition (DDA) mode by top 20 method. Xcaliber (version 2.5) was used to set the mass spectrometer parameters as follows: mass range to 350–1650 m/z, resolution of 70,000 at 200 m/z for detected precursor ions, automatic gain control

(AGC) at 3 x 106, isolation window for MS2 at 1.2 m/z, automatic gain control (AGC) for MS2 at 2 x 105, higher-energy collisional dissociation (HCD) scans at a resolution of 35,000, and normalized collision energy (NCE) of 32. The maximum ion injection time (maximum IT) for the full-MS and MS2 scans was 30 ms and 120 ms, respectively. Dynamic exclusion with an exclusion time of 40 s was used.

8. MS Data Search

Proteome Discoverer, version 2.2 (Thermo Fisher Scientific, Waltham, MA) was used to search the resulting RAW files. The full-MS and MS/MS spectra search was conducted using the SEQUEST HT algorithm against a modified version of the Uniprot human database (December 2014, 88,717 protein entries; http://www.uniprot.org), which included chicken ovalbumin. The database search was performed using the target-decoy strategy. The search parameters were as follows: a precursor ion mass tolerance value of 20 ppm (monoisotopic mass); a fragment ion mass tolerance value of 0.02 Da (monoisotopic mass); full enzyme digest with trypsin (after KR/−) and up to 2 missed cleavages; static modification values of 229.163 Da for lysine residues and peptide N-termini for TMT labeling and 57.02 Da for cysteine residues with carbamidomethylation; and dynamic modification values of 42.01 Da for protein N-terminal acetylation, 0.984 Da for asparagine deamidation, and 15.99 Da for methionine oxidation.

A false discovery rate (FDR) of less than 1% at the peptide and protein levels was used as the confidence criteria. Proteins were quantified by computing reporter ion relative intensities with the “Reporter Ions Quantifier” node in Proteome Discoverer. The co-isolation threshold value was 70%. The mass spectrometry-based proteome data lists of all identified proteins and peptides have been deposited into ProteomeXchange (http://proteomecentral.proteomexchange.org) through the PRIDE partner repository: dataset identifier PXD016061 [48, 69-71].

9. Quantification of protein abundance and statistical analysis

Protein levels were normalized, based on the ovalbumin content in each TMT channel. Fold-change values were calculated by dividing the average value of the normalized protein abundance in the dis-meta group by that of the non dis-meta group. Statistical analysis for the proteomic data was performed for the normalized protein levels using Perseus (version 1.5.8.5). Student’s t-test was used to identify differentially expressed proteins (DEPs) for selecting biomarker candidates that differentiate distant metastasis from nondistant metastasis of breast cancer. The statistical cutoff for the student’s t-test was a p-value < 0.05. In addition, ANOVA was used to determine DEPs for analyzing the molecular characteristics of distant metastatic breast cancer between molecular subtypes using bioinformatic tools. Specifically, 9 samples in each group were classified as HER2, TNBC, and luminal, resulting in 6 subtype groups (HER2 nondis-meta, TNBC nondis-meta, luminal nondis-meta, HER2 dis-meta, TNBC dis-meta, and luminal dis-meta). Next, the quantified proteins in these groups were analyzed to detect statistically significant proteins. The statistical cutoff for the ANOVA was p-value < 0.05. Receiver operating characteristic (ROC) analyses of biomarker performance were performed using MedCalc (version 12.5.0) and Prism (version 6.0).

10. Bioinformatics analysis

The Gene Ontology (GO) of the proteins was classified using the DAVID bioinformatics tool (version 6.8). GO classification was assessed by Fisher's exact test to obtain a series of p-values that were filtered, based on a statistical significance of 0.05. Canonical pathways and downstream biological functions were enriched by Ingenuity Pathway Analysis (IPA, QIAGEN, Redwood City, CA). The analytical algorithms in IPA were used to predict the downstream effects on known biological pathways and functions, based on the inputted list of DEPs. IPA allocates activation scores on activated or inhibited status to biological functions and pathways that underlie the quantitative values of proteins. Fisher's exact test was used to acquire p-values, whereas the degree of activation was measured using Z-scores. The p-value cutoff was set to 0.05, and the predictive activation Z-score cutoff was set to a magnitude of 1.

11. RNA extraction and real-time polymerase chain reaction (RT-PCR)

Total RNA was isolated from the following breast cancer cell lines using TRIzol (Invitrogen, Carlsbad, CA, USA) per the manufacturer's instructions:

MCF10A, MCF7, T47D, BT474, skBR3, MDA-MB-453, BT-20, MDA-MB-468, HCC70, HCC38, MDA-MB-157, MDA-MB-436, MDA-MB-231, and HS578T. Two micrograms of total RNA from each cell line was used for the reverse-transcription reaction. First-strand cDNA was synthesized by standard random priming with RNA inhibitor (Promega, Madison, WI) and Moleney murine leukemia virus reverse transcripts (Promega, Madison, WI). Following cDNA synthesis, target genes were amplified using specific primers and HIPI plus Master mix (ElpisBio, Daejeon, Korea).

12. Cell lines and culture conditions for invasion and migration assays

The MDA-MB-231 and Hs578T cell lines were obtained from American Type Culture Collection (ATCC; Manassas, VA, USA) and the Korean Cell Line Bank (KCLB, Seoul, Korea), respectively. The cells were cultured in DMEM (Gibco, CA, USA), containing 10% fetal bovine serum (FBS; Invitrogen, Carlsbad, CA, USA) and 1% penicillin/streptomycin (Gibco, CA, USA). The cells were maintained at 37°C in a humidified atmosphere of 95% air and 5% CO2 and screened periodically for mycoplasma contamination. Both cell lines were confirmed by DNA profiling of short tandem repeats (STRs) by the KCLB (Seoul, Korea).

13. Small interfering RNA (siRNA) transfection

siRNAs that targeted LTF and TUBB2A and AccuTarget Negative Control siRNA were purchased from Bioneer (Daejeon, Korea). The siRNA sequences for LTF and TUBB2A were as follows: siLTF-1, 5’-GAGAUCAGACACUACCUU-3’; siLTF-2, 5’-CACACUGUUGAUGUAAUGA-3’; siTUBB2A-1, ’-CUCAAGCAUGGUCUUUCA-3’; siTUBB2A-2, 5’-CACACUGUUGAUGUAAUGA-3’. Cells were transfected using Lipofectamine RNAiMAX (Invitrogen, Carlsbad, CA, USA) per the manufacturer’s instructions. After a 48-hour incubation, silencing of LTF and TUBB2A was confirmed by measuring their respective mRNA levels.

14. Cell migration and invasion assays

Quantitative cell migration and invasion were assessed using 24-well inserts (Corning Incorporated, NY, USA) with 8-μm pores according to the manufacturer’s instructions. In brief, for the transwell migration assay, transfected cells (5×104 cells) were seeded into the upper chamber, and medium that contained 10% FBS was added to the lower chamber. After a 24-hour incubation, the cells on the top of the membrane were removed using a cotton swab. The remaining migrant cells were washed with PBS, fixed in 4% paraformaldehyde, stained with 1% crystal violet for 10 min, and imaged and counted in 3 randomly selected fields under a microscope (Nikon, Tokyo, Japan). These experiments were performed in triplicate.

For the in vitro invasion assay, the upper wells of Boyden chambers were coated with 2 mg/ml of Matrigel (Corning Incorporated, NY, USA) at 37°C in a 5%

CO2 incubator for 2 hours. The cells (1×105 cells) were seeded into the upper chamber, and medium that contained 10% FBS was added to the lower chamber. The rest of the assay was performed as described above.

All experiments related to RT PCR, invasion, migration, and cell proliferation assays were performed by Dr. kyungmin Lee, Professor Han Suk Ryu’s laboratory, Department of Pathology, Seoul National University Hospital.

RESULTS

1. Construction of distant metastatic breast cancer proteomic datasets

In the pooled sample set, 9441 proteins were identified, and 7179 proteins were quantified across all samples. In the individual sample set, 8746 proteins were identified, and 6642 proteins were quantified in all samples (Figure 2 and Figure 3a). In addition, the number of identifications in each sample was calculated, resulting in a range from 7515 to 7798 identified proteins in the individual sample set and 8287 to 8309 proteins in the pooled sample set. Overall, the numbers of proteins in the samples of each sample set were similar (Figure 3b-c).

Our proteomic platform enabled us to perform an in-depth analysis of the distant metastatic breast cancer proteome, as evidenced by a dynamic range that spanned over 6 orders of magnitude (Figure 4). This comprehensive dataset included many established biomarkers for breast cancer, including the receptor tyrosine kinase erbB-2 (HER2), estrogen receptor (ESR1), progesterone receptor (PGR), and androgen receptor (AR). Notably, established protein biomarkers for metastatic breast cancer, such as EGFR, HSPD1, PRDX6, and TPM4, which are related to lymph node and regional metastasis, were also detected [50]. Moreover, this proteome encompassed most of the identified proteins in our previous study and included an additional 3757 and 3126 newly identified proteins in the pooled and individual sample sets, respectively (Figure 5) [44]. Consequently, our in-depth proteomic profiling generated a comprehensive dataset that is suitable for biomarker discovery and analysis with regard to determining the underlying mechanisms of distant metastasis in breast cancer.

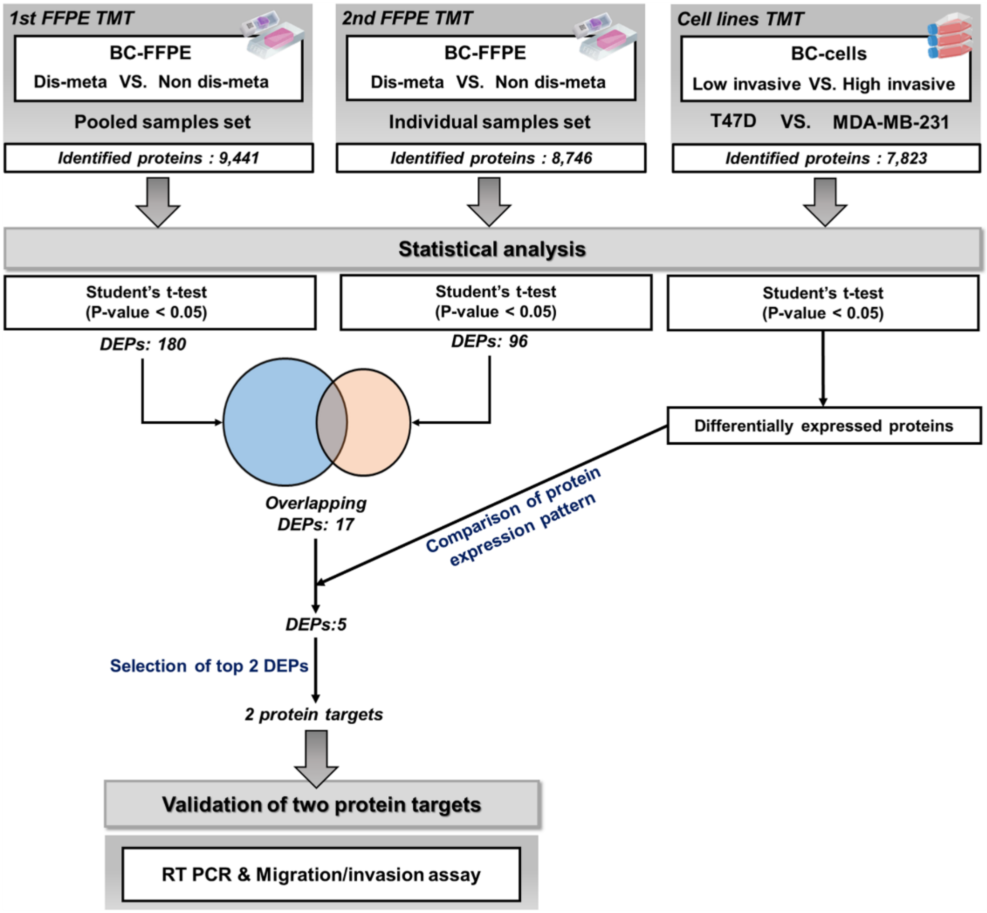


Figure 2. Schematic of overall proteomic results of the TMT-based proteomic analysis. Number of identified proteins; pooled sample set: 9441, individual sample set: 8746, and cell line set: 7823. Number of DEPs by statistical analysis and the steps for selection of protein targets. Validation phase of protein targets; real-time polymerase chain reaction (RT-PCR) and migration/invasion assay.

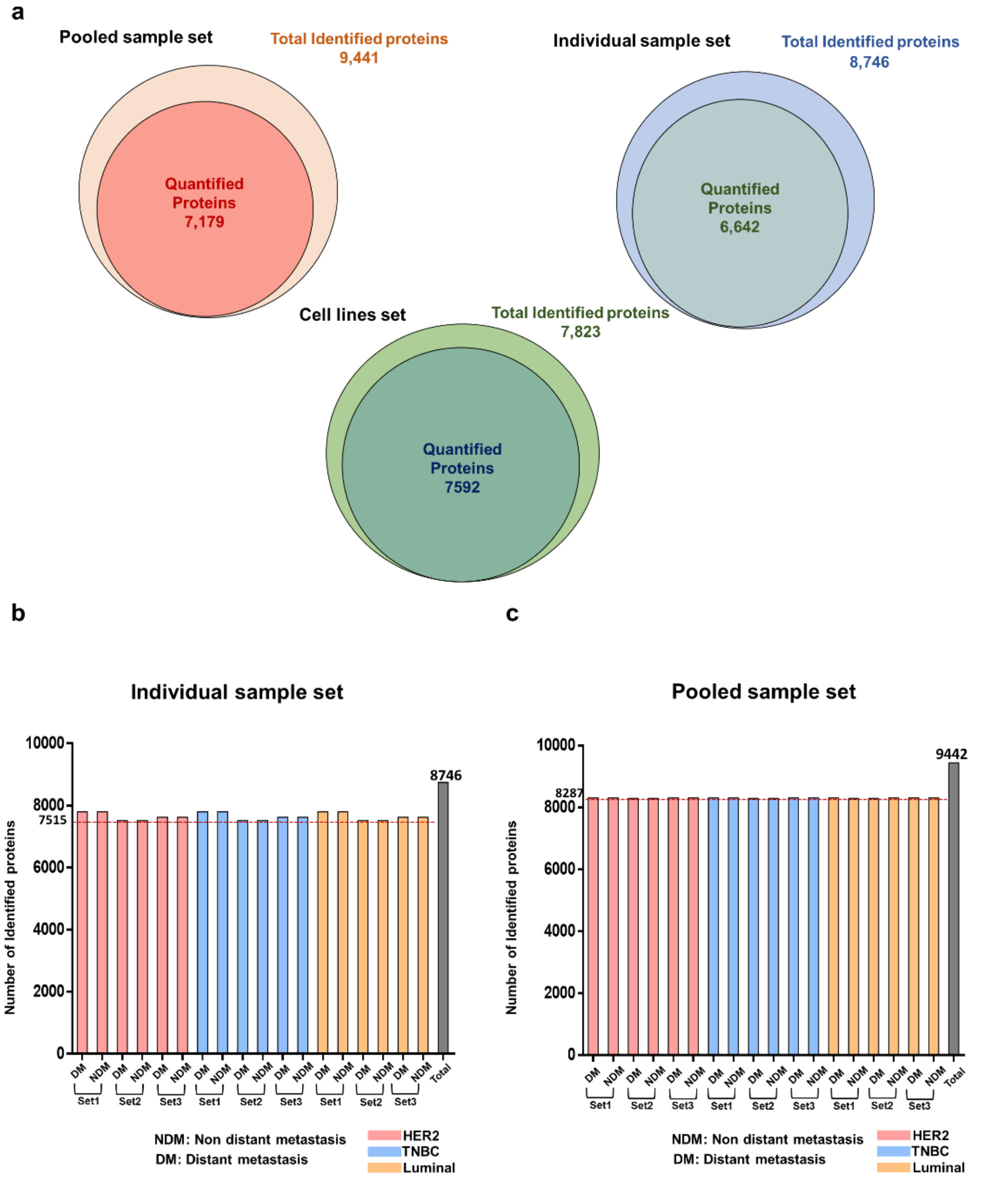


Figure 3. Identified and quantified proteins in TMT experiments. (a) The number of identified and quantified proteins in the pooled sample set, individual sample set, and cell line set. (b) The number of identified proteins in each sample of the individual sample set. (c) The number of identified proteins in each sample of

the pooled sample set.

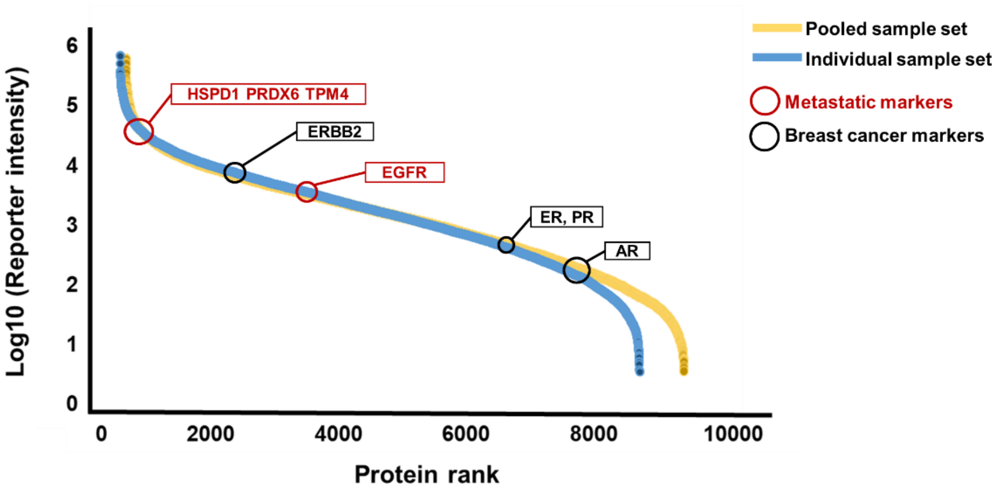


Figure 4. Dynamic ranges of protein abundance in pooled sample set and individual sample set. The dynamic range of the pooled sample set is marked in yellow, and that of the individual sample set is marked in blue. Known metastatic biomarkers are indicated in red, and breast cancer markers are marked in black.

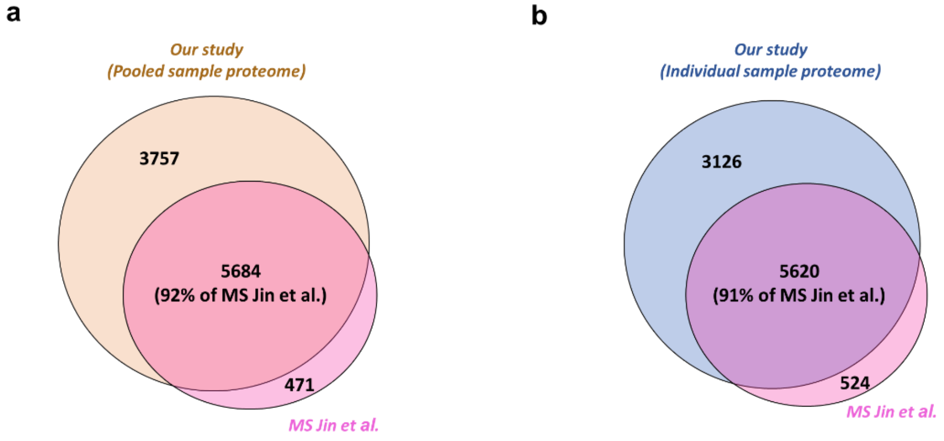
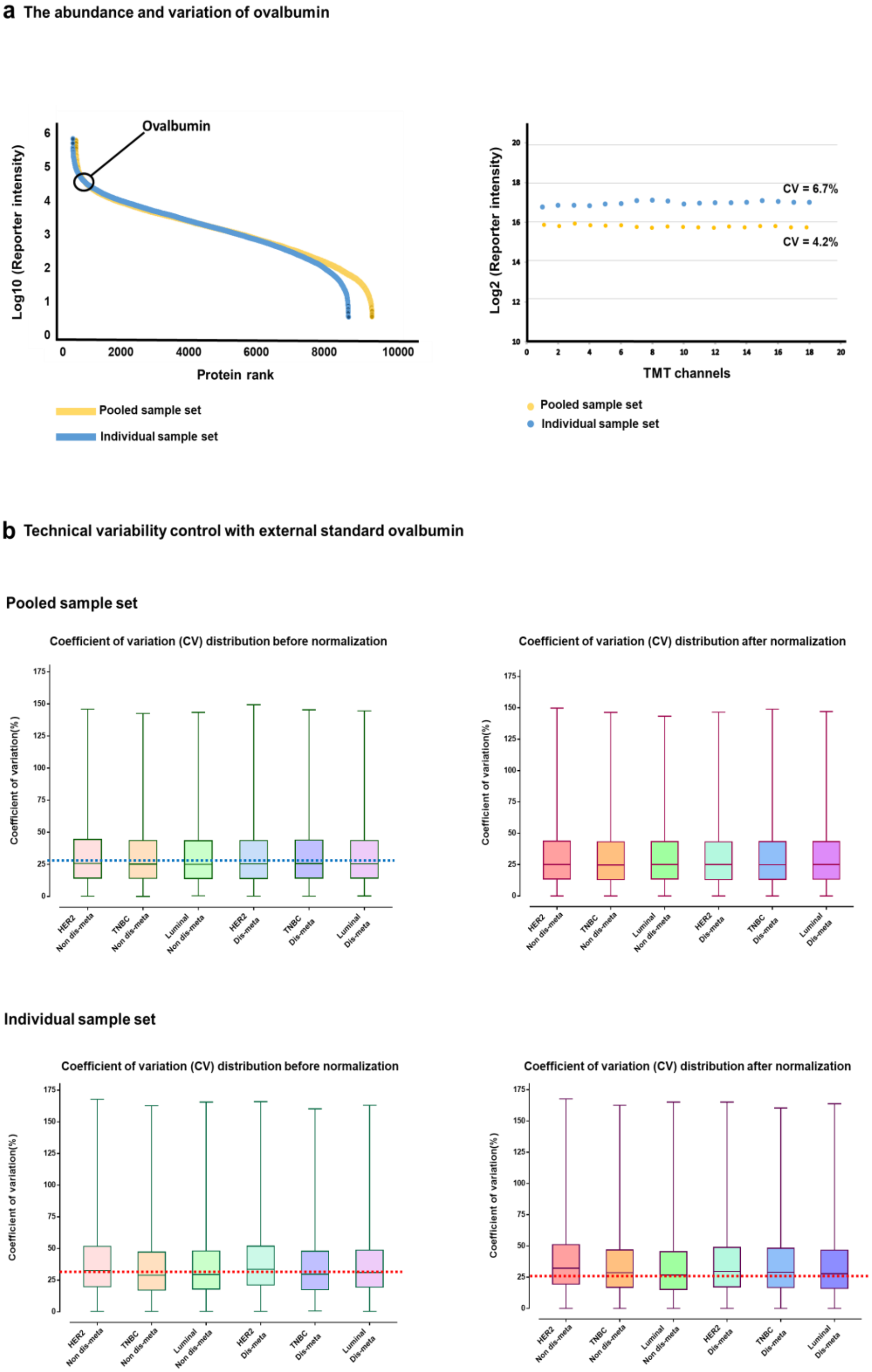


Figure 5. Comparative analysis between our FFPE tissue proteome and those of our previous studies. (a) Comparison of identified proteins between our pooled sample proteome data and those of *MS Jin et al.* (b) Comparison of identified proteins between our individual sample proteome data and those of *MS Jin et al.*

2. Quality assessment of proteomic data

The multiplexing feature of the TMT-based strategy allowed us to examine the quantitative variation within and between our samples. Interbatch and intrabatch variation was assessed using an internal standard, ovalbumin. As a result, the interbatch and intrabatch normalization produced coefficients of variation of 4.17% and 6.7% in the pooled and individual sample sets, respectively (Figure 6a). Although the variation in non-normalized intensities reflected excellent reproducibility, a slight improvement in reproducibility was observed when the levels of proteins were normalized to ovalbumin (Figure 6b-c).

Next, correlation values were calculated to assess the variation between technical replicates in the pooled sample set. MS analysis of the pooled sample set showed excellent correlation, with Pearson's correlation values ranging from 0.993 to 0.994 and averaging 0.993 (Figure 6d). In addition, the correlation between the quantitative levels of all samples was calculated to assess the variation across individual samples. MS analysis of the individual sample set revealed a wider range of correlation values than that of the pooled sample set, with Pearson's correlation values ranging from 0.647 to 0.988 and averaging 0.927 (Figure 6e). One sample, a HER2 type in the non dis-meta group, had low correlation values when paired with other individual samples, resulting in a range of 0.647 to 0.778. Slight differences in protein abundance between individual samples were observed.



a The abundance and variation of ovalbumin

b Technical variability control with external standard ovalbumin

c **Median CV values in each sample**

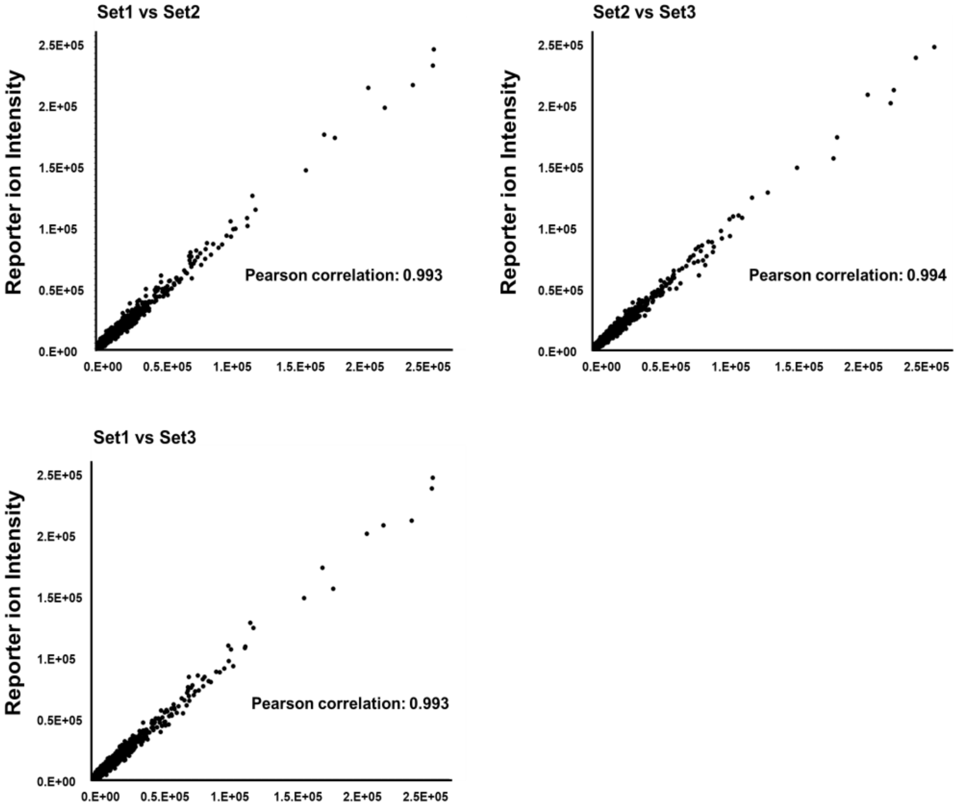
Pooled sample set

Case	HER2 Non dis-meta	TNBC Non dis-meta	Luminal Non dis-meta	HER2 Dis-meta	TNBC Dis-meta	Luminal Dis-meta	Median
No normalization	25.63	25.15	25.03	25.40	25.53	25.31	25.36
Ovalbumin normalization	25	24.76	25.06	24.99	24.87	25.15	25.00

Individual sample set

Case	HER2 Non dis-meta	TNBC Non dis-meta	Luminal Non dis-meta	HER2 Dis-meta	TNBC Dis-meta	Luminal Dis-meta	Median
No normalization	32.55	28.89	29.41	33.5	29.58	31.05	30.32
Ovalbumin normalization	32.2	28.58	26.77	29.63	28.98	27.8	28.78

d **Cross-correlation among experimental sets of pooled sample set**



Set1 vs Set2

Set2 vs Set3

Set1 vs Set3

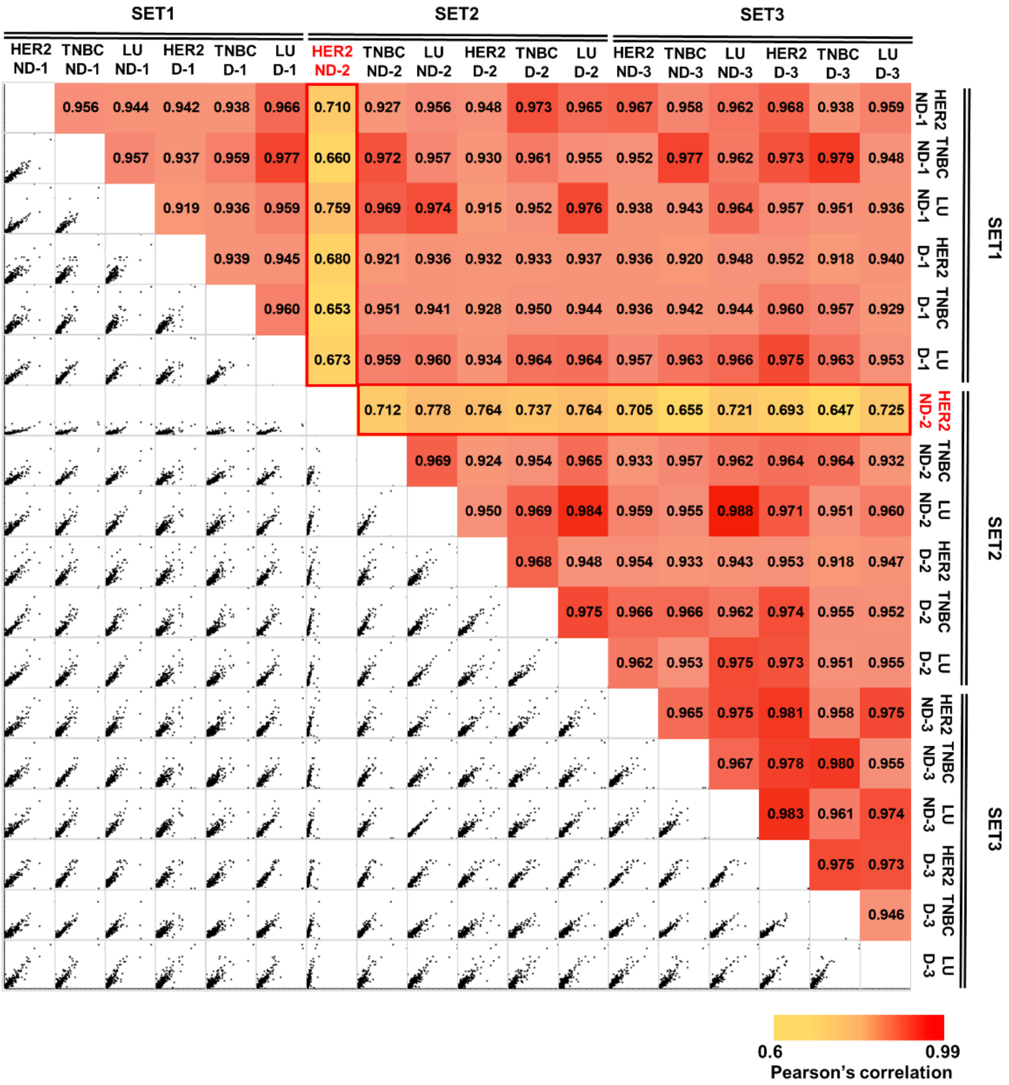

e Cross-correlation among individual samples of individual sample set

Figure 6. The quality assessment of MS analysis. (a) Abundance and technical variation of the external standard, ovalbumin. Ovalbumin was quantified in the middle-high abundance interval and had a CV of 4.2% and 6.7% in the pooled and individual sample sets in 18 TMT channels, respectively. (b), (c) The quantitative reproducibility of all proteins was improved slightly on normalization with the

external standard, ovalbumin; the median CV value of the biological replicates of the pooled and individual sample sets decreased by 0.36% and 1.54%, respectively. (d) Cross-correlation analysis using the protein levels to confirm the repeatability of the MS analyses between experimental sets of the pooled sample set. (e) Variabilities in individual samples in our MS analysis are depicted in a multiscatter plot. Reproducibility between individual samples is represented by Pearson's correlation value. Values of correlation with HER2 ND-2 are marked in red. (ND; non dis-meta, D; dis-meta, LU; luminal, - #; number of TMT set)

3. Determination of protein targets to validate distant metastatic potential

To select important protein targets to verify distant metastatic potential of breast cancer, the quantified proteins in the BC FFPE tissues datasets (i.e., the pooled and individual sample sets) were examined separately by statistical analysis. For the proteomic datasets of BC FFPE tissues, student's t-test was performed to determine differentially expressed proteins (DEPs) between the nondistant metastasis and distant metastasis groups. When a Benjamini-Hochberg false discovery rate (BH-FDR) cutoff of 0.05 was applied to the proteins in the pooled and individual sample sets respectively, however, none of the proteins in nondis-meta and dis-meta was significantly differentially expressed. Nonetheless, to determine protein targets for validation of distant metastatic breast cancer, alternative criteria were applied to the datasets.

The criteria were as follows: 1. The quantified proteins in our BC FFPE tissue proteomic datasets must pass a p-value (unadjusted for multiple comparison) cutoff of 0.05 by student's t-test for determining DEPs in nondis-meta versus dis-meta. 2. Overlapping DEPs in both BC FFPE tissue datasets were selected. 3. Overlapping DEPs that were also identified in the BC cell line proteomic dataset and demonstrated a consistent expression pattern in all 3 datasets were selected. 4. Overlapping DEPs that passed a fold-change cutoff of 1.2 were selected. 5. The most highly up-regulated and down-regulated DEPs were selected. Therefore, DEPs that satisfied all of the requirements were selected as protein targets for validation of distant metastatic potential (Figure 2).

Specifically, a total of 180 and 96 proteins were initially selected as DEPs by student's t-test (p-value < 0.05) in the pooled and individual sample sets, respectively (Figure 2). Next, overlapping proteins in DEPs of each sample set were selected.

As a result, 17 overlapping DEPs in both sets were selected. The results of the statistical analysis for these proteins are listed in Table 2. Of the 17 proteins, 5 (HSPA9, PSMB4, CTNNA1, XPO5, and PAFAH1B3) functioned in the growth, proliferation, metastasis, and recurrence of cancer [51-56]. Specifically, HSPA9 was associated with metastasis of hepatocellular carcinoma (HCC), and overexpression of HSPA9 increased the malignancy and aggressive behavior of HCC [51, 52]. Overexpression of PSMB4 increases cellular growth and the viability of breast cancer and ovarian cancer, leading to a poor prognosis [53, 54]. The deletion of CTNNA1 effects the loss of cell-to-cell adhesion, enhancing the growth and mobility of breast cancer cells [55]. XPO5 exports pre-miRNAs through the nuclear membrane to the cytoplasm and is thus important in breast cancer tumorigenesis [56]. PAFAH1B3 is a critical driver of the pathogenicity of breast cancer by inhibiting tumor-suppressing signaling lipids [72]. These 5 proteins were upregulated in our distant metastasis group, which I propose stimulate the distant metastatic potential of breast cancer.

Subsequently, I examined whether the overlapping 17 proteins were also differentially expressed in the proteomic dataset of BC cell lines, comparing less-invasive T47D and highly invasive MDA-MB-231 cells. This examination was performed to identify proteins that might have molecular features that are related to the distant metastasis of breast cancer by comparing the BC FFPE and BC cell line proteomes. Five proteins had consistent expression patterns between all proteomic datasets: tubulin beta-2A chain (TUBB2A); lactotransferrin (LTF); acyl-coenzyme a dehydrogenase, C-4 to C-12 straight chain, isoform CRA_a (ACADM); proteasome subunit beta type-4 (PSMB4); and mitotic checkpoint protein BUB3 (BUB3) (Table 1). Next, with regard to the five proteins, the fold-change in expression between nondistant metastatic and distant metastatic groups was calculated. When the fold-change cutoff was set to 1.2, two proteins were selected: LTF was the most extensively downregulated protein, whereas TUBB2A was the most highly upregulated (Figure 2 and Table 2). The normalized abundance of LTF and TUBB2A distinguished the 2 sample groups significantly (Figure 7a). Based on the criteria, LTF and TUBB2A were selected as important protein targets for validation of their function in relation to distant metastasis of breast cancer.

Table 2. Detailed statistical analysis of 17 overlapping proteins.

Protein name	Dis-meta vs non dis-meta in pooled sample set				Dis-meta vs non dis-meta in individual sample set				High invasive vs low invasive in cell lines set					
	t test significance	*p* value	Adjusted *p* value (BH FDR < 0.05)	Fold-change	*t* test significance	*p* value	Adjusted *p* value (BH FDR < 0.05)	Fold-change	*t* test significance	*p* value	Adjusted *p* value (BH FDR < 0.05)	Fold-change	Consistency of protein expression	Fold-change > 1.2
Glyceraldehyde-3-phosphate dehydrogenase	+	0.01308	1	1.211	+	0.03681	1	1.260	+	0.01580	0.02242	0.815	N	N
Tubulin beta-2A chain	+	0.01730	1	1.219	+	0.01980	1	1.298	+	0.00076	0.00173	2.329	Y	Y
Lactotransferrin	+	0.00000	9.269E-09	0.581	+	0.02619	1	0.546	+	0.00529	0.00866	0.551	Y	Y
Stress-70 protein, mitochondrial	+	0.00251	0.85696	1.114	+	0.04264	1	1.160	+	0.00003	0.00019	0.742	N	N
Catenin alpha-1	+	0.04027	1	1.137	+	0.03017	1	1.189	+	0.00000	0.00003	0.473	N	N
Bifunctional purine biosynthesis protein PURH	+	0.03371	1	1.150	+	0.03078	1	1.180	+	0.00047	0.00119	0.710	N	N
Heterogeneous nuclear ribonucleoprotein H	+	0.04529	1	1.046	+	0.02779	1	1.111	N/D	N/D	N/D	N/D	N/D	N
Isoform 2 of Multifunctional protein ADE2	+	0.01657	1	1.149	+	0.02412	1	1.147	+	0.00184	0.00355	0.742	N	N
ADP/ATP translocase 3	+	0.01062	1	0.827	+	0.04718	1	0.833	+	0.00022	0.00068	1.416	N	N
Exportin-5	+	0.00720	1	1.177	+	0.03871	1	1.227	+	0.00359	0.00619	0.868	N	N
RNA-binding protein 39	+	0.04237	1	1.074	+	0.02435	1	1.119	+	0.00171	0.00334	0.916	N	N
Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain, isoform CRA_a	+	0.04319	1	0.922	+	0.01636	1	0.834	+	0.00049	0.00124	0.889	Y	N
Proteasome subunit beta type-4	+	0.00136	0.60958	1.111	+	0.03592	1	1.143	+	0.00066	0.00155	1.213	Y	N
Beta-glucuronidase	+	0.00266	0.83128	0.643	+	0.02986	1	0.630	N/D	N/D	N/D	N/D	N/D	N
Mitotic checkpoint protein BUB3	+	0.00728	1	1.113	+	0.04702	1	1.112	+	0.03503	0.04592	1.057	Y	N
Platelet-activating factor acetylhydrolase IB subunit gamma	+	0.04161	1	1.221	+	0.02117	1	1.266	+	0.01903	0.02649	0.850	N	N
2-hydroxyacyl-CoA lyase 1	+	0.01337	1	1.258	+	0.04677	1	1.314	+	0.00001	0.00010	0.575	N	N

(N/D- not detection, N- no, Y- yes)

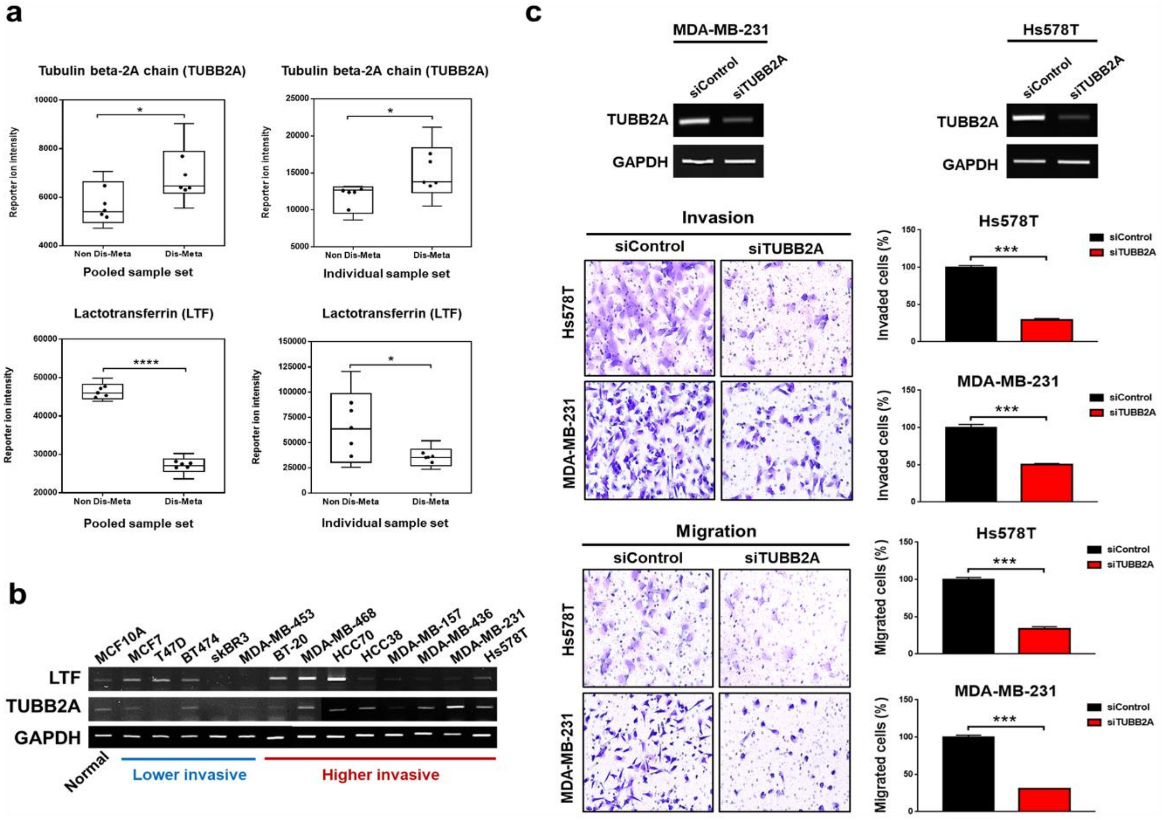


Figure 7. Validation of TUBB2A and LTF as protein targets. (a) Protein expression patterns of TUBB2A and LTF by mass spectrometry; expression pattern of reporter ion intensity of TUBB2A (upper panel) and LTF (lower panel) in pooled sample set (left panel) and individual sample set (right panel), respectively. The data in the interquartile range are displayed as black dots (* < p-value 0.05; **** < p-value 0.0001). (b) Expression patterns of TUBB2A and LTF in various breast cancer cell lines by RT PCR. Higher expression levels are lighter than lower levels (red line; higher invasive BC cell lines, blue line; lower invasive BC cell lines). (c) Results of invasion and migration assays for TUBB2A using Hs578T and MDA-MB-231 BC cell lines. RT-PCR of TUBB2A, downregulated by siRNA transfection in both cell lines (upper panel). Images of invading and migrating cells (lower left panel) and percentage (%) of invading and migrating cells (lower right panel) (*** < p-value 0.001). The RT-PCR, invasion, and migration assays were performed by Dr. kyungmin Lee, Professor Han Suk Ryu’s laboratory, Department of Pathology, Seoul National University Hospital.

4. Expression levels of TUBB2A and LTF verified by RT-PCR

The difference in the expression of TUBB2A and LTF was validated by RT-PCR in 1 normal breast cell line and 13 breast cancer cell lines, the relative invasiveness of which was determined per other studies [74-81]. The expression of LTF was lower in the higher invasive group than in the lower invasive group, except in 3 cell lines (BT20, MDA-MB-368, and HCC70). In particular, HCC70 expressed the most LTF (Figure 7b). The level of TUBB2A was generally higher in the higher invasive group compared with the lower invasive group. Specifically, MDA-MB-231 had the highest expression of TUBB2A (Figure 7b). The expression level of TUBB2A by MS was consistent with that by RT-PCR. The patterns of LTF by MS were not consistent with the RT-PCR results.

5. Distant metastatic potential of TUBB2A

The correlation between TUBB2A and metastatic characteristics was validated by invasion and migration assay. Two highly invasive BC cell lines (Hs578T and MDA-MB-231) were used to examine invasion and migration, based on the levels of TUBB2A. As a result, by siRNA transfection, TUBB2A was downregulated in both cell lines by RT-PCR. The number of invading cells fell significantly by over 50% when TUBB2A was knocked down compared with the control group (siControl), as did the number of migrating cells (Figure 7c). Conversely, because the relative cell proliferation did not differ significantly on the day when the invasion and migration assays were conducted (Figure 8), the decreased invasiveness of the cells did not result from the altered cell proliferation. Thus, the distant metastatic potential of TUBB2A was verified, independent of the influence of cell proliferation.

To determine the ability of TUBB2A as a novel protein biomarker candidate of distant metastatic breast cancer, its performance was evaluated in the individual sample set. The sensitivity, specificity, and positive predictive value (PPV) by receiver operating characteristic

(ROC) analysis were 78%, 100%, and 88%, respectively. Furthermore, the area under curve (AUC) value was 0.852, based on the ROC curve, and the threshold value, expressed as reporter ion intensity, that corresponded to the highest Youden's index was 13,178 (Figure 9). Based on these results, we expected TUBB2A to perform well in the diagnosis and prediction of distant metastatic breast cancer.

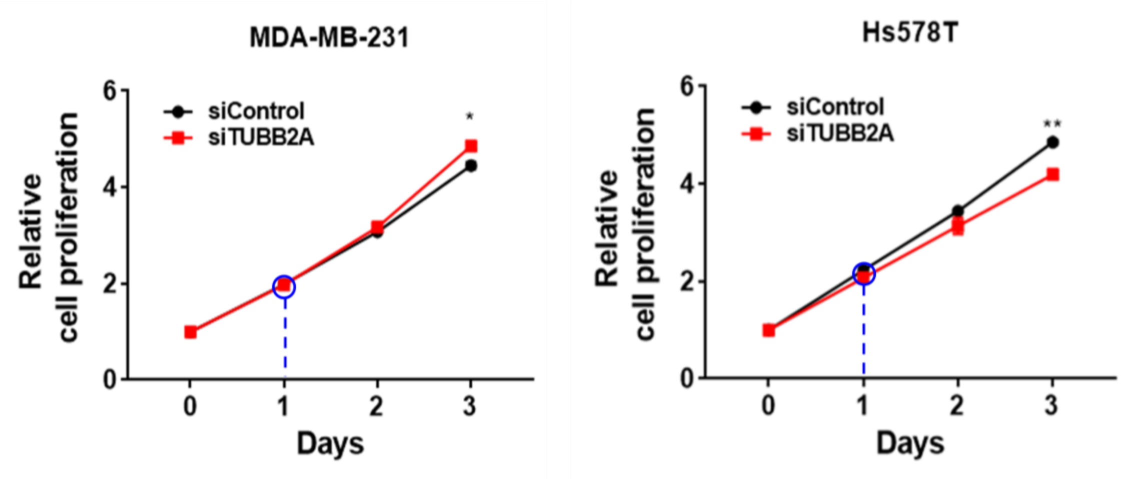
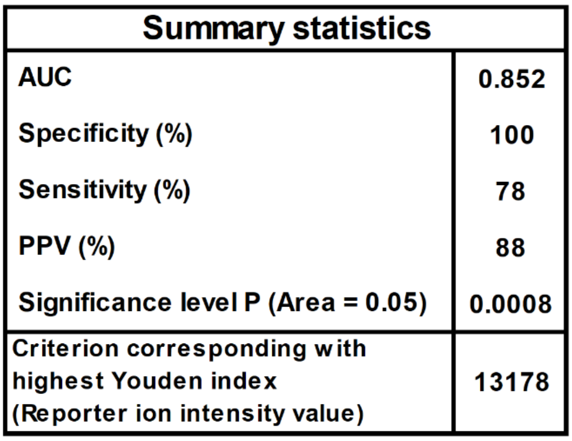


Figure 8. Cell proliferation of MDA-MB-231 and Hs578T cell lines. Relative cell proliferation was observed for 3 days, when TUBB2A was knocked down compared with the control group (siControl) (* < p-value 0.05; ** < p-value 0.01). The time point at which migration and invasion assays were performed is indicated in blue circle. Cell proliferation assay was performed by Dr. kyungmin Lee, Professor Han Suk Ryu's laboratory, Department of Pathology, Seoul National University Hospital.



Summary statistics	
AUC	0.852
Specificity (%)	100
Sensitivity (%)	78
PPV (%)	88
Significance level P (Area = 0.05)	0.0008
Criterion corresponding with highest Youden index (Reporter ion intensity value)	13178

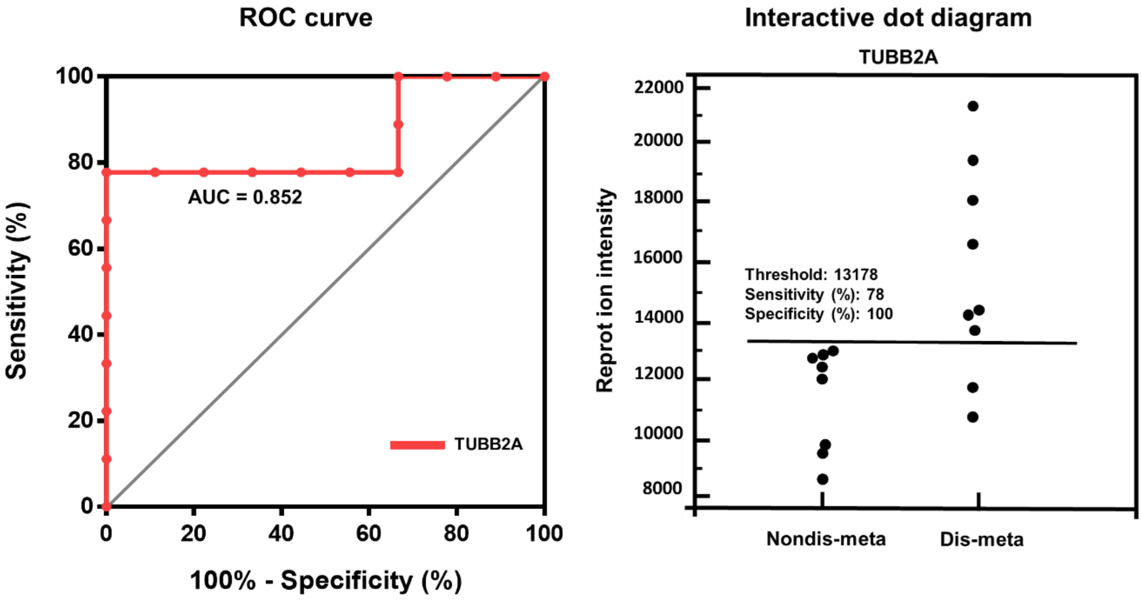


Figure 9. Performance of the novel biomarker TUBB2A in the individual sample set. Table of summary statistics in ROC analysis, ROC curve with AUC = 0.852, and interactive dot diagram with sensitivity = 78%, specificity = 100%, and reporter ion intensity threshold = 13,178.

6. Biological functions of distant metastatic breast cancer

To examine the functional signatures of distant metastatic breast cancer, I performed a bioinformatics analysis using 259 DEPs from the 2 sample sets. By gene ontology (GO) enrichment analysis, the 177 upregulated proteins in the distant metastasis group were assigned to various biological processes, such as cell-cell adhesion, proteolysis during cellular protein catabolism, NIK/NK-kappa B signaling, microtubule-based processes, and retrograde vesicle-mediated transport,- Golgi-to-ER (Fisher's exact test p-value < 0.05) (Figure 10a and Table 3). The most significant biological process in upregulated proteins was the regulation of mRNA stability (p-value = 7.82E-07). Conversely, the 82 downregulated proteins were allocated to various biological processes, including oxidation-reduction, organization of actin cytoskeleton, response to hydrogen peroxide, thrombin receptor signaling, sequestering of actin monomers, and positive regulation of toll-like receptor 4 signaling (Fisher's exact test p-value < 0.05) (Figure 10b and Table 3). The most significant biological process in downregulated proteins was oxidation-reduction (p-value = 2.89E-04).

In the enrichment of biological functions and pathways, the 259 DEPs were assigned to 6 canonical pathways and 11 downstream biological functions (Fisher's exact test p-value < 0.05, and Z-score > 1). Canonical pathways included acute phase response signaling, ILK signaling, actin cytoskeletal signaling, leukocyte extravasation signaling, and tRNA charging (Figure. 11a, and Table 4). The most significant and activated canonical pathway was glycolysis I (p-value = 1.74E-06, and activation Z-score = 2.236). Biological functions included polarization of tumor cell lines, orientation of cells, adhesion of BC cell lines, binding of NFkB sites, glycolysis in tumor cell lines, and proliferation of tumor/carcinoma cell lines (Figure. 11b, and Table 4). The most significant and activated biological function was cell proliferation of tumor cell lines (p-value = 1.69E-08, and activation Z-score = 2.451). Based on our results, I propose that the interaction of various biological functions induces distant metastatic breast cancer.

Of the 2 protein targets, the result showed that the TUBB2A has association with the proliferation of tumor/carcinoma cell lines, microtubule-based processes, epithelial adherens junction signaling, 14-3-3-mediated signaling, and phagosome maturation. The most significant function of TUBB2A was cell proliferation of tumor cell lines (p-value = 1.69E-08). LTF was involved in the binding of NFkB sites, negative regulation of apoptotic process, positive regulation of I-KappaB kinase/NF-kappaB signaling, negative regulation of ATPase activity, and positive regulation of toll-like receptor 4 signaling pathway. Binding of NFkB sites was the most significant function (p-value = 2.17E-04) (Figure 12 and Table 5). Thus, these candidates had distinct and independent biological characteristics.

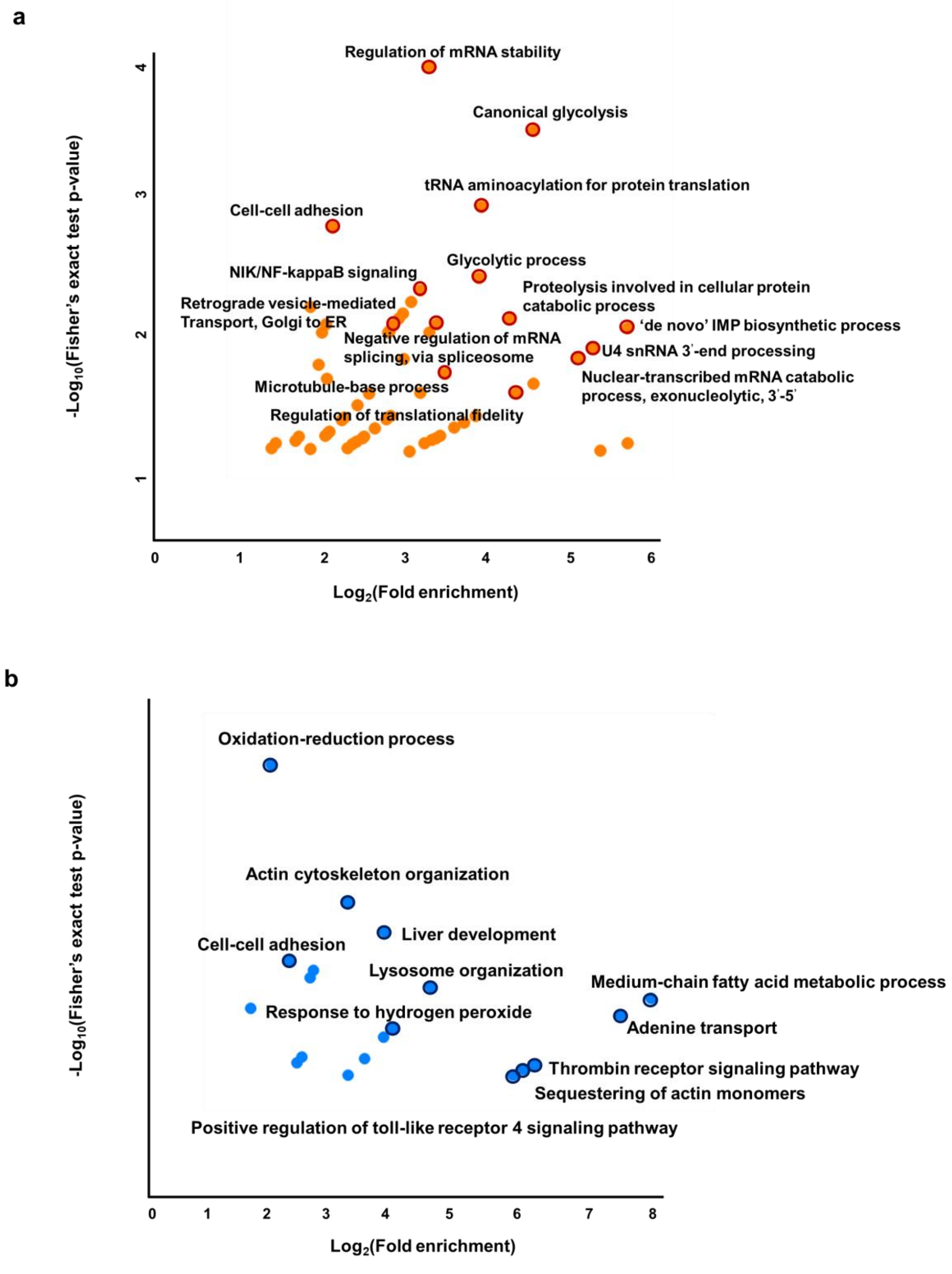


Figure 10. Gene ontology analysis of all 259 DEPs in the two sample sets using The Database for Annotation, Visualization and Integrated Discovery (DAVID). (a) Biological process terms of 177 upregulated DEPs. (b) Biological process terms of 82 downregulated DEPs (Fisher's exact test p-value < 0.05).

Table 3. GO analysis using the DAVID bioinformatics tool. Biological processes of upregulated DEPs by student's t-test are listed. The p-value (modified Fisher exact p-value) cutoff for the GO annotation was set to < 0.05. Genes that were associated with each GO term are represented as official gene symbols. 'GO direct' filters extensive GO terms, based on the measured specificity of each term.

Category	Terms	Count	%	P-Value	Genes	Fold Enrichment
GOTERM_BP_DIRECT	GO:0043488~regulation of mRNA stability	10	5.682	7.82E-07	P20618, P28074, P08107, Q13868, Q5RKV6, P28070, Q06265, P31946, P49720, Q99436	9.821
	GO:0061621~canonical glycolysis	6	3.409	4.80E-06	P04406, P08237, P52789, P04075, P06733, P14618	23.344
	GO:0006418~tRNA aminoacylation for protein translation	6	3.409	4.30E-05	P07814, Q12904, Q9P2J5, P26639, P41252, P54136	15.173
	GO:0098609~cell-cell adhesion	12	6.818	7.75E-05	Q9H4G0, P31947, P06733, P04075, P31946, P31939, E9PL19, P54136, P08107, P22234, P62258, Q9UHX1, P14618	4.479
	GO:0006096~glycolytic process	5	2.841	3.31E-04	P04406, P08237, P52789, P04075, P06733	14.876
	GO:0038061~NIK/NF-kappaB signaling	6	3.409	4.75E-04	P20618, P28074, P28070, P61081, P49720, Q99436	9.196
	GO:0051436~negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	6	3.409	6.65E-04	P20618, P28074, O43684, P28070, P49720, Q99436	8.548
	GO:0016032~viral process	11	6.250	7.60E-04	P20618, P28074, Q16531, P28070, P62258, Q07021, P20340, P31946, Q8TAE8, P49720, Q99436	3.721
	GO:0051437~positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition	6	3.409	9.08E-04	P20618, P28074, O43684, P28070, P49720, Q99436	7.986
	GO:0031145~anaphase-promoting complex-dependent catabolic process	6	3.409	0.00108	P20618, P28074, O43684, P28070, P49720, Q99436	7.683
	GO:0048025~negative regulation of mRNA splicing, via spliceosome	4	2.273	0.00109	Q15287, P26368, X6RAL5, Q07021	19.268
	GO:0051603~proteolysis involved in cellular protein catabolic process	5	2.841	0.00125	P20618, P28074, P28070, P49720, Q99436	10.537
	GO:0006364~rRNA processing	9	5.114	0.00127	P42285, Q13868, Q5RKV6, Q9NWS0, Q9UQ80, Q06265, P78346, P62244, Q9Y5J1	4.254
	GO:0006890~retrograde vesicle-mediated transport, Golgi to ER	6	3.409	0.00128	P42858, Q9BVK6, Q8WVM8, P20340, Q9H0N0, O43264	7.402
	GO:0006189~'de novo' IMP biosynthetic process	3	1.705	0.00140	P22234, P22102, P31939	50.578
	GO:0070125~mitochondrial translational elongation	6	3.409	0.00150	Q9Y2Q9, Q9Y3B7, Q9BYD3, Q8TAE8, Q9NRX2, Q9BYC9	7.140
	GO:0006521~regulation of cellular amino acid metabolic process	5	2.841	0.00157	P20618, P28074, P28070, P49720, Q99436	9.917
	GO:0070126~mitochondrial translational termination	6	3.409	0.00158	Q9Y2Q9, Q9Y3B7, Q9BYD3, Q8TAE8, Q9NRX2, Q9BYC9	7.057
	GO:0000398~mRNA splicing, via spliceosome	9	5.114	0.00160	Q15287, P42285, B8ZZ98, P26368, E9PB61, P62318, G8JLB6, P09234, Q15393	4.101
	GO:0034475~U4 snRNA 3'-end processing	3	1.705	0.00258	Q13868, Q5RKV6, Q06265	37.934
	GO:0034427~nuclear-transcribed mRNA catabolic process, exonucleolytic, 3'-5'	3	1.705	0.00330	Q13868, Q5RKV6, Q06265	33.719
	GO:0002479~antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	5	2.841	0.00341	P20618, P28074, P28070, P49720, Q99436	8.028
	GO:0043161~proteasome-mediated ubiquitin-dependent protein catabolic process	8	4.545	0.00399	P20618, P28074, Q16531, O43684, P28070, P49720, P61077, Q99436	3.986
	GO:0007017~microtubule-based process	4	2.273	0.00524	Q71U36, Q13885, P43034, Q08426	11.240
	GO:0008380~RNA splicing	7	3.977	0.00600	Q15287, Q14498, Q9UHX1, X6RAL5, P62318, Q07021, Q15393, E9PL19	4.266
	GO:0009168~purine ribonucleoside monophosphate biosynthetic process	3	1.705	0.00697	P22234, P22102, P31939	23.344
	GO:0006094~gluconeogenesis	4	2.273	0.00919	P04406, P04075, P06733, P00505	9.196
	GO:0006635~fatty acid beta-oxidation	4	2.273	0.00919	P51659, O14975, Q08426, Q15067	9.196
	GO:0006450~regulation of translational fidelity	3	1.705	0.00927	Q9BTE6, Q9P2J5, P41252	20.231
	GO:0051897~positive regulation of protein kinase B signaling	5	2.841	0.00943	P52895, P36222, P10599, Q07021, P24593	6.021
	GO:0060071~Wnt signaling pathway, planar cell polarity pathway	5	2.841	0.01287	P20618, P28074, P28070, P49720, Q99436	5.498
	GO:0006749~glutathione metabolic process	4	2.273	0.01766	P0CG29, P82970, P09488, O43708	7.225
	GO:0021987~cerebral cortex development	4	2.273	0.01766	P22102, P62258, P43034, P31939	7.225
	GO:0000132~establishment of mitotic spindle orientation	3	1.705	0.01783	P42858, P43034, O43264	14.451
	GO:0007062~sister chromatid cohesion	5	2.841	0.01875	O43684, O75122, P43034, Q99623, O43264	4.911
	GO:0001649~osteoblast differentiation	5	2.841	0.01935	P51659, P35232, E9PB61, P41252, P24593	4.863

GO:0051259~protein oligomerization	4	2.273	0.01938	P08237, Q9UJ83, Q13263, Q32Q12	6.976
GO:0002223~stimulatory C-type lectin receptor signaling pathway	5	2.841	0.01997	P20618, P28074, P28070, P49720, Q99436	4.817
GO:0000715~nucleotide-excision repair, DNA damage recognition	3	1.705	0.02121	Q16531, Q99627, P09874	13.194
GO:0015949~nucleobase-containing small molecule interconversion	3	1.705	0.02483	P10599, Q32Q12, P15531	12.139
GO:0006369~termination of RNA polymerase II transcription	4	2.273	0.02508	Q15287, P26368, E9PB61, P62318	6.322
GO:0006461~protein complex assembly	5	2.841	0.02756	P07814, Q9Y697, J3KNA0, Q15393, O43264	4.360
GO:0033209~tumor necrosis factor-mediated signaling pathway	5	2.841	0.02910	P20618, P28074, P28070, P49720, Q99436	4.286
GO:0090263~positive regulation of canonical Wnt signaling pathway	5	2.841	0.03069	P20618, P28074, P28070, P49720, Q99436	4.215
GO:0043388~positive regulation of DNA binding	3	1.705	0.03069	P10599, Q13263, P15531	10.838
GO:0098869~cellular oxidant detoxification	4	2.273	0.03158	P10599, P11678, Q7LBC6, O43708	5.780
GO:0006397~mRNA processing	6	3.409	0.03196	Q14498, Q9UHX1, P26368, X6RAL5, Q07021, Q15393, E9PL19	3.391
GO:0000226~microtubule cytoskeleton organization	4	2.273	0.03275	P04406, Q66K74, O75122, P43034	5.699
GO:0043928~exonucleolytic nuclear-transcribed mRNA catabolic process involved in deadenylation-dependent decay	3	1.705	0.03276	Q13868, Q5RKV6, Q06265	10.464
GO:0006754~ATP biosynthetic process	3	1.705	0.03276	P04075, P14618, P36542	10.464
GO:0045739~positive regulation of DNA repair	3	1.705	0.03487	Q9NQ88, P12004, Q13263	10.116
GO:1900740~positive regulation of protein insertion into mitochondrial membrane involved in apoptotic signaling pathway	3	1.705	0.03487	P62258, P31947, P31946	10.116
GO:0000209~protein polyubiquitination	6	3.409	0.03532	P20618, P28074, P28070, P49720, P61077, Q99436	3.299
GO:0007030~Golgi organization	4	2.273	0.03636	P42858, Q9BVK6, O75122, O43264	5.468
GO:0006412~translation	7	3.977	0.03852	Q9BRX2, P26639, Q9Y3B7, Q9BYD3, Q9NRX2, P62244, Q9BYC9	2.799
GO:0002762~negative regulation of myeloid leukocyte differentiation	2	1.136	0.03873	Q32Q12, P15531	50.578
GO:0061024~membrane organization	3	1.705	0.03926	P62258, P31947, P31946	9.483
GO:0007005~mitochondrion organization	4	2.273	0.04018	P35232, P09874, Q8WVM0, Q99623	5.255
GO:0042060~wound healing	4	2.273	0.04419	P02751, P31431, P21860, Q5JRA6	5.058
GO:0000165~MAPK cascade	7	3.977	0.04497	P20618, P28074, P28070, P21860, P31946, P49720, Q99436	2.703
GO:0050821~protein stabilization	5	2.841	0.04525	P04406, P35232, P08107, P17987, Q99623	3.719
GO:0071051~polyadenylation-dependent snoRNA 3'-end processing	2	1.136	0.04818	Q13868, Q5RKV6	40.463
GO:0071038~nuclear polyadenylation-dependent tRNA catabolic process	2	1.136	0.04818	Q13868, Q06265	40.463
GO:0009113~purine nucleobase biosynthetic process	2	1.136	0.04818	P22234, P22102	40.463
GO:0010035~response to inorganic substance	2	1.136	0.04818	P22102, P31939	40.463
GO:0042769~DNA damage response, detection of DNA damage	3	1.705	0.04862	Q16531, P09874, P12004	8.430

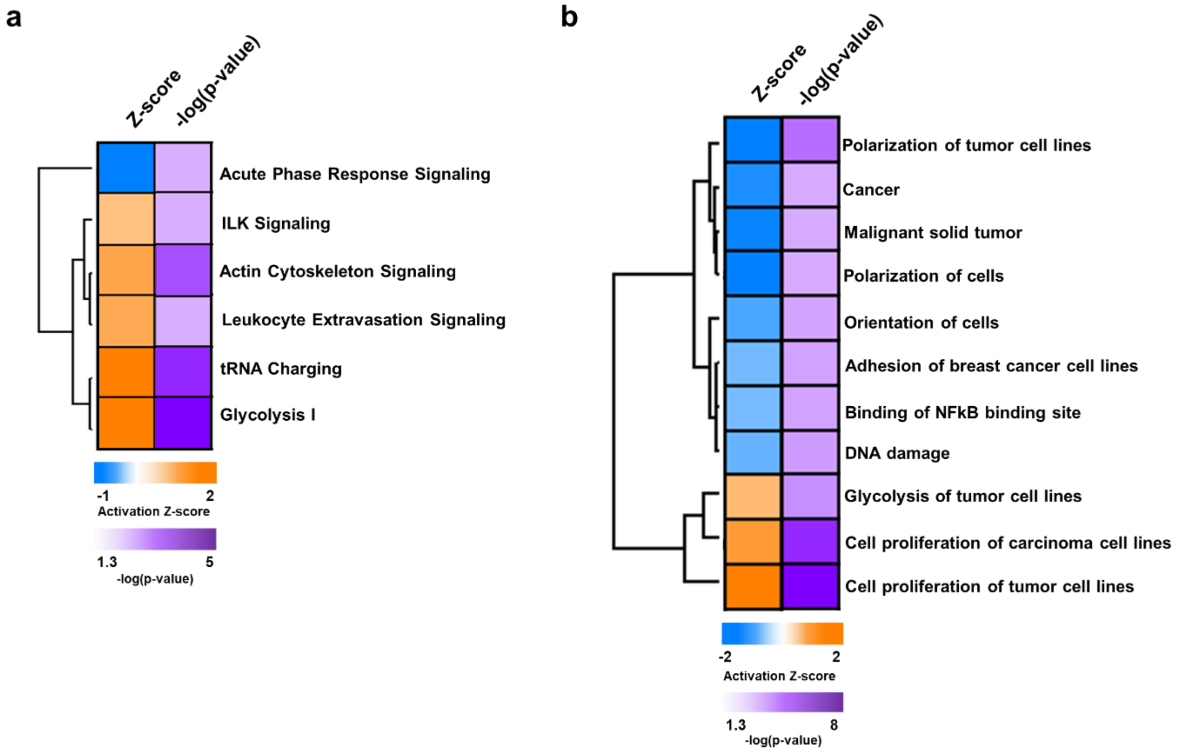


Figure 11. IPA analysis of total 259 proteins that were sum of DEPs in the two sample sets on canonical pathway and downstream biological functions. (a) Canonical pathway enrichment of all 259 DEPs in the two sample sets. (b) Hierarchical clustering of downstream biological functions assessed by IPA using the 259 DEPs. The significant pathways, and downstream biological functions (Fisher's exact test p-value <0.05) were deduced using Ingenuity Pathway Analysis (IPA), and their activation and inhibition states are expressed as Z-score.

Table 4. Downstream biological functions and canonical pathways of DEPs by student t-test by IPA analysis. Downstream biological functions were examined using the IPA informatics tool. The p-value cutoff was set to < 0.05, and the activation Z-score was set to > 1. Proteins in each biological function and pathway are listed. P-values and Z-scores of biological functions and pathways are shown.

Categories	Disease and function annotation	-log(P value)	Activation z-score	Molecules	# Molecules
Cellular Development,Cellular Growth and Proliferation	Cell proliferation of tumor cell lines	7.772	2.451	ACTN4,ALOX15B,ANO1,BUB3,C1QBP,CNDP2,COPS8,CYP1B1,DCLK1,ERBB3,FKBP5,FN1,GAPDH,GGT1,GSTM1,HPGD,HTT,IGFBP5,IQGAP2,ITGA3,LGALS3BP,MCM2,MUCL1,MYH14,NAMPT,NDUFAF2,NFS1,NME1,PA2G4,PAFAH1B1,PARP1,PCNA,PKM,PPIF,PSMB7,PTPRF,RAC1,SDC1,SDC4,SFN,SLC25A6,SOD2,TGM2,TMSB10/TMSB4X,TRIM28,TRIO,TUBB2A,TXN,USP9X	49
DNA Replication, Recombination, and Repair	DNA damage	6.542	-1.188	BUB3,FN1,GAPDH,HTT,IDH1,MCM2,PARP1,PCNA,SOD2,USP9X,YWHAE	11
Cell Death and Survival	Apoptosis of neuroblastoma cell lines	5.218	-1.029	ENO1,HTT,IGFBP5,IKBKB,PARP1,TGM2,TXN,UTP18	8
Cell Morphology	Orientation of cells	4.872	-1.408	ACTN4,FN1,ITGA3,NAMPT,PDLIM1,RAC1,SDC4	7
Cell-To-Cell Signaling and Interaction	Interaction of colorectal cancer cell lines	4.129	2.213	ERBB3,FN1,ITGA3,NME1,PKM,RAC1	6
Cell Morphology	Polarization of cells	4.012	-2.236	ACTN4,FN1,ITGA3,NAMPT,PDLIM1,RAC1	6
Cell-To-Cell Signaling and Interaction	Adhesion of colorectal cancer cell lines	3.869	2.219	FN1,ITGA3,NME1,PKM,RAC1	5
Cellular Movement	Migration of pancreatic cancer cell lines	3.798	1.109	CNDP2,FN1,PPIF,RAC1,SFN	5
Cell Cycle,Gene Expression	Binding of NFkB binding site	3.663	-1.067	FN1,LTF,RAC1,SOD2,TRIM28	5
Cellular Development,Cellular Growth and Proliferation	Cell proliferation of carcinoma cell lines	3.500	1.582	C1QBP,ERBB3,GAPDH,HPGD,LGALS3BP,MYH14,NAMPT,NDUFAF2,PKM,PTPRF,RAC1,SLC25A6,TGM2,TRIM28,TUBB2A	15
Cell-To-Cell Signaling and Interaction	Adhesion of melanoma cell lines	3.493	-1.067	ALCAM,ALOX15B,ERBB3,NME1	4
Cell Morphology	Polarization of tumor cell lines	3.493	-2	ACTN4,ITGA3,PDLIM1,RAC1	4
Cancer,Organismal Injury and Abnormalities	Cancer	3.444	-1.771	AACS,ACADM,ACOX1,ACSM1,ACTN4,AIMP1,AKR1A1,AKR1C1/AKR1C2,ALCAM,ALDOA,ALOX15B,ANO1,ANXA4,APOD,APOL3,ARFRP1,ARHGAP1,ATIC,BCAM,BPNT1,BST1,BUB3,C1QBP,CALB2,CAPS,CEP152,CHI3L1,CLSTN2,CNDP2,COL2A1,COPS8,CORO1B,CRABP1,CROT,CTNNA1,CTSZ,CXCL17,CYP1B1,CYP4X1,DCLK1,EHHADH,EML2,ENO1,EPB41L1,EPRS,EPX,ERBB3,ERP29,ETF1,EXOSC6,FABP7,FKBP5,FN1,GALM,GAPDH,GART,GGT1,GOT2,GSTM1,GSTT2/GSTT2B,GUSB,HAAO,HACL1,HAGHL,HBD,HEXB,HLA-DRB1,HMGN5,HNRNPH1,HPGD,HSD17B4,HSPA4,HSPA9,HTT,IARS,IDH1,IGFBP5,IKBKB,IQGAP2,ISYNA1,ITGA3,KRT15,LANCL1,LARS,LBP,LGALS3BP,LOXL2,LRBA,LTF,LYPLA1,MAP1S,MBLAC2,MCM2,MCM6,MIA3,MTHFD1,MUCL1,MYH14,NAMPT,NANS,NDUFAF2,NFS1,NME1,NME1-NME2,OPTN,PA2G4,PAFAH1B1,PAFAH1B3,PAICS,PARP1,PCNA,PDIA4,PDLIM1,PDLIM5,PEA15,PELO,PFDN4,PFKM,PGLS,PKM,PLIN3,PPIF,PRRC1,PSMB4,PSMB7,PTPRF,RAB6C/RAB6D,RAC1,RARS,RBM39,RCN2,SAP18,SCARB2,SCFD1,SCP2,SDC1,SDC4,SDR16C5,SEC11C,SEC23B,SELENBP1,SF3B3,SFN,SLC25A1,SLC25A6,SLC27A2,SOD2,SRP54,STAU2,SUSD2,TARS,TGM2,TM7SF2,TMED9,TMSB10/TMSB4X,TRIM28,TRIO,TRIP13,TUBA1A,TUBB2A,TXN,USP14,USP9X,UTP18,VWA5A,XPO5,YBX3,YWHAB,YWHAE,ZW10	170
Carbohydrate Metabolism,Cellular Function and Maintenance	Glycolysis of tumor cell lines	3.284	1.091	C1QBP,IKBKB,PFKM,PKM	4
Cell-To-Cell Signaling and Interaction	Adhesion of breast cancer cell lines	3.204	-1.091	ALOX15B,ERBB3,ERP29,FN1,NME1	5

Cancer,Organismal Injury and Abnormalities	Extracranial solid tumor	3.169411331	-1.201	AACS,ACADM,ACOX1,ACSM1,ACTN4,AIMP1,AKR1A1,AKR1C1/AKR1C2,ALCAM,ALDOA,ALOX15B,ANO1,ANXA4,APOD,APOL3,ARFRP1,ARHGAP1,ATIC,BCAM,BPNT1,BST1,BUB3,C1QBP,CALB2,CAPS,CEP152,CHI3L1,CLSTN2,CNDP2,COL2A1,COPS8,CORO1B,CRABP1,CROT,CTNNA1,CTSZ,CXCL17,CYP1B1,CYP4X1,DCLK1,EHHADH,EML2,ENO1,EPB41L1,EPRS,EPX,ERBB3,ERP29,ETF1,EXOSC6,FABP7,FKBP5,FN1,GALM,GAPDH,GART,GGT1,GOT2,GSTM1,GSTT2/GSTT2B,GUSB,HAAO,HACL1,HAGHL,HBD,HEXB,HLA-DRB1,HMGN5,HNRNPH1,HPGD,HSD17B4,HSPA4,HSPA9,HTT,IARS,IDH1,IGFBP5,IKBKB,IQGAP2,ISYNA1,ITGA3,KRT15,LANCL1,LARS,LBP,LGALS3BP,LOXL2,LRBA,LTF,LYPLA1,MAP1S,MBLAC2,MCM2,MCM6,MIA3,MTHFD1,MUCL1,MYH14,NAMPT,NANS,NDUFAF2,NFS1,NME1,NME1-NME2,OPTN,PA2G4,PAFAH1B1,PAFAH1B3,PAICS,PARP1,PCNA,PDIA4,PDLIM1,PDLIM5,PEA15,PELO,PFDN4,PFKM,PGLS,PKM,PLIN3,PPIF,PRRC1,PSMB4,PSMB7,PTPRF,RAC1,RARS,RBM39,RCN2,SAP18,SCARB2,SCFD1,SCP2,SDC1,SDC4,SDR16C5,SEC11C,SEC23B,SELENBP1,SF3B3,SFN,SLC25A1,SLC25A6,SLC27A2,SOD2,SRP54,STAU2,SUSD2,TARS,TGM2,TM7SF2,TMED9,TMSB10/TMSB4X,TRIM28,TRIO,TRIP13,TUBA1A,TUBB2A,TXN,USP14,USP9X,UTP18,VWA5A,XPO5,YBX3,YWHAB,YWHAE,ZW10	169
Cancer,Organismal Injury and Abnormalities	Malignant solid tumor	3.137868621	-1.926	AACS,ACADM,ACOX1,ACSM1,ACTN4,AIMP1,AKR1A1,AKR1C1/AKR1C2,ALCAM,ALDOA,ALOX15B,ANO1,ANXA4,APOD,APOL3,ARFRP1,ARHGAP1,ATIC,BCAM,BPNT1,BST1,BUB3,C1QBP,CALB2,CAPS,CEP152,CHI3L1,CLSTN2,CNDP2,COL2A1,COPS8,CORO1B,CRABP1,CROT,CTNNA1,CTSZ,CXCL17,CYP1B1,CYP4X1,DCLK1,EHHADH,EML2,ENO1,EPB41L1,EPRS,EPX,ERBB3,ERP29,ETF1,EXOSC6,FABP7,FKBP5,FN1,GALM,GAPDH,GART,GGT1,GOT2,GSTM1,GSTT2/GSTT2B,GUSB,HAAO,HACL1,HAGHL,HBD,HEXB,HLA-DRB1,HMGN5,HNRNPH1,HPGD,HSD17B4,HSPA4,HSPA9,HTT,IARS,IDH1,IGFBP5,IKBKB,IQGAP2,ISYNA1,ITGA3,KRT15,LANCL1,LARS,LBP,LGALS3BP,LOXL2,LRBA,LTF,LYPLA1,MAP1S,MBLAC2,MCM2,MCM6,MIA3,MTHFD1,MUCL1,MYH14,NAMPT,NANS,NDUFAF2,NFS1,NME1,NME1-NME2,OPTN,PA2G4,PAFAH1B1,PAFAH1B3,PAICS,PARP1,PCNA,PDIA4,PDLIM1,PDLIM5,PEA15,PELO,PFDN4,PFKM,PGLS,PKM,PLIN3,PPIF,PRRC1,PSMB4,PSMB7,PTPRF,RAC1,RARS,RBM39,RCN2,SAP18,SCARB2,SCFD1,SCP2,SDC1,SDC4,SDR16C5,SEC11C,SEC23B,SELENBP1,SF3B3,SFN,SLC25A1,SLC25A6,SLC27A2,SOD2,SRP54,STAU2,SUSD2,TARS,TGM2,TM7SF2,TMED9,TMSB10/TMSB4X,TRIM28,TRIO,TRIP13,TUBA1A,TUBB2A,TXN,USP14,USP9X,UTP18,VWA5A,XPO5,YBX3,YWHAB,YWHAE,ZW10	169
Cellular Development,Cellular Growth and Proliferation	Proliferation of pancreatic cancer cell lines	3.096910013	1.617	CNDP2,ERBB3,NAMPT,PARP1,SFN,SOD2,USP9X	7

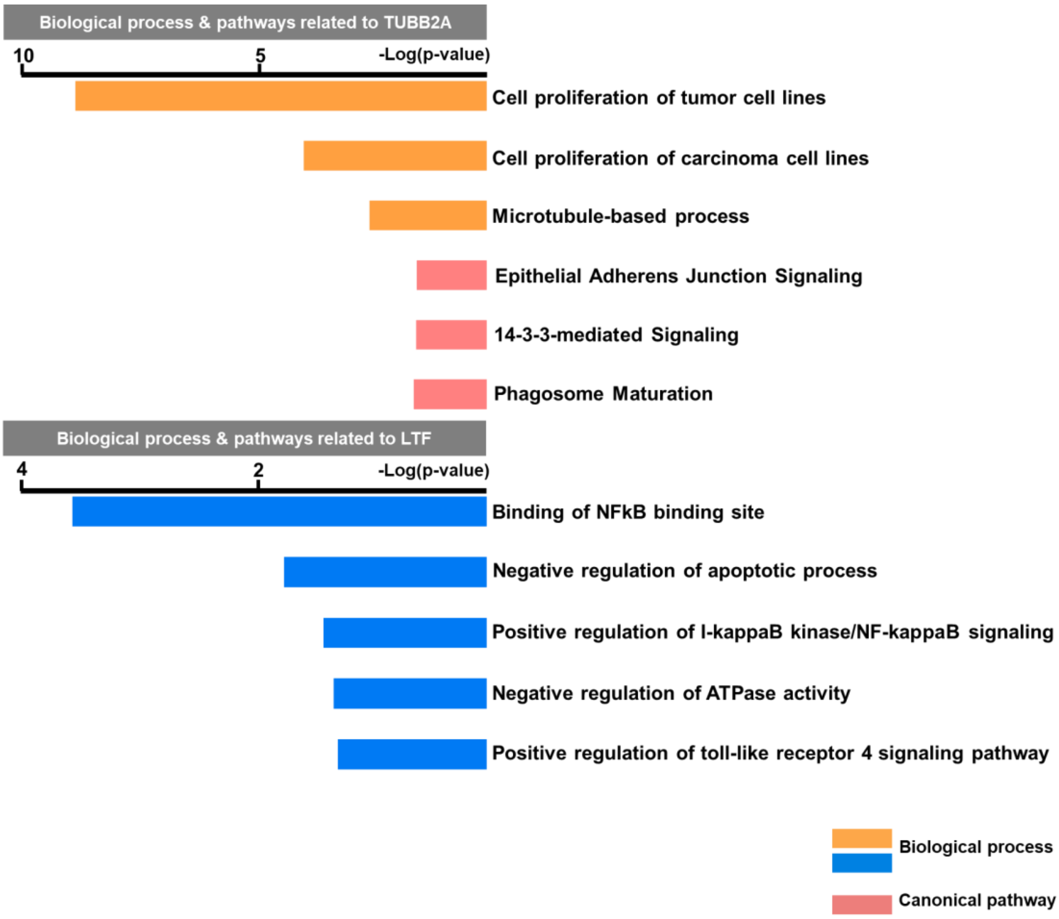


Figure 12. Biological functions and canonical pathways related to two biomarker candidates by IPA and DAVID analysis. Biological functions and pathways of TUBB2A (upper panel) and LTF (lower panel) (Fisher's exact test p-value < 0.05 for DAVID and IPA analysis).

Table 5. Biological functions of TUBB2A and LTF. Biological functions of TUBB2A and LTF were examined using the IPA and DAVID bioinformatics tools. Biological processes and canonical pathways of TUBB2A and LTF are listed. The p-value cutoff was set to < 0.05 for the IPA analysis. The p-value (modified Fisher exact p-value) cutoff for the GO annotation was set to < 0.05. Proteins in each biological function are listed.

TUBB2A		
Diseases or Functions Annotation(IPA)	**p-value**	**Molecules**
Cell proliferation of tumor cell lines	1.69E-08	ACTN4,ALOX15B,ANO1,BUB3,C1QBP,CNDP2,COPS8,CYP1B1,DCLK1,ERBB3,FKBP5,FN1,GAPDH,GGT1,GSTM1,HPGD,HTT,IGFBP5,IQGAP2,ITGA3,LGALS3BP,MCM2,MUCL1,MYH14,NAMPT,NDUFAF2,NFS1,NME1,PA2G4,PAFAH1B1,PARP1,PCNA,PKM,PPIF,PSMB7,PTPRF,RAC1,SDC1,SDC4,SFN,SLC25A6,SOD2,TGM2,TMSB10/TMSB4X,TRIM28,TRIO,TUBB2A,TXN,USP9X
Cell proliferation of carcinoma cell lines	0.00032	C1QBP,ERBB3,GAPDH,HPGD,LGALS3BP,MYH14,NAMPT,NDUFAF2,PKM,PTPRF,RAC1,SLC25A6,TGM2,TRIM28,TUBB2A
Canonical Pathways(IPA)	**p-value**	**Molecules**
Epithelial Adherens Junction Signaling	0.00200	ACTN4,CTNNA1,MYH14,RAC1,TUBA1A,TUBB2A
14-3-3-mediated Signaling	0.00501	SFN,TUBA1A,TUBB2A,YWHAB,YWHAE
Phagosome Maturation	0.00759	CTSZ,HLA-DRB1,HLA-DRB3,TUBA1A,TUBB2A
Biological Process(GO)	**p-value**	**Molecules**
microtubule cytoskeleton organization	0.03275	P04406, Q66K74, O75122, P43034

LTF		
Diseases or Functions Annotation(IPA)	**p-value**	**Molecules**
Binding of NFkB binding site	0.00022	FN1,LTF,RAC1,SOD2,TRIM28
Biological Process(GO)	**p-value**	**Molecules**
negative regulation of apoptotic process	0.01600	P02788, P21980, P14625, P09525, P04179, O14920, P30405
positive regulation of I-kappaB kinase/NF-kappaB signaling	0.03537	O95236, P02788, P21980, O14920
negative regulation of ATPase activity	0.04379	P02788, P30405
positive regulation of toll-like receptor 4 signaling pathway	0.04806	P02788, P18428

7. Proteomic alterations in distant metastatic breast cancer between molecular subtypes

According to the results of a previous study, pooling biological groups can reduce the variation that originates from the sample while retaining the defining features of the group itself [57]. I expected our pooled samples for each molecular subtype to reveal distinct information on the molecular characteristics between the HER2, TNBC, and luminal groups. For these reasons, a pooled sample set was used to identify the changes in proteins between distinct breast cancer molecular subtypes in the distant metastasis and nondistant metastasis groups.

By ANOVA, 1086 proteins were differentially expressed between breast cancer molecular subtypes (p-value < 0.05) (Figure 13a). These DEPs were then analyzed by hierarchical clustering to determine their expression patterns between breast cancer molecular subtypes, resulting in 6 groups: upregulated proteins in HER2-non-distant metastasis (cluster 1; 176 DEPs), upregulated proteins in HER2-distant metastasis (cluster 2; 124 DEPs), upregulated proteins in TNBC-non-distant metastasis (cluster 3; 193 DEPs), upregulated proteins in TNBC-distant metastasis (cluster 4; 342 DEPs), upregulated proteins in luminal-non-distant metastasis (cluster 5; 29 DEPs), and upregulated proteins in luminal-distant metastasis (cluster 6; 184 DEPs).

8. Biological functions of distant metastatic breast cancer between molecular subtypes

To gain greater insight into the molecular features of distant metastatic breast cancer between molecular subtypes, pathway enrichment analysis was conducted for clusters 2, 4, and 6, which comprised proteins that were upregulated in the distant metastasis group of each molecular subtype. By Ingenuity Pathway Analysis (IPA), 2 canonical pathways were derived for cluster 2, versus 14 for cluster 4 and 11 for cluster 6 (p-value < 0.05, Z-score > 1) (Figure 13b-d and Table 6). Specifically, in cluster 2, only PI3K/AKT signaling and BAG signaling were deduced and activated between three subtypes. PI3K/AKT signaling was the most highly activated pathway (Z-score = 2) in the HER2 type (Figure 13b and Table 6). In cluster 4, all 14 pathways were activated—glycolysis 1, gluconeogenesis 1, and tRNA charging were extensively activated in the TNBC types (Figure 13c and Table 6). tRNA charging was the most highly activated pathway (Z-score = 2.828), whereas EIF2 signaling was the least activated (Z-score = 0.333) in TNBC types (Figure 13c and Table 6). In cluster 6, most pathways were activated, such as actin cytoskeleton signaling, acute phase response signaling, intrinsic prothrombin activation, and GP6 signaling, in the luminal type. Among them, GP6 signaling was the most highly activated (Z-score = 3.464). However, LXR/RXR signaling was inhibited in the luminal type (Z-score = -0.707) (Figure 13d and Table 6). Based on our results, distinct activation states exist between the HER2, TNBC, and luminal types.

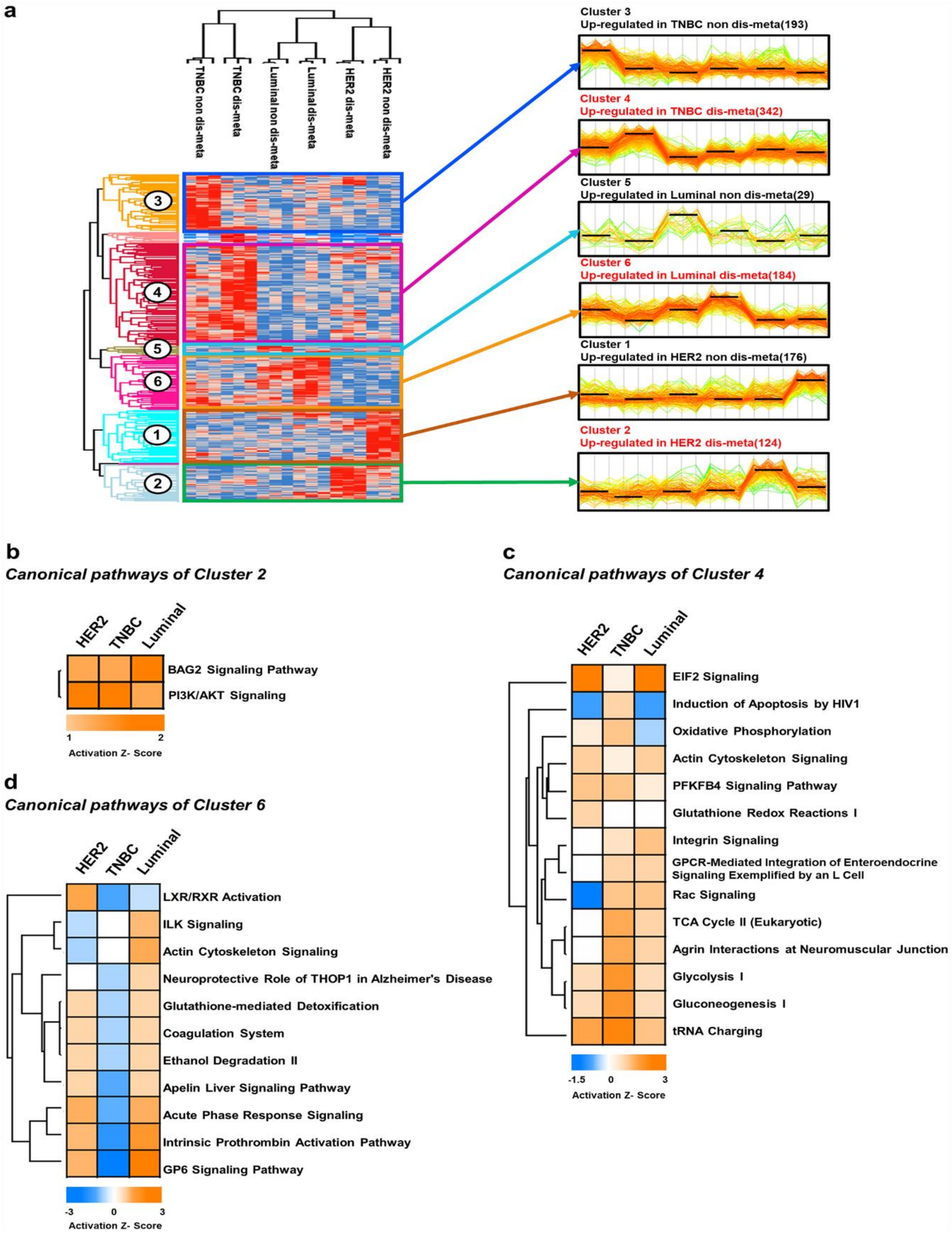


Figure 13. Proteomic alteration in distant metastatic breast cancer between molecular subtypes. (a) Hierarchical clustering of differentially expressed proteins (DEPs) between distant metastatic breast cancer molecular subtypes (ANOVA, p-value<0.05). The DEPs (1086) from the

pooled sample set were divided into 6 groups. Clusters of upregulated proteins are marked in red. (b)-(d) Canonical pathway enrichment of clusters 2, 4, and 6. The significant pathways (Fisher's exact test p-value <0.05) were deduced using Ingenuity Pathway Analysis (IPA), and their activation and inhibition states are expressed as Z-scores.

Table 6. Canonical pathways of clusters enriched by IPA analysis. Canonical pathways in clusters 2, 4, and 6 of the pooled sample set were investigated using the IPA informatics tool. Canonical pathways between molecular subtypes are listed. The p-value cutoff was set to < 0.05, and the activation Z-score was set to > 1. P-values and Z-scores of the canonical pathways are listed.

Cluster 2	Canonical Pathway	HER2 (Dis-meta/Non dis-meta)_Activation Z-score	TNBC (Dis-meta/Non dis-meta)_Activation Z-score	Luminal (Dis-meta/Non dis-meta)_Activation Z-score	-log(P value)
	PI3K/AKT Signaling	2	2	1	2.093
	BAG2 Signaling Pathway	1	1	2	3.923

Cluster 4	Canonical Pathway	HER2 (Dis-meta/Non dis-meta)_Activation Z-score	TNBC (Dis-meta/Non dis-meta)_Activation Z-score	Luminal (Dis-meta/Non dis-meta)_Activation Z-score	-log(P value)
	tRNA Charging	2.121	2.828	1.414	6.733
	EIF2 Signaling	3	0.333	3	9.190
	Gluconeogenesis I	0.816	2.449	0.816	5.535
	Glycolysis I	0.816	2.449	0.816	5.648
	Rac Signaling	-1.342	1.342	1.342	1.403
	PFKFB4 Signaling Pathway	1.342	1.342	0.447	2.993
	Agrin Interactions at Neuromuscular Junction	0	2	1	1.407
	TCA Cycle II (Eukaryotic)	0	2	1	3.195
	Induction of Apoptosis by HIV1	-1	1	-1	1.738
	Actin Cytoskeleton Signaling	1.134	0.378	1.134	1.542
	Oxidative Phosphorylation	0.447	1.342	-0.447	1.507
	Integrin Signaling	0	0.707	1.414	1.627
	GPCR-Mediated Integration of Enteroendocrine Signaling Exemplified by an L Cell	0	1	1	1.481
	Glutathione Redox Reactions I	1	0	0	3.345

Cluster 6	Canonical Pathway	HER2 (Dis-meta/Non dis-meta)_Activation Z-score	TNBC (Dis-meta/Non dis-meta)_Activation Z-score	Luminal (Dis-meta/Non dis-meta)_Activation Z-score	-log(P value)
	GP6 Signaling Pathway	1.732	-3.464	3.464	9.305
	Intrinsic Prothrombin Activation Pathway	1.633	-2.449	2.449	5.835
	Acute Phase Response Signaling	1.897	-1.897	1.897	10.272
	LXR/RXR Activation	2.121	-2.121	-0.707	4.934
	Apelin Liver Signaling Pathway	1	-2	1	4.132
	Ethanol Degradation II	1	-1	1	4.066
	Coagulation System	1	-1	1	3.615
	Actin Cytoskeleton Signaling	-1	0	2	1.390
	Glutathione-mediated Detoxification	1	-1	1	4.202
	ILK Signaling	-0.816	0	1.633	3.613
	Neuroprotective Role of THOP1 in Alzheimer's Disease	0	-1	1	3.395

DISCUSSION

One of the goals of our study was to discover novel protein biomarker candidates of distant metastatic breast cancer. Initially, I considered the potential problem with multiple comparisons, which can generate false positives if unaddressed, in selecting the protein targets. Therefore, I applied a multiple testing correction to our datasets. However, none of proteins was able to pass the BH FDR cutoff. Thus, we proposed alternative criteria to compensate for the statistically insufficient significance of proteins in determining the protein targets.

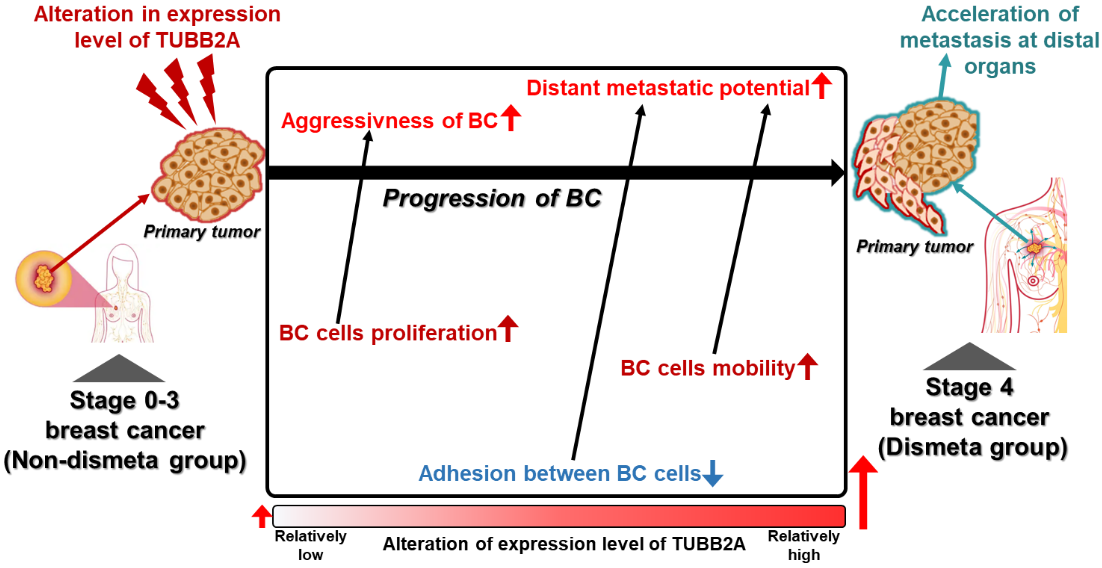
When the criteria were applied to our in-depth proteome data, LTF (p-value < 0.001) and TUBB2A (p-value < 0.05) appeared as important protein targets for validation of distant metastatic potential. TUBB2A was upregulated and LTF was downregulated in the distant metastasis group. TUBB2A was upregulated in more invasive breast cancer cell lines (i.e., BC cell lines in the higher invasive group), whereas the expression patterns of LTF were perturbed across breast cancer cell lines by RT-PCR. Considering the expression level of TUBB2A in the higher-invasiveness group and the high malignancy of distant metastatic breast cancer [4, 58, 59], the upregulation of TUBB2A might promote the invasion of breast cancer cells, inducing the potential of distant metastatic breast cancer. In addition, based on the results of the invasion and migration assay, we verified that the high expression of TUBB2A increases the mobility of breast cancer cells, providing further support for TUBB2A as a novel biomarker candidate of distant metastatic breast cancer.

Regarding performance of TUBB2A, TUBB2A could distinguish between distant metastasis and nondistant metastasis (i.e., 78% sensitivity, 100% specificity, and an AUC value of 0.852) and might predict distant metastasis (i.e., 88% PPV) in the individual sample set. However, because our TMT-based data were obtained from a small cohort (n=36), future studies should evaluate the performance of TUBB2A by absolute quantitation in a large cohort to assess its clinical applicability, which lies beyond the scope of our current study. One possible design would be to quantify TUBB2A using targeted proteomic techniques, such as multiple reaction monitoring (MRM) and parallel reaction monitoring (PRM).

Another goal was to determine the overall biological functions that exist in distant metastatic breast cancer. Biological functions that are related to proliferation and movement of cancer cells were activated. Specifically, cell polarization/orientation was related to cell adhesion, and actin-based signaling was associated with migration [60-62]. NF-kappa B modulates the immune response, but its inhibition and dysregulation are linked to improper immune development [63, 64]. Thus, the inhibition of polarization of tumor cell lines and adhesion of BC cell lines might weaken the adhesion between cells in primary breast tumors, and the activation of actin cytoskeletal signaling and proliferation of tumor cell lines might enhance the movement of breast cancer cells. In addition, blocking NF kappa B binding sites might allow breast cancer cells to migrate to other distal sites without activating the immune system.

I noted proteins that were associated with distant metastatic breast cancer, based on our bioinformatics analysis. By GO analysis, ‘cell-cell adhesion’ terms were observed in upregulated and downregulated DEPs. However, each term consisted of different proteins. Furthermore, proteins in ‘adhesion of BC cell lines’ term did not overlap with those in the ‘cell-cell adhesion’ term. Thus, adhesion between breast cancer cells in primary tumors might be weakened, but that between breast cancer cells and cells in other organs could be strengthened, due to various proteins with potentially distinct functions in cell adhesion. In our pathway enrichment analysis, FN1 overlapped between activated leukocyte extravasation signaling and inhibited acute phase response signaling. Considering the opposing states of these pathways, the former might enhance the mobility of breast cancer cells to other organs, shuttling leukocytes out of the circulatory system. In parallel, inhibition of acute phase response signaling might suppress the immune response. Thus, FN1 might create a suitable microenvironment that is conducive to distant metastasis of breast cancer.

With regard to our protein targets, TUBB2A was associated with cellular proliferation, movement, and adhesion, and LTF was involved in cell death, the immune response, and metabolism. The exact role TUBB2A plays in distant metastasis of breast cancer remain unclear to our knowledge. However, based on these functions, TUBB2A might control the mobility of distant metastatic breast cancer by regulating the adhesion and proliferation of breast cancer cells, and LTF might govern the death of breast cancer cells and the immune system during distant metastasis. Thus, TUBB2A might be a key protein that controls the migration of breast cancer cells from a primary tumor. LTF might be an auxiliary protein that helps breast cancer cells survive during movement toward distal sites by disrupting the immune system. A schematic model of biological regulation of distant metastasis of breast cancer according to alterations of the expression level of TUBB2A (a novel biomarker candidate of this study) was presented in Figure 14.



Alteration in expression level of TUBB2A
Primary tumor
Stage 0-3 breast cancer (Non-dismeta group)
Aggressivness of BC↑
Distant metastatic potential↑
Progression of BC
BC cells proliferation↑
BC cells mobility↑
Adhesion between BC cells↓
Acceleration of metastasis at distal organs
Primary tumor
Stage 4 breast cancer (Dismeta group)
Relatively low
Alteration of expression level of TUBB2A
Relatively high

Figure 14. A schematic model of regulation of distant metastasis of breast cancer.

Another goal was to determine the characteristics of distant metastatic breast cancer between molecular subtypes. In cluster 2, the most highly activated pathway was PI3K/AKT signaling in the HER2 type. A previous study that used transcriptome data revealed that PI3K/AKT kinases are expressed in circulating breast tumor cells and that the activation of this signal regulates their metastatic and malignant state [68]. Compared with our proteomic results, the activation states of PI3K/AKT signaling were consistent. Thus, our PI3K/AKT signaling proteins might be associated with the regulation of distant metastatic potential and function as targets for the eradication of HER2-type distant metastatic breast cancer.

In cluster 4, the most highly activated pathway was tRNA charging signaling in the TNBC type. The exact functions of this pathway in distant metastatic breast cancer have not been determined. However, based on a previous study, tRNA overexpression in breast tumor cells might increase the translational efficiency of genes that are related to the progression and development of breast cancer [67]. The tRNA charging-related proteins that I recorded might be upregulated and translationally modified products of such genes, influencing the distant metastatic potential and progression of breast cancer. Thus, these proteins might be targets for removal or suppression in slowing the malignancy of TNBC-type distant metastatic breast cancer.

In cluster 6, the most highly activated pathway was glycoprotein 6 (GP6) signaling in the luminal type. GP6 is a platelet membrane glycoprotein that functions as a receptor for collagen and regulates the collagen-induced activation and aggregation of platelets [65, 66]. The detailed functions of this pathway in distant metastatic breast cancer have not been described. However, based on its functions, breast cancer cells could migrate easily to distal sites, masking their aggregate forms with platelet-combined forms. Furthermore, breast cancer cell complexes might adhere to collagen and subsequently to platelets, leading to additional platelet aggregation. Thus, GP6 signaling and its factors might facilitate the circulation of breast cancer cells with little activation of the immune systems due to their disguised forms, allowing them to settle at distal sites. Furthermore, the expression level of these proteins could be used to monitor the progression of luminal-type distant metastatic breast cancer.

Although I performed pathway enrichment analysis using the upregulated DEPs in the 3 clusters, one of the benefits of our study was that it could have considered the downregulated DEPs in the remaining 3 clusters (clusters 1, 3, and 5) in the analysis. These proteins might be related to distinct biological activities that suppress the activation of distant metastatic breast cancer between subtypes. Consequently, our proteomic clusters might expand our understanding of the effects of molecular subtype on distant metastatic breast cancer.

Without our in-depth proteomic data, most of our DEPs might be unable to be identified or detected in other studies, because I are the first to collect proteomic data in distant metastatic breast cancer, analyzing clinical FFPE tissues from primary breast tumors. Our results indicate that the pathological relevance of our FFPE tissues in BC research is valid at the proteomic level and in severe breast cancer pathologies. Through our latent data, I discovered a novel protein biomarker candidate that has the potential to distinguish distant metastatic breast cancer and demonstrated distinct molecular features between BC subtypes. I expect that the discoverd biomarker candidate can be used to diagnose and predict distant metastatic breast cancer. Furthermore, our molecular pathways should provide insights into the relationship between molecular subtypes and distant metastatic breast cancer.

In conclusion, I have constructed a comprehensive proteome of distant metastatic breast cancer by analyzing FFPE tissue slides using TMT-based mass spectrometric techniques. Our study demonstrates that the TMT-based approach is beneficial, because its greater quantitative ability generates a larger selection of proteins from which to choose novel biomarker candidates. This finding was verified by our proteomic dataset, which comprised the largest number of proteins in distant metastatic breast cancer. Through our criteria, I selected 2 important protein targets for distant metastatic breast cancer and performed functional studies to validate them. Finally, I was able to determine a novel protein biomarker candidate. Furthermore, our bioinformatics analysis revealed specific molecular characteristics between molecular subtypes. Thus, our in-depth proteomic data and analyses can be an important resource for distant metastatic breast cancer research. In future studies, I hope to assemble a larger cohort of breast cancer FFPE samples to test the performance of our novel biomarker candidate using targeted proteomics techniques, such as parallel reaction monitoring (PRM) and multiple reaction monitoring (MRM).

CHAPTER II

Quantitative Proteomic Approach for Discriminating Major Depressive Disorder and Bipolar Disorder by Multiple Reaction Monitoring-Mass Spectrometry

INTRODUCTION

Major depressive disorder (MDD) and bipolar disorder (BD) are common psychiatric disorders, with overall prevalence rates of 3% to 10% and 2% to 4%, respectively [82, 83]. Both disorders are debilitating, with the highest and fourth-highest Disability-Adjusted Life Year values in the Korean population for MDD and BD among all mental and substance use disorders [84].

Distinguishing MDD and BD has been challenging, because their diagnosis relies on behavioral observations and subjective symptoms. The complexity, heterogeneity, and commonality between these disorders further complicate the diagnosis. Nearly 40% of BD patients are initially misdiagnosed with MDD [85-87]. Thus, their misdiagnosis represents as serious problem, because it leads to the erroneous prescription of antidepressant monotherapy, which can worsen outcomes by inducing hypomanic/manic states and rapid cycling during the course of the disorder [88, 89]. For these reasons, the discovery of biomarkers that differentiate MDD and BD has garnered significant interest from clinicians and researchers alike.

In the research of mood disorders, proteomics, which studies proteins that reflect functions and phenotypes, has focused on discovering candidate biomarkers, given the limitations of genomic studies in this endeavor [90]. Although traditional proteomic studies have centered on proteomic alterations in the central nervous system, several limitations, such as invasiveness and accessibility issues, have impeded the collection of brain and cerebrospinal fluid (CSF) samples [90, 91]. Thus, the application of quantitative proteomic studies to analyze peripheral blood in mood disorders has increased. Among such studies, most have focused on differentiating MDD from healthy controls (HCs) or BD from HCs [90, 92, 93].

Recently, studies have attempted to differentiate MDD from BD using various proteomic technologies [94-97]. An immunoassay-based proteomic approach has proposed a predictive model to distinguish MDD from BD by quantifying multiple proteins [97]. Furthermore, based on recent advances in mass spectrometry (MS)-based proteomics in improving high-throughput techniques for proteomic quantitation [98, 99], such techniques (eg, MALDI-TOF/TOF MS and LC-MS/MS) have also contributed to discriminating between MDD and BD by examining proteomic profiles or identifying biomarker candidates [94-96]. Although these studies proposed many candidates, such as C3, C4BPA, CFI, B2RAN2, ENG, RAB7A, ROCK2, XPO7, PDGF-BB, and TSP-1, their clinical relevance and significance remain unknown, necessitating further validation of these biomarkers [94-97]. Moreover, although MS-based targeted proteomic techniques have been used in studies to differentiate mood disorders from HCs or between HCs and severity of a specific mood disorder [100, 101], no study has attempted to discriminate MDD from BD using MS-based quantitative targeted proteomic methods, such as LC-MRM-MS, at least to our knowledge.

LC-MRM-MS is a highly selective and sensitive targeted proteomic technology for quantifying targeted proteins or peptides in biological samples [102, 103]. In contrast to conventional technologies, such as immunoassays, this technology can measure at least 300 protein targets per sample simultaneously, with precision [104]. In addition, LC-MRM-MS generates consistent and reproducible data from highly complex samples between laboratories [105]. Furthermore, high-throughput LC-MRM-MS technology has been used to quantify potential protein or peptide targets that are associated with diseases or disorders [100, 101, 105-112]. Recently, LC-MRM-MS has been applied in research on psychiatric disorders [100, 105, 112]. Consequently, the LC-MRM-MS-based proteomic approach has also necessitated in the efforts of discriminating between MDD and BD.

In this study, using LC-MRM-MS technology, I developed a model for distinguishing MDD from BD patients, based on proteomic data on their blood specimens. I performed several methods to reduce model overfitting and considered the generalizability of the model, with regard to feature extraction and model averaging. In addition, I determined the performance of the model for patients without current hypomanic/manic/mixed symptoms and those who were drug-free and compared its performance for MDD or BD patients against HCs. Furthermore, I studied the biological interactions across the proteins that constituted the model with MDD and BD patients. As a result, I propose the application of quantitative targeted proteomics to blood samples in mood disorders, supporting the applicability of MS-based proteomic approaches in the diagnosis of MDD and BD.

MATERIALS AND METHODS

1. Study design

Overall, this study comprised five steps—sample collection from matched groups, MRM-MS analysis, model development in the training set, model evaluation in the test set and combined set, and model application to different subgroups of patients and HC group. First, plasma samples of 270 individuals (90 MDD, 90 BD, and 90 HCs) whose gender, age, and BMI were matched were collected. Second, protein targets for MDD and BD were quantified by MRM-MS in plasma samples of 270 individuals. Third, a model for discriminating MDD and BD was developed in the training set. Fourth, the model was evaluated in the test and combined sets. Lastly, the developed model was applied to MRM-MS data of patients' subgroups and that of HCs. The overall scheme of this study was presented in Figure 1.

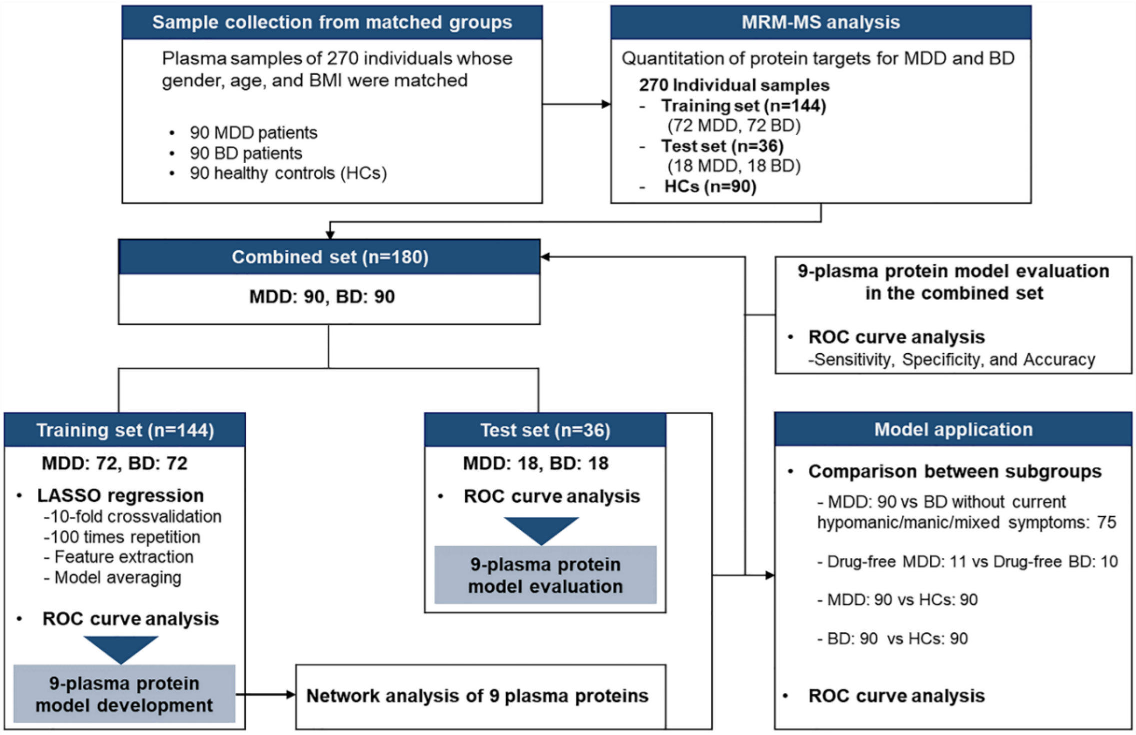

Sample collection from matched groups
Plasma samples of 270 individuals whose gender, age, and BMI were matched
• 90 MDD patients
• 90 BD patients
• 90 healthy controls (HCs)
MRM-MS analysis
Quantitation of protein targets for MDD and BD
270 Individual samples
- Training set (n=144)
(72 MDD, 72 BD)
- Test set (n=36)
(18 MDD, 18 BD)
- HCs (n=90)
Combined set (n=180)
MDD: 90, BD: 90
9-plasma protein model evaluation in the combined set
• ROC curve analysis
-Sensitivity, Specificity, and Accuracy
Training set (n=144)
MDD: 72, BD: 72
• LASSO regression
-10-fold crossvalidation
-100 times repetition
- Feature extraction
- Model averaging
• ROC curve analysis
9-plasma protein model development
Test set (n=36)
MDD: 18, BD: 18
• ROC curve analysis
9-plasma protein model evaluation
Network analysis of 9 plasma proteins
Model application
• Comparison between subgroups
- MDD: 90 vs BD without current hypomanic/manic/mixed symptoms: 75
- Drug-free MDD: 11 vs Drug-free BD: 10
- MDD: 90 vs HCs: 90
- BD: 90 vs HCs: 90
• ROC curve analysis

Figure 1. Overall scheme of this study. This study consists of "sample collection", "MRM-MS analysis", "model development", "model evaluation" and "model application". The number of

subjects of this study and methods corresponding to each step are illustrated. MDD, major depressive disorder; BD, bipolar disorder; LASSO, least absolute shrinkage and selection operator; AUROC, area under receiver operating characteristics.

2. Clinical samples

The cohort comprised 180 patients and 90 HCs, matched for age and sex. The patients and HCs were enrolled from August 2018 to November 2019, aged 19 to 65 years. Within the cohort, 90 patients had BD [43 BD-I, 43 BD-II, and 4 BD-not otherwise specified (NOS)], and 90 patients had MDD. BD-I, formerly known as manic-depressive disorder, is characterized by episodes of mania and depression. Although BD-II is similar to BD-I, it is hallmarked by episodes of depression and hypomania that never reach the severity of mania. BD-NOS is known as subthreshold BD or a type that does not fulfill the criteria of BD-I or BD-II. MDD is characterized by at least 1 episode of depression, with no mania/hypomania episodes to rule out BD-I and BD-II and no subthreshold hypomanic episodes to rule out BD-NOS.

Patients were enrolled from 1) Seoul National University Hospital, 2) Seoul Metropolitan Government Seoul National University Boramae Medical Center, 3) Nowon Eulji Medical Center, Eulji University, 4) Hanyang University Seoul Hospital, 5) Cha University Bundang Medical Center, and 6) Inha University Hospital. The diagnoses of the patients were made per the Diagnosis and Statistical Manual of Mental Disorders 5th version (DSM-5) and confirmed using the Mini-International Neuropsychiatric Interview (MINI). HCs were recruited from Seoul National University Hospital by advertisement. HCs had to have no psychiatric diagnosis according to the MINI and no psychiatric history in second-degree relatives.

Patients and HCs were excluded per the following criteria: those who took anti-inflammatory analgesics, including non-steroidal anti-inflammatory drugs (NSAIDs) and steroids (acetaminophen was allowed) for the past 2 weeks; those who had received intensive psychotherapy for the past 2 months; history of neuromodulation [electroconvulsive therapy

(ECT), transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS), or deep brain stimulation (DBS)]; neurosurgery; central nervous system disease, including epilepsy, stroke, parkinsonism, and meningitis; cancer; tuberculosis; history of substance abuse (exception for nicotine, caffeine, and alcohol); pregnancy/lactation; and those who were predicted to have an intellectual disability or had difficulty interpreting the Korean language. Patients who were on anti-inflammatory analgesics and steroids were excluded, because immune and inflammatory pathways are linked to mood disorders [113, 114], and those who were undergoing intensive psychotherapy were excluded to confine treatment influences to psychotropic medications. Most of the other exclusion criteria were based on previous studies of certain diseases and conditions and their association with altered protein expression [115-121].

The study was carried out in accordance with the Declaration of Helsinki, and the study design was reviewed by the institutional review boards of Seoul National University Hospital (IRB No. 1806-106-951) and all other participating hospitals. Informed consent was obtained from each participant.

Plasma samples were collected from each subject in a 6-ml EDTA tube (Ref. 367863, Becton Dickinson and Company, Trenton, NJ, USA) and centrifuged at 1100-1300 *g* for 10-15 minutes at room temperature or 4°C; the supernatant was collected and stored in Eppendorf tubes at < -70°C until use.

3. Demographics and clinical features

Demographics, such as gender, age, body mass index (BMI), current smoking status, current exercise status, current alcohol use, blood collection time, fasting time, duration from first onset (years), and duration from first medication (years), were considered. Age, BMI, duration from first onset (years), and duration from first medication (years) were analyzed as continuous variables. Gender (male/female), current smoking status (yes/no), current exercise status (yes/no), current alcohol use (yes/no), blood collection time (AM, PM), and fasting time (< 8 hours, ≥ 8 hours) were analyzed as dichotomous variables. Current exercise status (yes/no) was based on the

World Health Organization (WHO) recommendation for moderate-intensity physical activity at least once per week, for 30 minutes [122]. Current alcohol use (yes/no) was defined as at least one drink, once per week.

Symptom severity was assessed using the Brief Psychiatric Rating Scale (BPRS) [123], Young Mania Rating Scale (YMRS) [124], Montgomery-Asberg Depression Rating Scale (MADRS) [125], and Hamilton Anxiety Scale (HAM-A) [126]. Because bipolar disorder has several mood states, we classified those whose YMRS scores were over 12 points as having current hypomanic/manic/mixed symptoms [127] and excluded them from the secondary analysis. Medication use was analyzed as a dichotomous variable with the following classifications: antipsychotics (AP), mood stabilizer (MS), antidepressants (AD), and benzodiazepines/hypnotics (BZD/HNT). Patients who had been drug-free for at least 1 month (11 MDD and 10 BD) were included in the secondary analysis.

4. Protein quantification by quantitative targeted proteomic approach (MRM-MS)

4.1. Determination of quantifiable targets for MRM-MS analysis

In this study, 210 proteins, corresponding to 671 peptides, were analyzed using targeted MRM-MS. These proteins and corresponding peptides were derived from various sources [92, 94-97, 128-137] and determined per the criteria in Figure 2. Specifically, three types of sources were compiled to generate the list of initial protein targets: 1) proteins that were obtained from our previous study that profiled the proteome of MDD and BD, 2) proteins that originated from previous proteomic studies on mood disorders, including significantly differentially expressed proteins between MDD and BD, between MDD and HC, and between BD and HC, and 3) proteins that have been approved by the US Food and Drug Administration (FDA) and designated as laboratory developed test (LDT). In total, 686 proteins were selected as initial targets after redundant entries were removed.

To investigate targets that had matching MS/MS spectra and unique peptides, 3 MS/MS spectral libraries—the Institute for Systems Biology (https://www.systemsbiology.org), National Institute of Standards and Technology (https://www.nist.gov), and the SWATHAtlas database (www. SWATHAtlas.org)—were used. In total, 648 proteins, corresponding to 7369 unique peptides, were selected.

To examine targets that were detected in blood samples reproducibly, MRM-MS analysis was performed on a pooled plasma sample that consisted of 51 healthy controls (HCs), 40 MDD, and 50 BD samples, which did not overlap with individual samples in our current study (90 MDD, 90 BD, and 90 HCs). Specifically, plasma samples of 3 HCs were collected twice at various time points—once each for the analysis of the pooled sample and the individual samples. Targets were considered detectable if: 1) at least 5 transitions for MRM-MS were observed; 2) they had the same elution patterns within the predictive retention time; and 3) the ratio of transition peaks was obtained as in the spectral library (dot product > 0.8). The dot product score represents the correlation between the peak intensities of transitions that originated from the endogenous peptides of interest and the corresponding entries in the spectral library. Regarding the 648 proteins that corresponded to 7369 unique peptides, 240 MRM-MS methods were generated, and 240 MS runs were performed in total. Reproducibility between MS runs was examined using indexed Retention Time (iRT) standard peptides. The technical reproducibility was evaluated, based on coefficient of variation (CV) values of the intensities and retention time (RT) of iRT peptides—CV value < 15% for intensity and < 1% for RT, respectively. Consequently, 412 proteins, corresponding to 1052 unique peptides, were selected as detectable targets in blood.

A total of 1052 SIS peptides, representing 1052 detectable endogenous peptides (412 proteins), were used to examine quantifiable targets. The targets were considered to be quantifiable, based on the following criteria: 1) co-elution of endogenous peptides of interest and corresponding SIS peptides; 2) top 5 peptides per protein, based on rank of peptide intensity; and 3) 1 representative transition per peptide, based on rank of intensity and Automated Detection of

Inaccurate and Imprecise Transitions (AuDIT) [138]. Consequently, 210 proteins, corresponding to 671 unique peptides and 671 transitions, were determined to be quantifiable targets (Table 1). These targets were applied to MRM-MS analysis of individual plasma samples of major depressive disorder (MDD) and bipolar disorder (BD) patients and healthy controls (HCs).

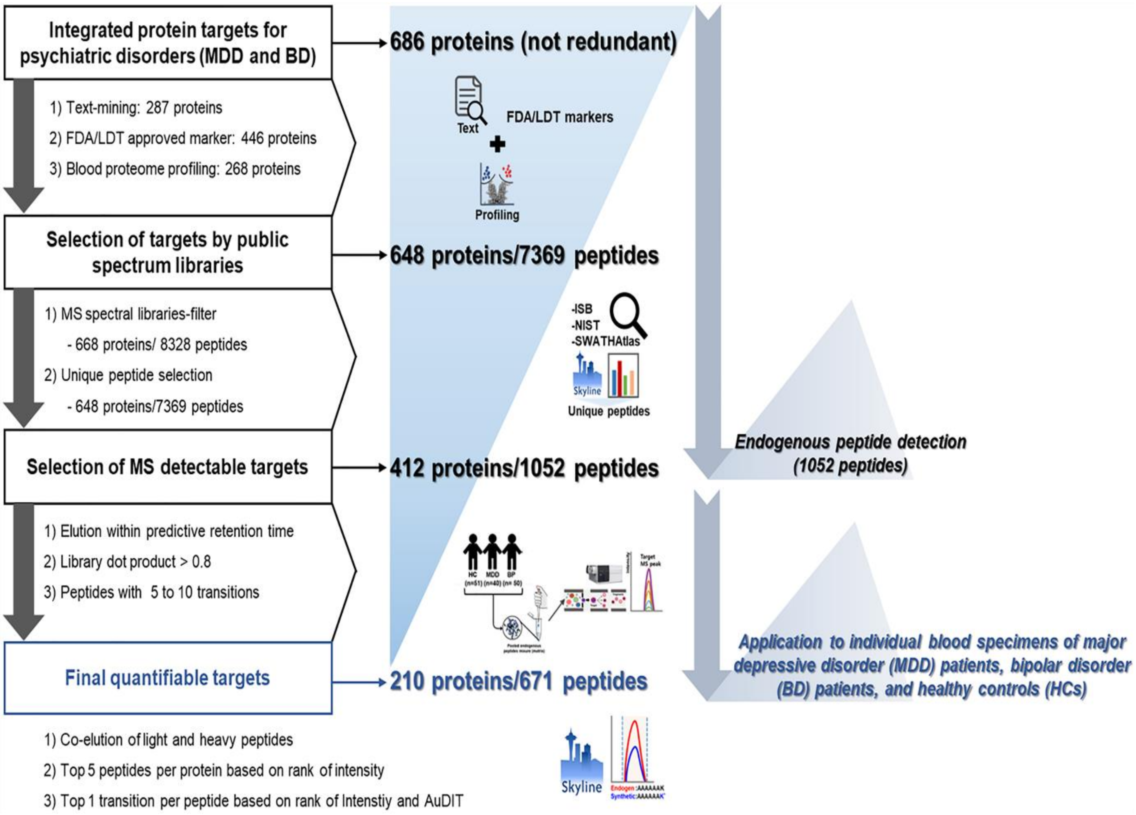

Integrated protein targets for psychiatric disorders (MDD and BD)
686 proteins (not redundant)
1) Text-mining: 287 proteins
2) FDA/LDT approved marker: 446 proteins
3) Blood proteome profiling: 268 proteins
Text
FDA/LDT markers
Profiling
Selection of targets by public spectrum libraries
648 proteins/7369 peptides
1) MS spectral libraries-filter
- 668 proteins/ 8328 peptides
2) Unique peptide selection
- 648 proteins/7369 peptides
-ISB
-NIST
-SWATHAtlas
Skyline
Unique peptides
Selection of MS detectable targets
412 proteins/1052 peptides
Endogenous peptide detection (1052 peptides)
1) Elution within predictive retention time
2) Library dot product > 0.8
3) Peptides with 5 to 10 transitions
HC MDD BP
(n=51) (n=40) (n= 50)
Pooled endogenous peptides mixture (matrix)
Target MS peak
Final quantifiable targets
210 proteins/671 peptides
Application to individual blood specimens of major depressive disorder (MDD) patients, bipolar disorder (BD) patients, and healthy controls (HCs)
1) Co-elution of light and heavy peptides
2) Top 5 peptides per protein based on rank of intensity
3) Top 1 transition per peptide based on rank of Intenstiy and AuDIT
Skyline
Endogen :AAAAAK
Synthetic:AAAAAK*

Figure 2. Determination of quantifiable targets for MDD and BD. Overall process, including integration of initial protein targets, selection of targets by public spectrum libraries, selection of MS detectable targets, and determination of final quantifiable targets. The number of selected targets for each step is represented. The final quantifiable targets were applied to MRM-MS analysis of individual blood samples of MDD patients, BD patients, and healthy controls. MS, mass spectrometry; MRM, multiple reaction monitoring; MDD, major depressive disorder; BD, bipolar disorder; HC, healthy control; FDA, US Food and Drug Administration; LDT, laboratory developed test; AuDIT, automated detection of inaccurate and imprecise transitions.

Table 1. The 671 peptides (210 proteins) examined by LC-MRM-MS[a]

The final quantifiable targets information			Mass Information											
Uniprot accession number	Protein	Peptide	Precursor ion. Light (m/z)	Precursor ion. Heavy (m/z)	Precursor ion charge	Product ion. Light (m/z)	Additional product ion.Light (m/z)[b]	Product ion. Heavy (m/z)	Product ion charge	Additional product ion charge[b]	Product ion type	Additional product ion type[b]	Product ion Quantifier[c]	Collision energy (volt)
P02763	A1AG1	EQLGEFYEALDCLR	871.9	876.9	2	258.1	876.4	258.1	1	1	b2	y7	Y	28
P02763	A1AG1	SDVVYTDWK	556.8	560.8	2	712.3		720.3	1		y5		Y	18.3
P19652	A1AG2	EQLGEFYEALDCLCIPR	1057	1062	2	272.2	933.4	282.2	1	1	y2	y7	Y	33.8
P04217	A1BG	ATWSGAVLAGR	544.8	549.8	2	730.4		740.4	1		y8		Y	17.9
P04217	A1BG	CEGPIPDVTFELLR	823.4	828.4	2	1089.6		1099.6	1		y9		Y	26.5
P04217	A1BG	ELLVPR	363.7	368.7	2	272.2	484.3	282.2	1	1	y2	y4	Y	12.3
P04217	A1BG	LLELTGPK	435.8	439.8	2	644.4		652.4	1		y6		Y	14.5
P04217	A1BG	VTLTCVAPLSGVDFQLR	938.5	943.5	2	1131.6		1141.6	1		y10		Y	30.1
P08697	A2AP	DFLQSLK	425.7	429.7	2	588.4		596.4	1		y5		Y	14.2
P08697	A2AP	LCQDLGPGAFR	617.3	622.3	2	547.3		557.3	1		y5		Y	20.1
P08697	A2AP	LFGPDLK	395.2	399.2	2	529.3		537.3	1		y5		Y	13.3
P08697	A2AP	QEDDLANINQWVK	786.9	790.9	2	674.4		682.4	1		y5		Y	25.4
P01023	A2MG	ALLAYAFALAGNQDK	783.4	787.4	2	561.3		569.3	1		y5		Y	25.3
P01023	A2MG	EQAPHCICANGR	471.5	474.9	3	494.2		499.2	2		y8		Y	12.2
P01023	A2MG	HNVYINGITYTPVSSTNEK	1069	1073	2	861.4		869.4	1		y8		Y	34.1
P01023	A2MG	NEDSLVFVQTDK	697.8	701.9	2	737.4		745.4	1		y6		Y	22.6
P01023	A2MG	QGIPFFGQVR	574.8	579.8	2	850.5		860.5	1		y7		Y	18.8
P01023	A2MG	TEVSSNHVLIYLDK	809.4	813.4	2	231.1	538.3	231.1	1	1	b2	y4	Y	26.1
P01023	A2MG	VTAAPQSVCALR	424.9	428.2	3	519.3		529.3	1		y4		Y	10.5
P01023	A2MG	VYDYYETDEFAIAEYNAPCSK	1274.5	1278.6	2	491.2		499.2	1		y4		Y	40.5
Q15848	ADIPO	GDIGETGVPGAEGPR	706.3	711.3	2	839.4		849.4	1		y9		Y	22.9
P43652	AFAM	FTFEYSR	475.2	480.2	2	701.3		711.3	1		y5		Y	15.7
P43652	AFAM	HFQNLGK	422.2	426.2	2	204.1	527.2	212.1	1	1	y2	b4	Y	14.1
P43652	AFAM	IAPQLSTEELVSLGEK	572	574.7	3	632.4		640.4	1		y6		Y	15.8
P43652	AFAM	TNFAFR	378.2	383.2	2	540.3		550.3	1		y4		Y	12.7
P43652	AFAM	YHYLIR	432.7	437.7	2	351.2		356.2	2		y5		Y	14.4
P02768	ALBU	LVNEVTEFAK	575.3	579.3	2	937.5		945.5	1		y8		Y	18.8
P04075	ALDOA	ALQASALK	401.2	405.3	2	157.1		157.1	2		b3		Y	13.4
P35858	ALS	ANVFVQLPR	522.3	527.3	2	759.5		769.5	1		y6		Y	17.2
P35858	ALS	DFALQNPSAVPR	657.8	662.8	2	626.4		636.4	1		y6		Y	21.4
P35858	ALS	LAELPADALGPLQR	732.4	737.4	2	314.2		314.2	1		b3		Y	23.7

P35858	ALS	LEALPNSLLAPLGR	732.4	737.4	2	1037.6		1047.6	1		y10		Y	23.7
P35858	ALS	LHSLHLEGSCLGR	493.6	496.9	3	614.8		619.8	2		y11		Y	13
P02760	AMBP	CVLFPYGGCQGNGNK	835.9	839.9	2	260.1	373.2	260.1	1	1	b2	b3	Y	26.9
P02760	AMBP	ECLQTCR	483.7	488.7	2	436.2		446.2	1		y3		Y	16
P02760	AMBP	ETLLQDFR	511.3	516.3	2	565.3		575.3	1		y4		Y	16.8
P02760	AMBP	GVCEETSGAYEK	665.3	669.3	2	884.4		892.4	1		y8		Y	21.6
P02760	AMBP	TVAACNLPIVR	607.3	612.3	2	484.3		494.3	1		y4		Y	19.8
P15144	AMPN	AQIINDAFNLASAHK	538.3	541	3	200.1	313.7	200.1	1	2	b2	y6	Y	14.6
P54802	ANAG	DFCGCHVAWSGSQLR	593.9	597.3	3	759.3		764.3	2		y13		Y	16.6
P01019	ANGT	DPTFIPAPIQAK	649.4	653.4	2	724.4		732.4	1		y7		Y	21.1
P01019	ANGT	LQAILGVPWK	562.8	566.9	2	883.5		891.6	1		y8		Y	18.4
P01019	ANGT	QPFVQGLALYTPVVLPR	633.4	636.7	3	680.4		690.5	1		y6		Y	18
P01019	ANGT	SLDFTELDVAAEK	719.4	723.4	2	316.2		316.2	1		b3		Y	23.3
P01019	ANGT	TSPVDEK	388.2	392.2	2	587.3		595.3	1		y5		Y	13
P01019	ANGT	VLSALQAVQGLLVAQGR	575	578.3	3	530.3		540.3	1		y5		Y	15.9
P01008	ANT3	ENAEQSR	417.2	422.2	2	390.2		400.2	1		y3		Y	13.9
P01008	ANT3	FATTFYQHLADSK	764.9	768.9	2	234.1	961.5	242.2	1	1	y2	y8	Y	24.7
P01008	ANT3	FDTISEK	420.2	424.2	2	692.3		700.4	1		y6		Y	14
P01008	ANT3	LQPLDFK	430.7	434.8	2	619.3		627.4	1		y5		Y	14.4
P01008	ANT3	VAEGTQVLELPFK	715.9	719.9	2	391.2		399.2	1		y3		Y	23.2
P01008	ANT3	VANPCVK	394.2	398.2	2	503.3		511.3	1		y4		Y	13.2
P01008	ANT3	VWELSK	381.2	385.2	2	476.3		484.3	1		y4		Y	12.8
P08519	APOA	LFLEPTQADIALLK	524.6	527.3	3	743.5		751.5	1		y7		Y	14.1
P08519	APOA	NPDAVAAPYCYTR	749.3	754.3	2	859.4		869.4	1		y6		Y	24.2
P08519	APOA	NPDPVAAPYCYTR	762.4	767.4	2	859.4		869.4	1		y6		Y	24.6
P02647	APOA1	AHVDALR	391.2	396.2	2	573.3		583.3	1		y5		Y	13.1
P02647	APOA1	DLATVYVDVLK	618.3	622.4	2	736.4		744.4	1		y6		Y	20.2
P02647	APOA1	EQLGPVTQEFWDNLEK	967	971	2	258.1	951.5	258.1	1	1	b2	y7	Y	31
P02647	APOA1	LLDNWDSVTSTFSK	806.9	810.9	2	971.5		979.5	1		y9		Y	26
P02647	APOA1	QGLLPVLESFK	410.9	413.6	3	623.3		631.4	1		y5		Y	10
P02647	APOA1	VQPYLDDFQK	626.8	630.8	2	228.1	765.4	228.1	1	1	b2	y6	Y	20.4
P02647	APOA1	VSFLSALEEYTK	693.9	697.9	2	853.4		861.4	1		y7		Y	22.5
P02652	APOA2	EPCVESLVSQYFQTVTDYGK	1175.5	1179.6	2	583.3		591.3	1		y5		Y	37.4
P02652	APOA2	SPELQAEAK	486.8	490.8	2	443.2		447.2	2		y8		Y	16.1
P06727	APOA4	GNTEGLQK	423.7	427.7	2	675.4		683.4	1		y6		Y	14.1
P06727	APOA4	IDQNVEELK	544.3	548.3	2	974.5		982.5	1		y8		Y	17.9

P06727	APOA4	IDQTVEELR	551.8	556.8	2	746.4		756.4	1		y6		Y	18.1
P06727	APOA4	ISASAEELR	488.3	493.3	2	775.4		785.4	1		y7		Y	16.1
P06727	APOA4	LGEVNTYAGDLQK	704.4	708.4	2	794.4		802.4	1		y7		Y	22.8
P06727	APOA4	VEPYGENFNK	598.8	602.8	2	484.7		488.7	2		y8		Y	19.6
P04114	APOB	ALVDTLK	380.2	384.2	2	575.3		583.4	1		y5		Y	12.8
P04114	APOB	FSVPAGIVIPSFQALTAR	937.5	942.5	2	1103.6		1113.6	1		y10		Y	30.1
P04114	APOB	ITLPDFR	431.2	436.2	2	534.3		544.3	1		y4		Y	14.4
P04114	APOB	LATALSLSNK	509.3	513.3	2	185.1	548.3	185.1	1	1	b2	y5	Y	16.8
P04114	APOB	QGFFPDSVNK	569.8	573.8	2	659.3		667.4	1		y6		Y	18.7
P04114	APOB	TSSFALNLPTLPEVK	808.9	813	2	472.3		480.3	1		y4		Y	26.1
P02655	APOC2	ESLSSYWESAK	643.8	647.8	2	957.4		965.4	1		y8		Y	21
P02655	APOC2	TAAQNLYEK	519.3	523.3	2	865.4		873.5	1		y7		Y	17.1
P02655	APOC2	TYLPAVDEK	518.3	522.3	2	658.3		666.4	1		y6		Y	17.1
P02656	APOC3	DALSSVQESQVAQQAR	858.9	863.9	2	573.3		583.3	1		y5		Y	27.6
P02656	APOC3	DYWSTVK	449.7	453.7	2	620.3		628.4	1		y5		Y	14.9
P02656	APOC3	GWVTDGFSSLK	598.8	602.8	2	244.1	854.4	244.1	1	1	b2	y8	Y	19.6
P05090	APOD	CPNPPVQENFDVNK	829.4	833.4	2	643.8		647.8	2		y11		Y	26.7
P05090	APOD	IPTTFENGR	517.8	522.8	2	461.2		466.2	2		y8		Y	17.1
P05090	APOD	NILTSNNIDVK	615.8	619.8	2	228.1	890.5	228.1	1	1	b2	y8	Y	20.1
P05090	APOD	NPNLPPETVDSLK	712.4	716.4	2	985.5		993.5	1		y9		Y	23.1
P05090	APOD	VLNQELR	436.3	441.3	2	659.3		669.4	1		y5		Y	14.5
P02649	APOE	AATVGSLAGQPLQER	749.4	754.4	2	827.4		837.4	1		y7		Y	24.2
P02649	APOE	AQAWGER	409.2	414.2	2	618.3		628.3	1		y5		Y	13.7
P02649	APOE	EQVAEVR	415.7	420.7	2	474.3		484.3	1		y4		Y	13.9
P02649	APOE	LAVYQAGAR	474.8	479.8	2	665.3		675.3	1		y6		Y	15.7
P02649	APOE	LQAEAFQAR	517.3	522.3	2	792.4		802.4	1		y7		Y	17
P02649	APOE	QQTEWQSGQR	624.3	629.3	2	761.4		771.4	1		y6		Y	20.4
P02649	APOE	QWAGLVEK	465.8	469.8	2	616.4		624.4	1		y6		Y	15.4
Q13790	APOF	SLPTEDCENEK	661.3	665.3	2	1121.4		1129.5	1		y9		Y	21.5
P02749	APOH	ATVVYQGER	511.8	516.8	2	652.3		662.3	1		y5		Y	16.9
P02749	APOH	CSYTEDAQCIDGTIEVPK	696	698.6	3	244.2	512.2	252.2	1	1	y2	b4	Y	20.3
P02749	APOH	FICPLTGLWPINTLK	887	891	2	421.2		421.2	1		b3		Y	28.5
P02749	APOH	VCPFAGILENGAVR	751.9	756.9	2	260.1	928.5	260.1	1	1	b2	y9	Y	24.3
P02749	APOH	VSFFCK	394.2	398.2	2	688.3		696.3	1		y5		Y	13.2
O14791	APOL1	ALDNLAR	386.7	391.7	2	588.3		598.3	1		y5		Y	13
O14791	APOL1	LNILNNNYK	553.3	557.3	2	228.1	652.3	228.1	1	1	b2	y5	Y	18.2

O14791	APOL1	VTEPISAESGEQVER	815.9	820.9	2	804.4		814.4	1		y7		Y	26.3
O95445	APOM	AFLLTPR	409.3	414.3	2	599.4		609.4	1		y5		Y	13.7
O95445	APOM	DGLCVPR	408.7	413.7	2	531.3		541.3	1		y4		Y	13.7
O95445	APOM	FLLYNR	413.2	418.2	2	565.3		575.3	1		y4		Y	13.8
O95445	APOM	SLTSCLDSK	505.7	509.8	2	810.4		818.4	1		y7		Y	16.7
O95445	APOM	WIYHLTEGSTDLR	530.9	534.3	3	288.2	648.3	298.2	1	1	y2	y6	Y	14.3
P00966	ASSY	IDIVENR	429.7	434.7	2	630.4		640.4	1		y5		Y	14.3
Q76LX8	ATS13	LFINVAPHAR	379.9	383.2	3	480.3		490.3	1		y4		Y	8.9
P61769	B2MG	VNHVTLSQPK	374.9	377.6	3	512.3		516.3	2		y9		Y	8.7
P02730	B3AT	LSVPDGFK	431.7	435.7	2	563.3		571.3	1		y5		Y	14.4
Q8TDL5	BPIB1	ALGFEAAESSLTK	662.3	666.4	2	806.4		814.4	1		y8		Y	21.5
P43251	BTD	LSSGLVTAALYGR	654.4	659.4	2	751.4		761.4	1		y7		Y	21.3
P43251	BTD	SHLIIAQVAK	360.6	363.2	3	451.3		451.3	1		b4		Y	8.2
P43251	BTD	VDLITFDTPFAGR	484.6	487.9	3	274.2		279.2	2		y5		Y	12.6
Q06187	BTK	LVQLYGVCTK	590.8	594.8	2	727.3		735.4	1		y6		Y	19.3
P02745	C1QA	SLGFCDTTNK	571.8	575.8	2	942.4		950.4	1		y8		Y	18.7
P02746	C1QB	IAFSATR	383.2	388.2	2	652.3		662.3	1		y6		Y	12.9
P02746	C1QB	LEQGENVFLQATDK	531.3	533.9	3	822.4		830.4	1		y7		Y	14.3
P02746	C1QB	TINVPLR	406.8	411.8	2	598.4		608.4	1		y5		Y	13.6
P02747	C1QC	FNAVLTNPQGDYDTSTGK	964.5	968.5	2	333.2		333.2	1		b3		Y	30.9
P02747	C1QC	FQSVFTVTR	542.8	547.8	2	809.5		819.5	1		y7		Y	17.8
P02747	C1QC	QTHQPPAPNSLIR	486.9	490.3	3	350.2		355.2	2		y6		Y	12.7
P02747	C1QC	TNQVNSGGVLLR	629.3	634.4	2	815.5		825.5	1		y8		Y	20.5
P02747	C1QC	VVTFCGHTSK	379.2	381.9	3	199.1	689.3	199.1	1	1	b2	y6	Y	8.9
P00736	C1R	FCGQLGSPLGNPPGK	764.9	768.9	2	398.2		406.3	1		y4		Y	24.7
P00736	C1R	GYGFYTK	418.2	422.2	2	615.3		623.3	1		y5		Y	14
P00736	C1R	NIGEFCGK	462.7	466.7	2	697.3		705.3	1		y6		Y	15.3
P00736	C1R	QDACQGDSGGVFAVR	783.9	788.9	2	964.5		974.5	1		y10		Y	25.3
P00736	C1R	VLNYVDWIK	575.3	579.3	2	937.5		945.5	1		y7		Y	18.8
P00736	C1R	YTTEIIK	434.2	438.3	2	603.4		611.4	1		y5		Y	14.5
Q9NZP8	C1RL	GSEAINAPGDNPAK	670.8	674.8	2	698.3		706.4	1		y7		Y	21.8
Q9NZP8	C1RL	WILTAAHTIYPK	707.4	711.4	2	1114.6		1122.6	1		y10		Y	22.9
P09871	C1S	CEYQIR	434.7	439.7	2	579.3		589.3	1		y4		Y	14.5
P09871	C1S	IIGGSDADIK	494.8	498.8	2	762.4		770.4	1		y8		Y	16.3
P09871	C1S	LQVIFK	374.2	378.2	2	506.3		514.3	1		y4		Y	12.6
P09871	C1S	SNALDIIFQTDLTGQK	882.5	886.5	2	273.1		273.1	1		b3		Y	28.4

P09871	C1S	SSNNPHSPIVEEFQVPYNK	729.4	732	3	521.3		529.3	1		y4		Y	21.5
P09871	C1S	TNFDNDIALVR	639.3	644.3	2	915.5		925.5	1		y8		Y	20.8
P04003	C4BPA	FSAICQGDGTWSPR	791.4	796.4	2	875.4		885.4	1		y8		Y	25.5
P04003	C4BPA	LSCSYSHWSAPAPQCK	626.9	629.6	3	700.3		708.4	1		y6		Y	17.8
P04003	C4BPA	YTCLPGYVR	564.8	569.8	2	591.3		601.3	1		y5		Y	18.5
P54289	CA2D1	VLLDAGFTNELVQNYWSK	699.7	702.4	3	825.4		833.4	1		y6		Y	20.4
P00915	CAH1	ESISVSSEQLAQFR	790.9	795.9	2	1065.5		1075.5	1		y9		Y	25.5
P00915	CAH1	GGPFSDSYR	493.2	498.2	2	774.3		784.3	1		y6		Y	16.3
P00915	CAH1	YSSLAEAASK	513.8	517.8	2	776.4		784.4	1		y8		Y	16.9
P00918	CAH2	YGDFGK	343.7	347.7	2	523.3		531.3	1		y5		Y	11.7
P27797	CALR	FVLSSGK	369.2	373.2	2	491.3		499.3	1		y5		Y	12.4
P27797	CALR	QIDNPDYK	496.7	500.7	2	522.3		530.3	1		y4		Y	16.4
P08185	CBG	HLVALSPK	432.8	436.8	2	251.2	614.4	251.2	1	1	b2	y6	Y	14.4
P08185	CBG	ITQDAQLK	458.8	462.8	2	702.4		710.4	1		y6		Y	15.2
P08185	CBG	QINSYVK	426.2	430.2	2	610.3		618.3	1		y5		Y	14.2
P22681	CBL	GTEPIVVDPFDPR	721.4	726.4	2	577.8		582.8	2		y10		Y	23.4
Q96IY4	CBPB2	DTGTYGFLLPER	684.8	689.8	2	401.2		411.2	1		y3		Y	22.2
Q96IY4	CBPB2	YPLYVLK	448.3	452.3	2	366.7		370.7	2		y6		Y	14.9
P15169	CBPN	IVQLIQDTR	543.3	548.3	2	873.5		883.5	1		y7		Y	17.8
P15169	CBPN	VQNECPGITR	587.3	592.3	2	946.4		956.4	1		y8		Y	19.2
P15169	CBPN	VYSIGR	347.7	352.7	2	432.3		442.3	1		y4		Y	11.8
P15169	CBPN	YDDLVR	390.7	395.7	2	502.3		512.3	1		y4		Y	13.1
P15169	CBPN	YGGPNHHLPLPDNWK	582.3	585	3	659.3		667.3	1		y5		Y	16.2
P30279	CCND2	ACQEQIEAVLLNSLQQYR	721.7	725	3	908.5		918.5	1		y7		Y	21.2
P08571	CD14	ATVNPSAPR	456.7	461.8	2	527.3		537.3	1		y5		Y	15.2
P08571	CD14	VLAYSR	354.7	359.7	2	496.3		506.3	1		y4		Y	12
P08571	CD14	VLDLSCNR	488.7	493.8	2	764.3		774.3	1		y6		Y	16.2
O43866	CD5L	CYGPGVGR	433.2	438.2	2	542.3		552.3	1		y6		Y	14.4
O43866	CD5L	LVGGDNLCSGR	574.3	579.3	2	935.4		945.4	1		y9		Y	18.8
P06731	CEAM5	TLTLFNVTR	532.8	537.8	2	850.5		860.5	1		y7		Y	17.5
P00450	CERU	DIASGLIGPLIICK	735.4	739.4	2	800.5		808.5	1		y7		Y	23.8
P00450	CERU	EVGPTNADPVCLAK	735.9	739.9	2	802.4		810.4	1		y7		Y	23.8
P00450	CERU	EYTDASFTNR	602.3	607.3	2	624.3		634.3	1		y5		Y	19.7
P00450	CERU	GAYPLSIEPIGVR	457.9	461.3	3	541.3		551.4	1		y5		Y	11.7
P00450	CERU	IGGSYK	312.7	316.7	2	511.3		519.3	1		y5		Y	10.7
P00450	CERU	QSEDSTFYLGER	716.3	721.3	2	637.3		647.3	1		y5		Y	23.2

P00751	CFAB	CLVNLIEK	494.8	498.8	2	715.4		723.4	1		y6		Y	16.3
P00751	CFAB	DISEVVTPR	508.3	513.3	2	787.4		797.4	1		y7		Y	16.8
P00751	CFAB	STGSWSTLK	483.7	487.8	2	778.4		786.4	1		y7		Y	16
P00751	CFAB	VSEADSSNADWVTK	754.8	758.9	2	248.2	920.4	256.2	1	1	y2	y8	Y	24.4
P00751	CFAB	YGLVTYATYPK	638.3	642.3	2	843.4		851.4	1		y7		Y	20.8
P08603	CFAH	CVEISCK	448.2	452.2	2	636.3		644.3	1		y5		Y	14.9
P08603	CFAH	DGWSAQPTCIK	631.8	635.8	2	904.5		912.5	1		y8		Y	20.6
P08603	CFAH	EYHFGQAVR	369.5	372.9	3	274.2	530.3	284.2	1	1	y2	y5	Y	8.5
P08603	CFAH	TGDEITYQCR	621.8	626.8	2	727.3		737.3	1		y5		Y	20.3
P08603	CFAH	TGESVEFVCK	578.3	582.3	2	682.3		690.3	1		y5		Y	18.9
P08603	CFAH	VGEVLK	322.7	326.7	2	545.3		553.3	1		y5		Y	11
P08603	CFAH	WQSIPLCVEK	630.3	634.3	2	745.4		753.4	1		y6		Y	20.5
P05156	CFAI	ACDGINDCGDQSDELCCK	706.3	708.9	3	467.2		475.2	1		y3		Y	20.6
P05156	CFAI	AQLGDLPWQVAIK	719.9	723.9	2	200.1	841.5	200.1	1	1	b2	y7	Y	23.3
P05156	CFAI	EANVACLDLGFQQGADTQR	698.3	701.7	3	775.4		785.4	1		y7		Y	20.3
P05156	CFAI	GLETSLAECTFTK	728.9	732.9	2	856.4		864.4	1		y7		Y	23.6
P05156	CFAI	HGNTDSEGIVEVK	462.2	464.9	3	644.4		652.4	1		y6		Y	11.8
P06276	CHLE	AEEILSR	409.2	414.2	2	617.4		627.4	1		y5		Y	13.7
P06276	CHLE	IFFPGVSEFGK	614.3	618.3	2	820.4		828.4	1		y8		Y	20
P06276	CHLE	TQILVGVNK	486.3	490.3	2	230.1	629.4	230.1	1	1	b2	y6	Y	16.1
P10909	CLUS	ASSIIDELFQDR	697.4	702.4	2	922.4		932.4	1		y7		Y	22.6
P10909	CLUS	EIQNAVNGVK	536.3	540.3	2	701.4		709.4	1		y7		Y	17.6
P10909	CLUS	ELDESLQVAER	644.8	649.8	2	802.4		812.5	1		y7		Y	21
P10909	CLUS	TLLSNLEEAK	559.3	563.3	2	215.1	790.4	215.1	1	1	b2	y7	Y	18.3
Q96KN2	CNDP1	AIHLDLEEYR	420.2	423.6	3	467.2		477.2	1		y3		Y	10.3
Q96KN2	CNDP1	EWVAIESDSVQPVPR	856.4	861.4	2	468.3		478.3	1		y4		Y	27.5
P02452	CO1A1	VLCDDVICDETK	733.8	737.8	2	627.8		631.8	2		y10		Y	23.7
P06681	CO2	AVISPGFDVFAK	625.8	629.8	2	880.5		888.5	1		y8		Y	20.4
P06681	CO2	CSSNLVLTGSSER	705.3	710.3	2	749.4		759.4	1		y7		Y	22.9
P06681	CO2	EILNINQK	486.3	490.3	2	616.3		624.4	1		y5		Y	16.1
P06681	CO2	GALISDQWVLTAAHCFR	649	652.3	3	242.1		242.1	1		b3		Y	18.6
P06681	CO2	HAFILQDTK	536.8	540.8	2	209.1	864.5	209.1	1	1	b2	y7	Y	17.6
P06681	CO2	LNINLK	357.7	361.7	2	601.4		609.4	1		y5		Y	12.1
P06681	CO2	SSGQWQTPGATR	638.3	643.3	2	501.3		511.3	1		y5		Y	20.8
P01024	CO3	ACEPGVDYVYK	650.8	654.8	2	940.5		948.5	1		y8		Y	21.2
P01024	CO3	DSCVGSLVVK	532.3	536.3	2	602.4		610.4	1		y6		Y	17.5

P01024	CO3	IWDVVEK	444.7	448.8	2	589.3		597.3	1		y5		Y	14.8
P01024	CO3	NEQVEIR	444.2	449.2	2	644.4		654.4	1		y5		Y	14.8
P01024	CO3	VLLDGVQNPR	555.8	560.8	2	898.5		908.5	1		y8		Y	18.2
P0C0L4	CO4A	ANSFLGEK	433.2	437.2	2	446.3		454.3	1		y4		Y	14.4
P0C0L4	CO4A	DSSTWLTAFVLK	684.4	688.4	2	791.5		799.5	1		y7		Y	22.2
P08572	CO4A2	IAVQPGTVGPQGR	427.2	430.6	3	457.3		467.3	1		y4		Y	10.6
P01031	CO5	AFDICPLVK	531.8	535.8	2	844.5		852.5	1		y7		Y	17.5
P01031	CO5	EYVLPHFSVSIEPEYNFIGYK	844.4	847.1	3	1130.6		1138.6	1		y9		Y	25.6
P01031	CO5	IDTALIK	387.2	391.2	2	660.4		668.4	1		y6		Y	13
P01031	CO5	IDTQDIEASHYR	483.2	486.6	3	667.8		672.8	2		y11		Y	12.6
P01031	CO5	ITHYNYLILSK	455.6	458.3	3	460.3		468.3	1		y4		Y	11.6
P01031	CO5	NFEITIK	432.7	436.8	2	603.4		611.4	1		y5		Y	14.4
P01031	CO5	QLPGGQNPVSYVYLEVVSK	1039.1	1043.1	2	837.5		845.5	1		y7		Y	33.2
P01031	CO5	TLLPVSKPEIR	418.3	421.6	3	463.3		468.3	2		y8		Y	10.3
P01031	CO5	VFQFLEK	455.8	459.8	2	664.4		672.4	1		y5		Y	15.1
P13671	CO6	ALNHLPLEYNSALYSR	621	624.3	3	538.3		548.3	1		y4		Y	17.6
P13671	CO6	DLHLSDVFLK	593.8	597.8	2	479.8		483.8	2		y8		Y	19.4
P13671	CO6	ENPAVIDFELAPIVDLVR	1005.5	1010.6	2	811.5		821.5	1		y7		Y	32.2
P13671	CO6	GEVLDNSFTGGICK	748.9	752.9	2	286.1		286.1	1		b3		Y	24.2
P13671	CO6	GFVVAGPSR	445.2	450.3	2	487.3		497.3	1		y5		Y	14.8
P13671	CO6	SEYGAALAWEK	612.8	616.8	2	845.5		853.5	1		y8		Y	20
P13671	CO6	TLNICEVGTIR	638.3	643.3	2	834.4		844.4	1		y7		Y	20.8
P10643	CO7	AASGTQNNVLR	565.8	570.8	2	143.1	743.4	143.1	1	1	b2	y6	Y	18.5
P10643	CO7	DSCTLPASAEK	589.8	593.8	2	602.3		610.3	1		y6		Y	19.3
P10643	CO7	ELSHLPSLYDYSAYR	605.3	608.6	3	774.3		784.3	1		y6		Y	17
P10643	CO7	SCVGETTESTQCEDEELEHLR	837	840.4	3	248.1	425.3	248.1	1	1	b2	y3	Y	25.3
P10643	CO7	VLFYVDSEK	550.3	554.3	2	887.4		895.4	1		y7		Y	18.1
P10643	CO7	YSAWAESVTNLPQVIK	903.5	907.5	2	584.4		592.4	1		y5		Y	29
P07357	CO8A	AIDEDCSQYEPIPGSQK	968.9	972.9	2	516.3		524.3	1		y5		Y	31
P07357	CO8A	LGSLGAACEQTQTEGAK	860.9	864.9	2	275.2		283.2	1		y3		Y	27.7
P07357	CO8A	LYYGDDEK	501.7	505.7	2	726.3		734.3	1		y6		Y	16.6
P07357	CO8A	STITYR	370.7	375.7	2	552.3		562.3	1		y4		Y	12.5
P07358	CO8B	CEGFVCAQTGR	642.8	647.8	2	692.3		702.3	1		y6		Y	20.9
P07358	CO8B	IPGIFELGISSQSDR	809.9	814.9	2	849.4		859.4	1		y8		Y	26.1
P07358	CO8B	LPLEYSYGEYR	695.3	700.3	2	211.1	774.3	211.1	1	1	b2	y6	Y	22.6
P07358	CO8B	QALEEFQK	496.8	500.8	2	680.3		688.3	1		y5		Y	16.4

P07358	CO8B	SGFSFGFK	438.7	442.7	2	585.3		593.3	1		y5		Y	14.6
P07358	CO8B	SVFLHAR	415.2	420.2	2	643.4		653.4	1		y5		Y	13.9
P07360	CO8G	AGQLSVK	351.7	355.7	2	631.4		639.4	1		y6		Y	11.9
P07360	CO8G	LDGICWQVR	573.8	578.8	2	748.4		758.4	1		y5		Y	18.8
P07360	CO8G	QLYGDTGVLGR	589.8	594.8	2	774.4		784.4	1		y8		Y	19.3
P07360	CO8G	SLPVSDSVLSGFEQR	810.9	815.9	2	836.4		846.4	1		y7		Y	26.1
P07360	CO8G	YGFCEAADQFHVLDEVR	686	689.3	3	221.1	403.2	221.1	1	1	b2	y3	Y	19.9
P02748	CO9	CLCACPFK	528.2	532.2	2	782.3		790.3	1		y6		Y	17.4
P02748	CO9	FEGIACEISK	577.3	581.3	2	877.4		885.5	1		y8		Y	18.9
P02748	CO9	SIEVFGQFNGK	613.3	617.3	2	797.4		805.4	1		y7		Y	20
P02748	CO9	TSNFNAAISLK	583.3	587.3	2	716.4		724.4	1		y7		Y	19.1
P02748	CO9	VVEESELAR	516.3	521.3	2	833.4		843.4	1		y7		Y	17
Q03692	COAA1	GTHVWVGLYK	387.2	389.9	3	480.3		488.3	1		y4		Y	9.1
Q9UMD9	COHA1	QAAYNADSGLK	569.3	573.3	2	704.4		712.4	1		y7		Y	18.6
P49747	COMP	DTDLDGFPDEK	626.3	630.3	2	488.2		496.2	1		y4		Y	20.4
P49747	COMP	EITFLK	375.7	379.7	2	508.3		516.3	1		y4		Y	12.6
P20815	CP3A5	DTINFLSK	469.3	473.3	2	721.4		729.4	1		y6		Y	15.5
P22792	CPN2	DHLGFQVTWPDESK	553.6	556.3	3	575.3		583.3	1		y5		Y	15.1
P22792	CPN2	GQVVPALNEK	527.8	531.8	2	671.4		679.4	1		y6		Y	17.4
P22792	CPN2	LTVSIEAR	444.8	449.8	2	575.3		585.3	1		y5		Y	14.8
P22792	CPN2	QLVCPVTR	486.8	491.8	2	731.4		741.4	1		y6		Y	16.1
P02741	CRP	ESDTSYVSLK	564.8	568.8	2	347.2		355.2	1		y3		Y	18.5
P02741	CRP	GYSIFSYATK	568.8	572.8	2	716.4		724.4	1		y6		Y	18.6
P02775	CXCL7	GTHCNQVEVIATLK	523.9	526.6	3	773.5		781.5	1		y7		Y	14.1
P02775	CXCL7	ICLDPDAPR	528.8	533.8	2	783.4		793.4	1		y7		Y	17.4
P02775	CXCL7	NIQSLEVIGK	550.8	554.8	2	873.5		881.5	1		y8		Y	18.1
P02775	CXCL7	TTSGIHPK	420.7	424.7	2	638.4		646.4	1		y6		Y	14
P01034	CYTC	ALDFAVGEYNK	613.8	617.8	2	780.4		788.4	1		y7		Y	20
O95822	DCMC	LCAWYLYGEK	651.8	655.8	2	772.4		780.4	1		y6		Y	21.2
P09172	DOPO	VISTLEEPTPQCPTSQGR	1000.5	1005.5	2	1228.6		1238.6	1		y11		Y	32
Q14126	DSG2	ILDVNDNIPVVENK	528	530.6	3	685.4		693.4	1		y6		Y	14.2
P32926	DSG3	LAEISLGVDGEGK	644.3	648.4	2	861.4		869.4	1		y9		Y	21
Q16610	ECM1	ELPSLQHPNEQK	473.9	476.6	3	540.8		544.8	2		y9		Y	12.3
Q16610	ECM1	EVGPPLPQEAVPLQK	801.4	805.5	2	485.3		493.3	1		y4		Y	25.8
Q16610	ECM1	FCEAEFSVK	558.8	562.8	2	809.4		817.4	1		y7		Y	18.3
Q16610	ECM1	FSCFQEEAPQPHYQLR	679.6	683	3	519.8		524.8	2		y8		Y	19.7

Q16610	ECM1	LTFINDLCGPR	653.3	658.3	2	831.4		841.4	1		y7		Y	21.3
Q16610	ECM1	NVALVSGDTENAK	659.3	663.3	2	821.4		829.4	1		y8		Y	21.4
Q16610	ECM1	QHVVYGPWNLPQSSYSHLTR	790.4	793.7	3	588.3		593.3	2		y10		Y	23.7
P00533	EGFR	CNLLEGEPR	544.3	549.3	2	272.2	501.2	282.2	1	1	y2	b4	Y	17.9
Q01780	EXOSX	SGPLPSAER	457.2	462.2	2	559.3		569.3	1		y5		Y	15.2
P00488	F13A	CGPASVQAIK	515.8	519.8	2	813.5		821.5	1		y8		Y	17
P00488	F13A	FQEGQEEER	576.3	581.3	2	876.4		886.4	1		y7		Y	18.9
P00488	F13A	GTYIPVPIVSELQSGK	844.5	848.5	2	322.1		322.1	1		b3		Y	27.2
P00488	F13A	LSIQSSPK	430.2	434.3	2	746.4		754.4	1		y7		Y	14.3
P00488	F13A	STVLTIPEIIIK	663.9	667.9	2	712.5		720.5	1		y6		Y	21.6
P05160	F13B	GDTYPAELYITGSILR	885	890	2	922.5		932.5	1		y8		Y	28.4
P05160	F13B	IQTHSTTYR	369.5	372.9	3	433.2		438.2	2		y7		Y	8.5
P05160	F13B	SGYLLHGSNEITCNR	574.3	577.6	3	550.2		560.2	1		y4		Y	15.9
P05160	F13B	VACEEPPFIENGAANLHSK	695	697.7	3	171.1	797.4	171.1	1	1	b2	y8	Y	20.2
P05160	F13B	VLHGDLIDFVCK	472.6	475.3	3	635.4		635.4	1		b6		Y	12.2
P05160	F13B	VQYECATGYYTAGGK	834.4	838.4	2	228.1	433.2	228.1	1	1	b2	y5	Y	26.9
P00742	FA10	ACIPTGPYPCGK	660.8	664.8	2	976.5		984.5	1		y9		Y	21.5
P00742	FA10	GYTLADNGK	469.7	473.7	2	504.2		512.3	1		y5		Y	15.6
P00742	FA10	NCELFTR	470.2	475.2	2	536.3		546.3	1		y4		Y	15.6
P00742	FA10	QEDACQGDSGGPHVTR	571.9	575.2	3	405.7		410.7	2		y8		Y	15.8
P00742	FA10	TGIVSGFGR	447.2	452.2	2	622.3		632.3	1		y6		Y	14.9
P00742	FA10	TNEFWNK	469.7	473.7	2	723.3		731.4	1		y5		Y	15.6
P03951	FA11	ALSGFSLQSCR	613.3	618.3	2	750.4		760.4	1		y6		Y	20
P03951	FA11	GGISGYTLR	462.3	467.3	2	696.4		706.4	1		y6		Y	15.3
P03951	FA11	SCALSNLACIR	632.8	637.8	2	319.1		319.1	1		b3		Y	20.6
P03951	FA11	TAAISGYSFK	522.8	526.8	2	688.3		696.3	1		y6		Y	17.2
P03951	FA11	VVSGFSLK	418.7	422.8	2	638.4		646.4	1		y6		Y	14
P00748	FA12	CFEPQLLR	531.8	536.8	2	626.4		636.4	1		y5		Y	17.5
P00748	FA12	LHEAFSPVSYQHDLALLR	699.4	702.7	3	251.2	837.5	251.2	1	1	b2	y7	Y	20.4
P00748	FA12	NGPLSCGQR	494.7	499.7	2	409.2		414.2	2		y7		Y	16.3
P00748	FA12	NWGLGGHAFCR	425.5	428.9	3	487.7		492.7	2		y9		Y	10.5
P00748	FA12	TEQAAVAR	423.2	428.2	2	615.4		625.4	1		y6		Y	14.1
P00748	FA12	VVGGLVALR	442.3	447.3	2	685.4		695.4	1		y7		Y	14.7
P12259	FA5	AVQPGETYTYK	628.8	632.8	2	299.2		299.2	1		b3		Y	20.5
P12259	FA5	EFNPLVIVGLSK	658.4	662.4	2	925.6		933.6	1		y9		Y	21.4
P12259	FA5	EVIITGIQTQGAK	679.4	683.4	2	903.5		911.5	1		y9		Y	22.1

P12259	FA5	GEYEEHLGILGPIIR	566	569.3	3	668.4		678.5	1		y6		Y	15.6
P12259	FA5	LAAALGIR	392.8	397.8	2	600.4		610.4	1		y6		Y	13.2
P12259	FA5	NFFNPPIISR	602.8	607.8	2	796.5		806.5	1		y7		Y	19.7
P08709	FA7	LHQPVVLTDHVVPLCLPER	741.4	744.7	3	251.2	585.3	251.2	1	1	b2	b5	Y	21.9
P00740	FA9	NCELDVTCNIK	683.3	687.3	2	275.1	404.1	275.1	1	1	b2	b3	Y	22.2
P00740	FA9	SALVLQYLR	531.8	536.8	2	692.4		702.4	1		y5		Y	17.5
P00740	FA9	SCEPAVPFPCGR	688.8	693.8	2	1000.5		1010.5	1		y9		Y	22.4
P00740	FA9	VSVSQTSK	418.2	422.2	2	550.3		558.3	1		y5		Y	14
P00740	FA9	VVCSCTEGYR	615.8	620.8	2	1032.4		1042.4	1		y8		Y	20.1
P23142	FBLN1	CVDVDECAPPAEPCGK	902.4	906.4	2	855.4		863.4	1		y8		Y	29
P23142	FBLN1	SQETGDLDVGGLQETDK	896.4	900.4	2	847.4		855.4	1		y8		Y	28.8
P23142	FBLN1	TGYYFDGISR	589.8	594.8	2	694.4		704.4	1		y6		Y	19.3
Q12805	FBLN3	NPCQDPYILTPENR	858.9	863.9	2	515.3		525.3	1		y4		Y	27.6
Q12805	FBLN3	SGNENGEFYLR	643.3	648.3	2	784.4		794.4	1		y6		Y	20.9
P35556	FBN2	FNLSHLGSK	501.8	505.8	2	262.1	628.3	262.1	1	1	b2	y6	Y	16.6
P22087	FBRL	NGGHFVISIK	536.3	540.3	2	706.4		714.5	1		y6		Y	17.6
P08637	FCG3A	AVVFLEPQWYR	704.4	709.4	2	749.4		759.4	1		y5		Y	22.8
Q9Y6R7	FCGBP	AIGYATAADCGR	613.3	618.3	2	821.4		831.4	1		y8		Y	20
Q9Y6R7	FCGBP	LASVSVSR	409.7	414.7	2	634.4		644.4	1		y6		Y	13.7
Q9Y6R7	FCGBP	VNGVLTALPVSVADGR	784.4	789.4	2	913.5		923.5	1		y9		Y	25.3
O75636	FCN3	YAVSEAAAHK	349.5	352.2	3	235.1	497.3	235.1	1	1	b2	y5	Y	7.8
O75636	FCN3	YGIDWASGR	512.7	517.8	2	691.3		701.3	1		y6		Y	16.9
P02765	FETUA	CNLLAEK	424.2	428.2	2	573.4		581.4	1		y5		Y	14.2
P02765	FETUA	EATEAAK	360.2	364.2	2	289.2		297.2	1		y3		Y	12.2
P02765	FETUA	EHAVEGDCDFQLLK	554.3	556.9	3	147.1		155.1	1		y1		Y	15.2
P02765	FETUA	FSVVYAK	407.2	411.2	2	666.4		674.4	1		y6		Y	13.6
P02765	FETUA	TVVQPSVGAAAGPVVPPCPGR	672.7	676	3	342.2		347.2	2		y6		Y	19.4
P02765	FETUA	VVHAAK	312.7	316.7	2	426.2		434.3	1		y4		Y	10.7
Q9UGM5	FETUB	DGYVLR	361.7	366.7	2	387.3		397.3	1		y3		Y	12.2
Q9UGM5	FETUB	SQASSCSLQSSDSVPVGLCK	1049	1053	2	673.4		681.4	1		y6		Y	33.5
Q9UGM5	FETUB	VNDAQEYR	497.7	502.7	2	781.3		791.4	1		y6		Y	16.4
Q03591	FHR1	STDTSCVNPPTVQNAHILSR	733	736.4	3	618.3		623.4	2		y11		Y	21.6
Q03591	FHR1	TGESAEFVCK	564.3	568.3	2	840.4		848.4	1		y7		Y	18.5
Q02985	FHR3	AQTTVTCTEK	569.8	573.8	2	200.1	638.3	200.1	1	1	b2	y5	Y	18.7
Q9BXR6	FHR5	IAGVNIK	357.7	361.7	2	601.4		609.4	1		y6		Y	12.1
Q9BXR6	FHR5	LQGSVTVTCR	560.8	565.8	2	879.4		889.4	1		y8		Y	18.4

Q9BXR6	FHR5	TGDAVEFQCK	577.8	581.8	2	274.1		274.1	1		b3		Y	18.9
P02671	FIBA	DLLPSR	350.7	355.7	2	359.2		369.2	1		y3		Y	11.9
P02671	FIBA	DNTYNR	391.7	396.7	2	553.3		563.3	1		y4		Y	13.1
P02671	FIBA	ESSSHHPGIAEFPSR	819.4	824.4	2	973.5		983.5	1		y9		Y	26.4
P02671	FIBA	GDFSSANNR	484.2	489.2	2	648.3		658.3	1		y6		Y	16
P02671	FIBA	GLIDEVNQDFTNR	760.9	765.9	2	894.4		904.4	1		y7		Y	24.6
P02671	FIBA	GSESGIFTNTK	570.8	574.8	2	610.3		618.3	1		y5		Y	18.7
P02671	FIBA	VQHIQLLQK	553.8	557.8	2	879.5		887.6	1		y7		Y	18.2
P02675	FIBB	DNENVVNEYSSELEK	884.9	888.9	2	572.2		572.2	1		b5		Y	28.4
P02675	FIBB	HQLYIDETVNSNIPTNLR	1064	1069	2	600.3		610.4	1		y5		Y	34
P02675	FIBB	NYCGLPGEYWLGNDK	893.4	897.4	2	1178.5		1186.6	1		y10		Y	28.7
P02675	FIBB	QGFGNVATNTDGK	654.8	658.8	2	706.3		714.4	1		y7		Y	21.3
P02675	FIBB	SILENLR	422.7	427.8	2	644.4		654.4	1		y5		Y	14.1
P02679	FIBG	ASTPNGYDNGIIWATWK	947.5	951.5	2	804.4		812.5	1		y6		Y	30.4
P02679	FIBG	DNCCILDER	597.7	602.8	2	532.3		542.3	1		y4		Y	19.5
P02679	FIBG	EGFGHLSPTGTTEFWLGNEK	736.4	739	3	447.2		455.2	1		y4		Y	21.7
P02679	FIBG	VELEDWNGR	559.3	564.3	2	229.1	776.3	229.1	1	1	b2	y6	Y	18.3
P02679	FIBG	YEASILTHDSSIR	746.4	751.4	2	815.4		825.4	1		y7		Y	24.1
P02679	FIBG	YLQEIYNSNNQK	757.4	761.4	2	867.4		875.4	1		y7		Y	24.5
P02751	FINC	DLQFVEVTDVK	646.8	650.8	2	789.4		797.4	1		y7		Y	21.1
P02751	FINC	FLATTPNSLLVSWQPPR	964	969	2	369.2		379.2	1		y3		Y	30.9
P02751	FINC	GEWTCIAYSQLR	742.4	747.4	2	503.3		513.3	1		y4		Y	24
P02751	FINC	HTSVQTTSSGSGPFTDVR	622	625.3	3	734.4		744.4	1		y6		Y	17.6
P02751	FINC	IGDTWR	374.2	379.2	2	634.3		644.3	1		y5		Y	12.6
P02751	FINC	ISCTIANR	467.7	472.7	2	734.4		744.4	1		y6		Y	15.5
P02751	FINC	IYLYTLNDNAR	678.4	683.4	2	277.2	803.4	277.2	1	1	b2	y7	Y	22
P02751	FINC	LLCQCLGFGSGHFR	551.3	554.6	3	660.3		670.3	1		y6		Y	15
P02751	FINC	LTVGLTR	380.2	385.2	2	545.3		555.3	1		y5		Y	12.8
P02751	FINC	QDGHLWCSTTSNYEQDQK	733	735.7	3	810.4		818.4	1		y6		Y	21.6
P02751	FINC	QYNVGPSVSK	539.8	543.8	2	574.3		582.3	1		y6		Y	17.7
P02751	FINC	SSPVVIDASTAIDAPSNLR	957	962	2	586.3		596.3	1		y5		Y	30.7
P02751	FINC	SYTITGLQPGTDYK	772.4	776.4	2	680.3		688.3	1		y6		Y	24.9
P02751	FINC	TYLGNALVCTCYGGSR	896.4	901.4	2	960.4		970.4	1		y8		Y	28.8
P02751	FINC	VPGTSTSATLTGLTR	731.4	736.4	2	761.5		771.5	1		y7		Y	23.7
P02751	FINC	YSFCTDHTVLVQTR	576.3	579.6	3	782.4		787.4	2		y13		Y	15.9
Q06787	FMR1	EPCCWWLAK	625.3	629.3	2	218.1	863.4	226.2	1	1	y2	y6	Y	20.4

O95954	FTCD	SDLQVAAK	416.2	420.2	2	629.4		637.4	1		y6		Y	13.9
P06396	GELS	EVQGFESATFLGYFK	861.9	865.9	2	875.5		883.5	1		y7		Y	27.7
P06396	GELS	HVVPNEVVVQR	425.9	429.2	3	501.3		511.3	1		y4		Y	10.5
P06396	GELS	QTQVSVLPEGGETPLFK	915.5	919.5	2	1074.5		1082.6	1		y10		Y	29.4
P06396	GELS	SEDCFILDHGK	660.8	664.8	2	569.3		577.3	1		y5		Y	21.5
P06396	GELS	TGAQELLR	444.3	449.3	2	159.1	658.4	159.1	1	1	b2	y5	Y	14.8
P06396	GELS	TPSAAYLWVGTGASEAEK	919.5	923.5	2	849.4		857.4	1		y9		Y	29.5
P06396	GELS	YIETDPANR	539.8	544.8	2	802.4		812.4	1		y7		Y	17.7
Q92820	GGH	YLESAGAR	433.7	438.7	2	590.3		600.3	1		y6		Y	14.4
P22352	GPX3	FYTFLK	409.7	413.7	2	508.3		516.3	1		y4		Y	13.7
P22352	GPX3	NSCPPTSELLGTSDR	817.4	822.4	2	636.8		641.8	2		y12		Y	26.3
P22352	GPX3	TTVSNVK	374.7	378.7	2	203.1	546.3	203.1	1	1	b2	y5	Y	12.6
Q14520	HABP2	FCEIGSDDCYVGDGYSYR	1081.9	1086.9	2	817.3		827.4	1		y7		Y	34.5
Q14520	HABP2	IYGGFK	342.7	346.7	2	408.2		416.2	1		y4		Y	11.6
Q14520	HABP2	LIANTLCNSR	581.3	586.3	2	1048.5		1058.5	1		y9		Y	19
Q14520	HABP2	LPGFDSCGK	490.7	494.7	2	434.2		438.2	2		y8		Y	16.2
P69905	HBA	VGAHAGEYGAEALER	510.6	513.9	3	617.3		627.3	1		y5		Y	13.6
P08397	HEM3	ELEHALEK	484.8	488.8	2	726.4		734.4	1		y6		Y	16
P02790	HEMO	ELISER	373.7	378.7	2	391.2		401.2	1		y3		Y	12.6
P02790	HEMO	GGYTLVSGYPK	571.3	575.3	2	650.4		658.4	1		y6		Y	18.7
P02790	HEMO	LLQDEFPGIPSPLDAAVECHR	788.7	792.1	3	676.3		681.3	2		y12		Y	23.6
P02790	HEMO	LYLVQGTQVYVFLTK	886.5	890.5	2	277.2	390.2	277.2	1	1	b2	b3	Y	28.5
P02790	HEMO	NFPSPVDAAFR	610.8	615.8	2	775.4		785.4	1		y7		Y	19.9
P02790	HEMO	SGAQATWTELPWPHEK	613.3	616	3	793.4		801.4	1		y6		Y	17.3
P02790	HEMO	YYCFQGNQFLR	748.3	753.3	2	862.5		872.5	1		y7		Y	24.2
P05546	HEP2	NFGYTLR	435.7	440.7	2	609.3		619.3	1		y5		Y	14.5
P05546	HEP2	NYNLVESLK	540.3	544.3	2	802.5		810.5	1		y7		Y	17.7
P05546	HEP2	QFPILLDFK	560.8	564.8	2	845.5		853.5	1		y7		Y	18.4
P05546	HEP2	SVNDLYIQK	540.3	544.3	2	893.5		901.5	1		y7		Y	17.7
P05546	HEP2	TLEAQLTPR	514.8	519.8	2	814.4		824.5	1		y7		Y	17
Q04756	HGFA	LCNIEPDER	573.3	578.3	2	872.4		882.4	1		y7		Y	18.8
Q04756	HGFA	LEACESLTR	539.8	544.8	2	836.4		846.4	1		y7		Y	17.7
Q04756	HGFA	SQFVQPICLPEPGSTFPAGHK	766.4	769.1	3	216.1	499.8	216.1	1	2	b2	y10	Y	22.8
Q04756	HGFA	TTDVTQTFGIEK	670.3	674.3	2	923.5		931.5	1		y8		Y	21.8
Q04756	HGFA	VANYVDWINDR	682.8	687.8	2	818.4		828.4	1		y6		Y	22.2
P00738	HPT	VGYVSGWGR	490.8	495.8	2	562.3		572.3	1		y5		Y	16.2

P00738	HPT	VSVNER	352.2	357.2	2	604.3		614.3	1		y5		Y	11.9
P00738	HPT	VTSIQDWVQK	602.3	606.3	2	1003.5		1011.5	1		y8		Y	19.7
P00739	HPTR	VVLHPNYHQVDIGLIK	615.7	618.4	3	698.9		702.9	2		y12		Y	17.4
P04196	HRG	ADLFYDVEALDLESPK	609	611.6	3	688.4		696.4	1		y6		Y	17.1
P04196	HRG	DGYLFQLLR	562.8	567.8	2	676.4		686.4	1		y5		Y	18.4
P04196	HRG	GGEGTGYFVDFSVR	745.8	750.9	2	869.5		879.5	1		y7		Y	24.1
P04196	HRG	QIGSVYR	411.7	416.7	2	581.3		591.3	1		y5		Y	13.8
P04196	HRG	YWNDCEPPDSR	719.8	724.8	2	571.3		581.3	1		y5		Y	23.3
P18065	IBP2	GECWCVNPNTGK	711.3	715.3	2	516.3		524.3	1		y5		Y	23.1
P18065	IBP2	LIQGAPTIR	484.8	489.8	2	742.4		752.4	1		y7		Y	16
P17936	IBP3	ALAQCAPPPAVCAELVR	912	917	2	1208.6		1218.7	1		y11		Y	29.3
P17936	IBP3	ETEYGPCR	506.2	511.2	2	489.2		499.2	1		y4		Y	16.7
P17936	IBP3	FLNVLSPR	473.3	478.3	2	685.4		695.4	1		y6		Y	15.7
P17936	IBP3	YGQPLPGYTTK	612.8	616.8	2	876.5		884.5	1		y8		Y	20
P24593	IBP5	AVYLPNCDR	554.3	559.3	2	661.3		671.3	1		y5		Y	18.2
P24593	IBP5	GVCLNEK	410.2	414.2	2	663.3		671.3	1		y5		Y	13.7
P05155	IC1	FQPTLLTLPR	395.9	399.2	3	486.3		496.3	1		y4		Y	9.5
P05155	IC1	GVTSVSQIFHSPDLAIR	914	919	2	771.4		781.4	1		y7		Y	29.3
P05155	IC1	LLDSLPSDTR	558.8	563.8	2	575.3		585.3	1		y5		Y	18.3
P05155	IC1	TLYSSSPR	455.7	460.7	2	696.3		706.3	1		y6		Y	15.1
P05155	IC1	TNLESILSYPK	632.8	636.8	2	216.1	807.5	216.1	1	1	b2	y7	Y	20.6
P05155	IC1	TTFDPK	354.7	358.7	2	244.2	375.6	252.2	1	2	y2	b3	Y	12
P05362	ICAM1	VELAPLPSWQPVGK	760.9	764.9	2	342.2		342.2	1		b3		Y	24.6
P22304	IDS	QSTEQAIQLLEK	694.4	698.4	2	814.5		822.5	1		y7		Y	22.5
P01344	IGF2	GIVEECCFR	585.3	590.3	2	900.3		910.3	1		y6		Y	19.1
P01344	IGF2	SCDLALLETYCATPAK	906.9	910.9	2	315.2		323.2	1		y3		Y	29.1
P01857	IGHG1	TPEVTCVVVDVSHEDPEVK	713.7	716.4	3	472.3		480.3	1		y4		Y	20.9
P01860	IGHG3	SCDTPPPCPR	593.8	598.8	2	723.4		733.4	1		y6		Y	19.4
P01871	IGHM	QIQVSWLR	515.3	520.3	2	788.4		798.4	1		y6		Y	17
P01871	IGHM	QVGSGVTTDQVQAEAK	809.4	813.4	2	1090.5		1098.6	1		y10		Y	26.1
P01871	IGHM	YAATSQVLLPSK	639.4	643.4	2	331.2		339.2	1		y3		Y	20.8
P01834	IGKC	DSTYSLSSTLTLSK	751.9	755.9	2	836.5		844.5	1		y8		Y	24.3
P01834	IGKC	VDNALQSGNSQESVTEQDSK	1068.5	1072.5	2	707.3		715.3	1		y6		Y	34.1
P05113	IL5	ETLALLSTHR	570.8	575.8	2	613.3		623.3	1		y5		Y	18.7
P06213	INSR	VCHLLEGEK	542.8	546.8	2	413.2		417.2	2		y7		Y	17.8
P05154	IPSP	AAAATGTIFTFR	613.8	618.8	2	214.1		214.1	1		b3		Y	20

P05154	IPSP	AVVEVDESGTR	581.3	586.3	2	763.4		773.4	1		y7		Y	19
P05154	IPSP	DFTFDLYR	538.8	543.8	2	814.4		824.4	1		y6		Y	17.7
P05154	IPSP	FSIEGSYQLEK	650.8	654.8	2	953.5		961.5	1		y8		Y	21.2
P05154	IPSP	TLYLADTFPTNFR	779.9	784.9	2	634.3		644.3	1		y5		Y	25.2
P08514	ITA2B	IVLLDVPVR	512.3	517.3	2	811.5		821.5	1		y7		Y	16.9
P19827	ITIH1	AAISGENAGLVR	579.3	584.3	2	902.5		912.5	1		y9		Y	19
P19827	ITIH1	FAHYVVTSQVVNTANEAR	669.3	672.7	3	775.4		785.4	1		y7		Y	19.3
P19827	ITIH1	GSLVQASEANLQAAQDFVR	1002.5	1007.5	2	806.4		816.4	1		y7		Y	32.1
P19827	ITIH1	QAVDTAVDGVFIR	695.9	700.9	2	706.4		716.4	1		y6		Y	22.6
P19827	ITIH1	QLVHHFEIDVDIFEPQGISK	784.4	787.1	3	667.3		667.3	2		b11		Y	23.4
P19827	ITIH1	QYYEGSEIVVAGR	735.9	740.9	2	887.5		897.5	1		y9		Y	23.8
P19823	ITIH2	AEDHFSVIDFNQNIR	902.9	907.9	2	791.4		801.4	1		y6		Y	29
P19823	ITIH2	IYLQPGR	423.7	428.7	2	570.3		580.3	1		y5		Y	14.1
P19823	ITIH2	TEVNVLPGAK	514.3	518.3	2	231.1	698.4	231.1	1	1	b2	y7	Y	16.9
P19823	ITIH2	TQVADAK	366.7	370.7	2	503.3		511.3	1		y5		Y	12.4
P19823	ITIH2	VVNNSPQPQNVVFDVQIPK	1061.6	1065.6	2	244.2	514.3	252.2	1	1	y2	b5	Y	33.9
Q06033	ITIH3	EHLVQATPENLQEAR	867.9	872.9	2	267.1	380.2	267.1	1	1	b2	b3	Y	27.9
Q06033	ITIH3	EVSFDVELPK	581.8	585.8	2	700.4		708.4	1		y6		Y	19
Q06033	ITIH3	SLPEGVANGIEVYSTK	555.3	558	3	726.4		734.4	1		y6		Y	15.2
Q14624	ITIH4	GPDVLTATVSGK	572.8	576.8	2	663.4		671.4	1		y7		Y	18.8
Q14624	ITIH4	ILDDLSPR	464.8	469.8	2	702.3		712.3	1		y6		Y	15.4
Q14624	ITIH4	LALDNGGLAR	500.3	505.3	2	815.4		825.4	1		y8		Y	16.5
Q14624	ITIH4	NPLVWVHASPEHVVVTR	970.5	975.5	2	424.3		424.3	1		b4		Y	31.1
Q14624	ITIH4	NVVFVIDK	467.3	471.3	2	720.4		728.4	1		y6		Y	15.5
O60674	JAK2	SDNIIFQFTK	606.8	610.8	2	670.4		678.4	1		y5		Y	19.8
P35527	K1C9	FSSSGGGGGGGR	491.7	496.7	2	748.3		758.3	1		y10		Y	16.2
P35527	K1C9	FSSSSGYGGGSSR	618.3	623.3	2	520.2		530.3	1		y6		Y	20.2
P35527	K1C9	TLLDIDNTR	530.8	535.8	2	215.1	733.3	215.1	1	1	b2	y6	Y	17.5
P35527	K1C9	VQALEEANNDLENK	793.9	797.9	2	917.4		925.4	1		y8		Y	25.6
P04264	K2C1	SLDLDSIIAEVK	651.9	655.9	2	874.5		882.5	1		y8		Y	21.2
P04264	K2C1	SLVNLGGSK	437.8	441.8	2	674.4		682.4	1		y7		Y	14.6
P04264	K2C1	TLLEGEESR	517.3	522.3	2	819.4		829.4	1		y7		Y	17
P13647	K2C5	ISISTSGGSFR	556.3	561.3	2	798.4		808.4	1		y8		Y	18.2
P29622	KAIN	GDATVFFILPNQGK	753.9	757.9	2	543.3		551.3	1		y5		Y	24.4
P29622	KAIN	GFQHLLHTLNLPGHGLETR	714.1	717.4	3	866.4		876.5	1		y8		Y	20.9
P29622	KAIN	WADLSGITK	495.8	499.8	2	733.4		741.4	1		y7		Y	16.4

P06732	KCRM	ELFDPIISDR	602.8	607.8	2	350.7		355.7	2		y6		Y	19.7
O75037	KI21B	AQEQGVAGPEFK	630.8	634.8	2	294.2	648.3	302.2	1	1	y2	y6	Y	20.6
P03952	KLKB1	FGCFLK	386.2	390.2	2	624.3		632.3	1		y5		Y	13
P03952	KLKB1	IAYGTQGSSGYSLR	730.4	735.4	2	826.4		836.4	1		y8		Y	23.6
P03952	KLKB1	LCNTGDNSVCTTK	735.3	739.3	2	509.2		517.3	1		y4		Y	23.8
P03952	KLKB1	LVGITSWGEGCAR	703.3	708.4	2	1023.4		1033.4	1		y9		Y	22.8
P03952	KLKB1	VNIPLVTNEECQK	772.4	776.4	2	214.1	908.4	214.1	1	1	b2	y7	Y	24.9
P01042	KNG1	AATGECTATVGK	583.3	587.3	2	204.1	736.4	212.1	1	1	y2	y7	Y	19.1
P01042	KNG1	ENFLFLTPDCK	692.3	696.3	2	880.4		888.4	1		y7		Y	22.5
P01042	KNG1	QVVAGLNFR	502.3	507.3	2	677.4		687.4	1		y6		Y	16.6
P01042	KNG1	TVGSDTFYSFK	626.3	630.3	2	173.1		173.1	2		b4		Y	20.4
P01042	KNG1	YNSQNQSNNQFVLYR	625.6	629	3	697.4		707.4	1		y5		Y	17.7
P05455	LA	IGCLLK	352.2	356.2	2	590.3		598.3	1		y5		Y	11.9
P11279	LAMP1	ALQATVGNSYK	576.3	580.3	2	839.4		847.4	1		y8		Y	18.9
P13473	LAMP2	IPLNDLFR	494.3	499.3	2	777.4		787.4	1		y6		Y	16.3
P18428	LBP	GLQYAAQEGLLALQSELLR	691.7	695.1	3	617.4		627.4	1		y5		Y	20.1
P18428	LBP	ITLPDFTGDLR	624.3	629.3	2	215.1	460.3	215.1	1	1	b2	y4	Y	20.4
P18428	LBP	LAEGFPLPLLK	599.4	603.4	2	680.5		688.5	1		y6		Y	19.6
P18428	LBP	VQLYDLGLQIHK	476.3	478.9	3	228.1	695.4	228.1	1	1	b2	y6	Y	12.3
P04180	LCAT	LAGYLHTLVQNLVNNGYVR	715.4	718.7	3	821.4		831.4	1		y7		Y	21
P04180	LCAT	SSGLVSNAPGVQIR	692.9	697.9	2	669.4		679.4	1		y6		Y	22.5
P04180	LCAT	STELCGLWQGR	653.8	658.8	2	876.4		886.4	1		y7		Y	21.3
P04180	LCAT	TYSVEYLDSSK	646.3	650.3	2	1027.5		1035.5	1		y9		Y	21
P07195	LDHB	IVVVTAGVR	457.3	462.3	2	701.4		711.4	1		y7		Y	15.2
P07195	LDHB	SADTLWDIQK	588.8	592.8	2	689.4		697.4	1		y5		Y	19.3
P51884	LUM	FNALQYLR	512.8	517.8	2	763.4		773.5	1		y6		Y	16.9
P51884	LUM	ILGPLSYSK	489.3	493.3	2	751.4		759.4	1		y7		Y	16.2
P51884	LUM	ISNIPDEYFK	613.3	617.3	2	798.4		806.4	1		y6		Y	20
P51884	LUM	LPSGLPVSLLTLYLDNNK	653	655.7	3	766.4		774.4	1		y6		Y	18.7
P51884	LUM	NIPTVNENLENYYLEVNQLEK	1268.6	1272.6	2	631.3		639.4	1		y5		Y	40.3
P51884	LUM	SLEDLQLTHNK	433.2	435.9	3	201.1	612.3	201.1	1	1	b2	y5	Y	10.8
P14151	LYAM1	AEIEYLEK	497.8	501.8	2	794.4		802.4	1		y6		Y	16.4
Q86UE4	LYRIC	WNSVSPASAGK	552.3	556.3	2	803.4		811.4	1		y9		Y	18.1
P61626	LYSC	STDYGIFQINSR	700.8	705.8	2	764.4		774.4	1		y6		Y	22.7
P61626	LYSC	WESGYNTR	506.7	511.7	2	697.3		707.3	1		y6		Y	16.7
P11226	MBL2	FQASVATPR	488.8	493.8	2	701.4		711.4	1		y7		Y	16.2

P11226	MBL2	TEGQFVDLTGNR	668.8	673.8	2	231.1	774.4	231.1	1	1	b2	y7	Y	21.7
P11226	MBL2	WLTFSLGK	476.3	480.3	2	765.5		773.5	1		y7		Y	15.8
O43772	MCAT	EGITGLYR	454.7	459.7	2	609.3		619.3	1		y5		Y	15.1
Q16674	MIA	GQVVYVFSK	513.8	517.8	2	186.1	480.3	186.1	1	1	b2	y4	Y	16.9
P43246	MSH2	DIYQDLNR	518.8	523.8	2	808.4		818.4	1		y6		Y	17.1
P61916	NPC2	SEYPSIK	412.2	416.2	2	607.3		615.4	1		y5		Y	13.8
Q13093	PAFA	ASLAFLQK	439.3	443.3	2	606.4		614.4	1		y5		Y	14.6
P36955	PEDF	ALYYDLISSPDIHGTYK	652.7	655.3	3	185.1	468.2	185.1	1	1	b2	y4	Y	18.7
P36955	PEDF	ELLDTVTAPQK	607.8	611.8	2	859.5		867.5	1		y8		Y	19.8
P36955	PEDF	LQSLFDSPDFSK	692.3	696.4	2	242.1	942.4	242.1	1	1	b2	y8	Y	22.5
P36955	PEDF	VLTGNPR	378.7	383.7	2	544.3		554.3	1		y5		Y	12.7
P36955	PEDF	YGLDSDLSCK	579.3	583.3	2	824.3		832.4	1		y7		Y	19
P05164	PERM	VFFASWR	456.7	461.7	2	666.3		676.3	1		y5		Y	15.2
P05164	PERM	VVLEGGIDPILR	640.9	645.9	2	840.5		850.5	1		y8		Y	20.9
P09619	PGFRB	LPGFHGLR	448.8	453.8	2	392.2		397.2	2		y7		Y	14.9
Q96PD5	PGRP2	DGSPDVTTADIGANTPDATK	973.5	977.5	2	874.4		882.4	1		y9		Y	31.2
Q96PD5	PGRP2	EFTEAFLGCPAIHPR	872.9	877.9	2	907.5		917.5	1		y8		Y	28.1
Q96PD5	PGRP2	TDCPGDALFDLLR	746.9	751.9	2	1116.6		1126.6	1		y10		Y	24.2
Q96PD5	PGRP2	TFTLLDPK	467.8	471.8	2	244.2	686.4	252.2	1	1	y2	y6	Y	15.5
P80108	PHLD	FGSSLITVR	490.3	495.3	2	205.1	775.5	205.1	1	1	b2	y7	Y	16.2
P80108	PHLD	HVSSPLASYFLSFPYAR	648	651.3	3	740.4		750.4	1		y6		Y	18.5
P80108	PHLD	IADVTSGLIGGEDGR	730.4	735.4	2	590.3		600.3	1		y6		Y	23.6
P80108	PHLD	NQVVIAAGR	464.3	469.3	2	586.4		596.4	1		y6		Y	15.4
P80108	PHLD	SWITPCPEEK	623.8	627.8	2	274.1	759.3	274.1	1	1	b2	y6	Y	20.3
P80108	PHLD	TLLLVGSPTWK	607.9	611.9	2	215.1	774.4	215.1	1	1	b2	y7	Y	19.8
Q6UXB8	PI16	WDEELAAFAK	590.3	594.3	2	878.5		886.5	1		y8		Y	19.3
P02776	PLF4	ICLDLQAPLYK	667.4	671.4	2	274.1	832.5	274.1	1	1	b2	y7	Y	21.7
P00747	PLMN	EAQLPVIENK	570.8	574.8	2	699.4		707.4	1		y6		Y	18.7
P00747	PLMN	LSSPAVITDK	515.8	519.8	2	743.4		751.4	1		y7		Y	17
P00747	PLMN	NPDGDVGGPWCYTTNPR	953.4	958.4	2	499.2		499.2	1		b5		Y	30.6
P00747	PLMN	VIPACLPSPNYVVADR	886	891	2	1117.6		1127.6	1		y10		Y	28.5
P00747	PLMN	WEYCNLK	506.7	510.7	2	697.3		705.3	1		y5		Y	16.7
P00747	PLMN	YEFLNGR	449.7	454.7	2	606.3		616.3	1		y5		Y	14.9
P55058	PLTP	ATYFGSIVLLSPAVIDSPLK	1046.1	1050.1	2	1026.6		1034.6	1		y10		Y	33.4
P55058	PLTP	AVEPQLQEEER	664.3	669.3	2	514.8		519.8	2		y8		Y	21.6
P55058	PLTP	FLEQELETITIPDLR	606.3	609.7	3	500.3		510.3	1		y4		Y	17

P55058	PLTP	TGLELSR	388.2	393.2	2	504.3		514.3	1		y4		Y	13
P27169	PON1	IFFYDSENPPASEVLR	942.5	947.5	2	868.5		878.5	1		y8		Y	30.2
P27169	PON1	IHVYEK	394.7	398.7	2	338.2		342.2	2		y5		Y	13.2
P27169	PON1	IQNILTEEPK	592.8	596.8	2	242.1	716.4	242.1	1	1	b2	y6	Y	19.4
P27169	PON1	SFNPNSPGK	474.2	478.2	2	599.3		607.3	1		y6		Y	15.7
P27169	PON1	STVELFK	412.2	416.2	2	635.4		643.4	1		y5		Y	13.8
P27169	PON1	VVAEGFDFANGINISPDGK	975.5	979.5	2	503.2		511.3	1		y5		Y	31.2
P32119	PRDX2	TDEGIAYR	462.7	467.7	2	217.1	579.3	217.1	1	1	b2	y5	Y	15.3
P04070	PROC	DTEDQEDQVDPR	723.8	728.8	2	272.2	858.4	282.2	1	1	y2	y7	Y	23.4
P04070	PROC	GDSPWQVVLLDSK	722.4	726.4	2	592.8		596.8	2		y10		Y	23.4
P04070	PROC	TFVLNFIK	491.3	495.3	2	733.5		741.5	1		y6		Y	16.2
P07737	PROF1	STGGAPTFNVTVTK	690.4	694.4	2	1006.6		1014.6	1		y9		Y	22.4
P07225	PROS	FSAEFDFR	509.7	514.7	2	784.4		794.4	1		y6		Y	16.8
P07225	PROS	HCLVTVEK	493.3	497.3	2	848.5		856.5	1		y7		Y	16.3
P07225	PROS	NNLELSTPLK	564.8	568.8	2	787.5		795.5	1		y7		Y	18.5
P07225	PROS	QSTNAYPDLR	582.8	587.8	2	500.3		510.3	1		y4		Y	19.1
P07225	PROS	SCEVVSVCLPLNLDTK	917.5	921.5	2	800.5		808.5	1		y7		Y	29.4
P07225	PROS	SFQTGLFTAAR	599.8	604.8	2	836.5		846.5	1		y8		Y	19.6
P07225	PROS	VYFAGFPR	478.8	483.8	2	694.4		704.4	1		y6		Y	15.8
P20742	PZP	ASPAFLASQNTK	617.8	621.8	2	648.3		656.3	1		y6		Y	20.2
P20742	PZP	GSFALSFPVESDVAPIAR	932	937	2	1153.6		1163.6	1		y11		Y	29.9
P20742	PZP	HQDGSYSTFGER	461.9	465.2	3	508.3		518.3	1		y4		Y	11.8
P20742	PZP	IQHPFTVEEFVLPK	562	564.6	3	244.2	732.4	252.2	1	1	y2	y6	Y	15.4
P20742	PZP	LPSNVVK	378.7	382.7	2	322.2		326.2	2		y6		Y	12.7
P20742	PZP	NALFCLESAWNVAK	811.9	815.9	2	299.2		299.2	1		b3		Y	26.2
P00797	RENI	LFDASDSSSYK	610.3	614.3	2	959.4		967.4	1		y9		Y	19.9
P02753	RET4	LIVHNGYCDGR	652.3	657.3	2	489.7		494.7	2		y8		Y	21.2
P02753	RET4	LLNLDGTCADSYSFVFSR	689	692.3	3	742.4		752.4	1		y6		Y	20
P02753	RET4	QEELCLAR	509.8	514.8	2	519.3		529.3	1		y4		Y	16.8
P02753	RET4	YWGVASFLQK	599.8	603.8	2	849.5		857.5	1		y8		Y	19.6
P06702	S10A9	LGHPDTLNQGEFK	485.9	488.6	3	621.3		621.3	1		b6		Y	12.7
P0DJI8	SAA1	FFGHGAEDSLADQAANEWGR	726.7	730	3	732.3		742.4	1		y6		Y	21.4
P02743	SAMP	AYSLFSYNTQGR	703.8	708.8	2	825.4		835.4	1		y7		Y	22.8
P02743	SAMP	DNELLVYK	497.3	501.3	2	522.3		530.3	1		y4		Y	16.4
P02743	SAMP	IVLGQEQDSYGGK	697.4	701.4	2	1068.5		1076.5	1		y10		Y	22.6
P02743	SAMP	QGYFVEAQPK	583.8	587.8	2	572.3		580.3	1		y5		Y	19.1

P02743	SAMP	VFVFPR	382.7	387.7	2	518.3		528.3	1		y4		Y	12.9
P02743	SAMP	VGEYSLYIGR	578.8	583.8	2	871.5		881.5	1		y7		Y	18.9
Q9BYB0	SHAN3	FEDHEIEGAHLPALTK	603	605.6	3	529.3		537.3	1		y5		Y	16.9
P04278	SHBG	DIPQPHAEPWAFSLDLGLK	712	714.7	3	892.5		900.5	1		y8		Y	20.8
P04278	SHBG	IALGGLLFPASNLR	721.4	726.4	2	804.4		814.4	1		y7		Y	23.4
P04278	SHBG	LPLVPALDGCLR	662.4	667.4	2	901.5		911.5	1		y8		Y	21.5
P04278	SHBG	TSSSFEVR	456.7	461.7	2	724.4		734.4	1		y6		Y	15.2
P04278	SHBG	VVLSQGSK	409.2	413.2	2	619.3		627.4	1		y6		Y	13.7
Q13103	SPP24	DYYVSTAVCR	617.3	622.3	2	693.3		703.3	1		y6		Y	20.1
P22105	TENX	LQGLIPGAR	462.8	467.8	2	683.4		693.4	1		y7		Y	15.3
P05452	TETN	CFLAFTQTK	558.3	562.3	2	695.4		703.4	1		y6		Y	18.3
P05452	TETN	EQQALQTVCLK	659.3	663.4	2	748.4		756.4	1		y6		Y	21.4
P05452	TETN	LDTLAQEVALLK	657.4	661.4	2	871.5		879.5	1		y8		Y	21.4
P05452	TETN	TENCAVLSGAANGK	696.3	700.3	2	717.4		725.4	1		y8		Y	22.6
P05452	TETN	TFHEASEDCISR	484.5	487.9	3	375.2		385.2	1		y3		Y	12.6
P05543	THBG	AQWANPFDPSK	630.8	634.8	2	200.1	446.2	200.1	1	1	b2	y4	Y	20.6
P05543	THBG	FLNDVK	368.2	372.2	2	475.3		483.3	1		y4		Y	12.4
P05543	THBG	GWVDLFVPK	530.8	534.8	2	244.1	718.4	244.1	1	1	b2	y6	Y	17.5
P05543	THBG	NALALFVLPK	543.3	547.3	2	787.5		795.5	1		y7		Y	17.8
P05543	THBG	SILFLGK	389.2	393.3	2	577.4		585.4	1		y5		Y	13.1
P00734	THRB	ELLESYIDGR	597.8	602.8	2	710.3		720.4	1		y6		Y	19.5
P00734	THRB	HQDFNSAVQLVENFCR	655.3	658.6	3	824.4		834.4	1		y6		Y	18.8
P00734	THRB	LAVTTHGLPCLAWASAQAK	665.7	668.4	3	832.4		840.4	1		y8		Y	19.2
P00734	THRB	NPDSSTTGPWCYTTDPTVR	1078	1083	2	472.3		482.3	1		y4		Y	34.4
P00734	THRB	VTGWGNLK	437.7	441.7	2	674.4		682.4	1		y6		Y	14.6
P00734	THRB	YTACETAR	486.2	491.2	2	707.3		717.3	1		y6		Y	16.1
P25942	TNR5	YCDPNLGLR	554.3	559.3	2	669.4		679.4	1		y6		Y	18.2
P02787	TRFE	ASYLDCIR	499.2	504.2	2	563.3		573.3	1		y4		Y	16.5
P02787	TRFE	CSTSSLLEACTFR	766.3	771.4	2	783.3		793.4	1		y6		Y	24.8
P02787	TRFE	FDEFFSEGCAPGSK	526.6	529.2	3	676.3		684.3	1		y7		Y	14.2
P02787	TRFE	HQTVPQNTGGK	389.5	392.2	3	351.2		355.2	2		y7		Y	9.2
P02787	TRFE	IECVSAETTEDCIAK	863.4	867.4	2	742.3		746.3	2		y13		Y	27.8
P02787	TRFE	SAGWNIPIGLLYCDLPEPR	724.4	727.7	3	498.3		508.3	1		y4		Y	21.3
P02787	TRFE	SASDLTWDNLK	417.2	419.9	3	675.3		683.4	1		y5		Y	10.2
P02788	TRFL	CSTSPLLEACEFLR	561.6	564.9	3	795.4		805.4	1		y6		Y	15.4
P02766	TTHY	AADDTWEPFASGK	697.8	701.8	2	606.3		614.3	1		y6		Y	22.6

P02766	TTHY	GSPAINVAVHVFR	456.3	459.6	3	558.3		568.3	1		y4		Y	11.6
P02766	TTHY	VLDAVR	336.7	341.7	2	460.3		470.3	1		y4		Y	11.4
P07911	UROM	SGSVIDQSR	474.7	479.7	2	618.3		628.3	1		y5		Y	15.7
P07911	UROM	VLNLGPITR	491.8	496.8	2	770.5		780.5	1		y7		Y	16.2
Q86UX7	URP2	VVLAGGVAPALFR	635.4	640.4	2	887.5		897.5	1		y9		Y	20.7
Q6EMK4	VASN	LAGLGLQQLDEGLFSR	573	576.3	3	823.4		833.4	1		y7		Y	15.8
Q6EMK4	VASN	SLTLGIEPVSPTSLR	785.4	790.5	2	856.5		866.5	1		y8		Y	25.3
Q6EMK4	VASN	YLQGSSVQLR	575.8	580.8	2	746.4		756.4	1		y7		Y	18.9
P35916	VGFR3	SGVDLADSNQK	567.3	571.3	2	662.3		670.3	1		y6		Y	18.6
P18206	VINC	QVATALQNLQTK	657.9	661.9	2	844.5		852.5	1		y7		Y	21.4
P02774	VTDB	FEDCCQEK	558.2	562.2	2	839.3		847.3	1		y6		Y	18.3
P02774	VTDB	HLSLLTTLSNR	418.9	422.2	3	590.3		600.3	1		y5		Y	10.3
P02774	VTDB	THLPEVFLSK	585.8	589.8	2	239.1	819.5	239.1	1	1	b2	y7	Y	19.2
P02774	VTDB	VCSQYAAYGEK	638.3	642.3	2	1016.5		1024.5	1		y9		Y	20.8
P02774	VTDB	YTFELSR	458.2	463.2	2	651.3		661.4	1		y5		Y	15.2
P04004	VTNC	CTEGFNVDK	535.2	539.2	2	808.4		816.4	1		y7		Y	17.6
P04004	VTNC	DVWGIEGPIDAAFTR	823.9	828.9	2	458.2		458.2	1		b4		Y	26.5
P04004	VTNC	FEDGVLDPDYPR	711.8	716.8	2	647.3		657.3	1		y5		Y	23.1
P04004	VTNC	GSQYWR	398.7	403.7	2	524.3		534.3	1		y3		Y	13.4
P04004	VTNC	VDTVDPPYPR	579.8	584.8	2	629.3		639.3	1		y5		Y	19
P04275	VWF	CHPLVDPEPFVALCEK	637.6	640.3	3	722.3		722.3	1		b6		Y	18.2
P04275	VWF	VTVFPIGIGDR	587.3	592.3	2	727.4		737.4	1		y7		Y	19.2
P25311	ZA2G	AGEVQEPELR	564.3	569.3	2	771.4		781.4	1		y6		Y	18.5
P25311	ZA2G	AYLEEECPATLR	726.3	731.3	2	235.1	846.4	235.1	1	1	b2	y7	Y	23.5
P25311	ZA2G	CLAYDFYPGK	617.3	621.3	2	301.2		309.2	1		y3		Y	20.1
P25311	ZA2G	QDPPSVVVTSHQAPGEK	592.6	595.3	3	766.9		770.9	2		y15		Y	16.5
P25311	ZA2G	SSGAFWK	391.7	395.7	2	608.3		616.3	1		y5		Y	13.1
P25311	ZA2G	YSLTYIYTGLSK	704.9	708.9	2	781.4		789.5	1		y7		Y	22.9

[a] UniProt accession number, protein ID, and information pertaining to mass spectra are listed for each peptide.

[b] For quantifiers (product ions) that are less than 300 m/z, additional quantifiers (>300 m/z) and their corresponding ion charge and type are listed.

[c] All product ions are specified as quantifiers. Y, yes.

4.2. Plasma sample preparation

Plasma samples were thawed on ice and centrifuged at 10,000 g for 10 min at 4°C. Supernatants were transferred to fresh tubes and vortexed. For each sample, a volume of 44 µl was diluted 1:4 with MARS buffer A (Agilent Technologies, Santa Clara, CA, USA) and passed through 0.22 µm Spin-X filters (Corning Costar, NY, USA). A volume of 176 µl of buffer A was added to each sample, and each diluted sample was centrifuged through a 0.22 µm filter (12,000 g, room temperature). Each plasma sample was depleted of 6 high-abundance human plasma proteins—albumin, IgG, IgA, transferrin, haptoglobin, and antitrypsin—using a multiple affinity removal system (MARS) column (Hu-6HC, 4.6 × 100 mm, Agilent Technologies, Santa Clara, CA, USA), loaded onto a high-performance liquid chromatography (HPLC) system (Shimadzu Co, Kyoto, Japan). A total of 200 µl was injected for each sample. Depleted plasma samples were concentrated by centrifugal filtration for 6 hours at 4°C using a 3000-Da molecular weight cutoff (MWCO) filter (Amicon Ultra-4 3K, Millipore, Burlington, MA, USA). Concentrated proteins of plasma samples were quantified by bicinchoninic acid assay (BCA assay) using the Pierce™ BCA Protein Assay Kit (Thermo Scientific, Rockford, IL, USA). BCA assay was performed to measure the concentration of individual samples. A 6-point standard curve was generated by serially diluting an initial concentration of 2 mg/mL BSA by a factor of 2. Standards and samples were placed on a 96-well plate, and a mixture of copper solution and BCA solution (1:50) was added. The proteins were digested using RapiGest surfactant and trypsin. A total of 40 µl solution of 0.2% RapiGest, 20 mM dithiothreitol (DTT), and 100 mM ABC buffer, pH 8.0 was added to the 40-µl plasma samples, adjusted with HPLC-grade water for a 100-µg digestion. After 1 hour in 60°C, 20 µl 100 mM iodoacetamide (IAA) was added. The samples were incubated in the dark for 30 min at room temperature. Following the incubation, the samples were incubated for 4 hours at 37°C after adding trypsin (Sequencing-grade modified, Promega, Madison, WI, USA), dissolved in 50 mM ABC, pH 8.0. Then, 10% formic acid was added to the samples to stop the enzymatic reaction (final concentration of formic acid: 1%), followed by incubation for 30 min at 37° to hydrolyze RapiGest surfactant in the acidified condition of the samples. After centrifugation at

15,000 rpm at 4°C for 1 hour, the precipitation of cleaved RapiGest surfactant was observed. Then, the supernatant without the precipitation was transferred to a new clean tube. The plasma peptide samples (the transferred supernatant for each sample) were spiked with crude stable isotope-labeled internal standard (SIS) peptide, with a C-terminal lysine or arginine heavy-isotope-labeled (13C615N2 or 13C615N4) [purity: crude (>70%), JPT, Berlin, Germany]. However, in this study, SIS peptide should have been added as early as possible after the enzymatic reaction in order to correct for any variability in the MS runs as well as any variation associated with sample preparation after the digestion step. The samples were randomly distributed in blocked batches and labeled with identification numbers to blind the researchers throughout the sample preparation.

4.3. LC-MRM-MS analysis

In this study, 671 peptides, representing 210 proteins (Supplementary Table 1), were analyzed by targeted multiple reaction monitoring-mass spectrometry (MRM-MS) analysis. MRM-MS analysis was performed on an Agilent 6490 triple quadrupole (QQQ) mass spectrometer (Agilent Technologies, Santa Clara, CA, USA), equipped with a Jetstream electrospray source that was coupled to a 1260 Infinity HPLC system (Agilent Technologies, Santa Clara, CA, USA). Solvents A and B for the HPLC consisted of 0.1% formic acid/water (v/v) and 0.1% formic acid/acetonitrile (v/v), respectively. Glass vials of the samples in the autosampler were maintained at 4°C. A total of 40 μl of digested sample was injected into a guard column (2.1 × 15.0 mm, 1.8 μm, 80 Å) (Agilent Technologies, Santa Clara, CA, USA). Online desalting was conducted with the effluent toward waste at 50 μl/min for 10 min in 3% solvent B, consisting of 0.1% formic acid/acetonitrile (v/v), at 40°C. After the position of valve was switched, the desalted sample was transferred from the guard column to the analytical column (0.5 × 35.0 mm, 3.5 μm, 80 Å) (Agilent Technologies, Santa Clara, CA, USA) in 3% solvent B, at a flow rate of 40 μL/min for 5 min. The analytical column was heated and maintained at 40°C by an oven.

The total run time per LC-MRM-MS analysis was 70 min. Approximately 10 μg of digested peptides was injected per LC-MRM-MS run. The peptides were separated on the column and eluted with a linear gradient of 3% to 35% acetonitrile (ACN) with 0.1% formic acid (FA) for 50 min at 40 μL/min. The mass spectra were generated in positive ion mode, based on the following parameters; 2500 V for the ion spray capillary voltage, 2000 V for the nozzle voltage, 5 V for the cell accelerator voltage, 200 V for the delta EMV, and 380 V for the fragmented voltage. The drying gas was sprayed at 15 L/min at 250°C, and the sheath gas flow was 12 L/min at 350°C. Collision energy (CE) was optimized by adding the intensities of individual transitions that resulted in the largest peak area. The default value of CE was calculated as follows: CE = 0.031 × (m/z of precursor) + 1 for double-charged precursor ions and CE = 0.036 × (m/z of precursor) – 4.8 for triple-charged ions. Five additional steps of adding or subtracting 2 V on each side of the default value of CE were predicted.

Five to 10 transition pairs (Q1 and Q3) were selected for each peptides. Subsequently, 1 representative transition for each peptide was determined by rank of intensity of transition and AuDIT [138]. All 671 SIS peptides, representing 210 proteins, were pooled and analyzed to check their retention time (RT). The RTs of the SIS peptides were compared with those of endogenous target peptides by spiking 100 fmol of the pooled mixture of SIS peptides. Subsequently, the final targets—210 proteins/671 peptides/671 transitions—were quantified in 270 individual blood samples, comprising 90 MDD, 90 BD, and 90 HCs. The 270 individual samples were randomly listed in blocked batches with an identification number for each sample. Subsequently, the LC-MRM-MS analysis was performed once per sample (1 replicate for each sample). A total of 270 LC-MRM-MS runs were performed.

4.4. Analysis of demographics and clinical variables of study subjects

Demographics and clinical differences between patient groups and HCs were analyzed by an unpaired Student's *T*-test for continuous variables and Fisher's exact test for dichotomous variables using SPSS, version 25.0 (IBM, Armonk, NY, USA). Statistical tests performed were two-tailed and *P*-value < 0.05 was considered statistically significant.

5. LC-MRM-MS data processing

The raw LC-MRM-MS data were processed using Skyline, version 19.1.0 (MacCoss Lab, Seattle, WA, USA). The quantification of each target was based on the peak area value of endogenous (light) and SIS (heavy) peptide transitions. Peak integrations were performed for a single representative transition for each respective peptide. Peak area ratio (PAR)—the light/heavy (L/H) ratio—was calculated for each target. This process, known as ratio normalization, was used to address any technical variation that occurred across MS runs. The ComBat algorithm—a nonparametric adjustment for reducing batch effects using an empirical Bayes framework—was used to counteract potential batch effects due to technical variabilities, such as sample collection, between institutions. PAR values were inserted into a web-based public server (https://genepattern.broadinstitute.org).

Then, all PAR values were log2-transformed to develop the model. There were no peptides with missing values. All LC-MRM-MS data for the 90 MDD and 90 BD patients were designated as the combined set. Subsequently, through the R package caret, the LC-MRM-MS data of the combined set were divided randomly at a ratio of 8:2 to determine the training set for development of the model and a test set for evaluation of its performance.

6. Selection of candidate features for development of the model

Regarding the LC-MRM-MS data for MDD and BD patients in the training set, area under the receiver operating characteristics (AUROC) analysis (for two groups) was performed to determine 1 representative peptide per protein, with the highest AUROC value, among 671 quantifiable peptides, using SPSS, version 25.0 (IBM, Armonk, NY, USA). In this analysis, the

MDD and BD groups were labeled “0” and “1” to indicate binary variables, respectively. Peptides with AUROC values greater than 0.5 could discriminate more elaborately BD group compared with MDD. Conversely, peptides with AUROC values below 0.5 could distinguish MDD group more precisely versus BD. For example, an AUROC value of 0.7 indicates that the BD group can be classified as the BD group itself, with a probability of 70%. In contrast, an AUROC value of 0.3 signifies that the MDD group can be categorized as MDD itself, with a probability of 70%, equivalent to an AUROC value of 0.7 (1-0.3) when the BD and MDD groups are designated 0 and 1, respectively. Thus, if the AUROC values of the peptides per protein are 0.3, 0.4, 0.5, and 0.6, respectively, the peptide with an AUROC value of 0.3 will be selected as the representative peptide per protein. The selected candidate features (210 proteins/210 peptides) by the AUROC analysis are presented in Table 2.

Table 2. The 210 proteins (210 peptides) selected as candidate features in the training set[a]

Uniprot accession number	Protein	Peptide	AUROC value[b] (MDD vs BD)
P02763	A1AG1	SDVVYTDWK	0.426
P19652	A1AG2	EQLGEFYEALDCLCIPR	0.467
P04217	A1BG	CEGPIPDVTFELLR	0.564
P08697	A2AP	QEDDLANINQWVK	0.557
P01023	A2MG	VYDYYETDEFAIAEYNAPCSK	0.55
Q15848	ADIPO	GDIGETGVPGAEGPR	0.521
P43652	AFAM	HFQNLGK	0.402
P02768	ALBU	LVNEVTEFAK	0.509
P04075	ALDOA	ALQASALK	0.386
P35858	ALS	LHSLHLEGSCLGR	0.461
P02760	AMBP	CVLFPYGGCQGNGNK	0.6
P15144	AMPN	AQIINDAFNLASAHK	0.39
P54802	ANAG	DFCGCHVAWSGSQLR	0.49
P01019	ANGT	LQAILGVPWK	0.425
P01008	ANT3	VWELSK	0.442
P08519	APOA	NPDAVAAPYCYTR	0.577
P02647	APOA1	QGLLPVLESFK	0.416
P02652	APOA2	SPELQAEAK	0.406
P06727	APOA4	GNTEGLQK	0.564
P04114	APOB	ITLPDFR	0.441
P02655	APOC2	ESLSSYWESAK	0.444
P02656	APOC3	DYWSTVK	0.405
P05090	APOD	VLNQELR	0.399
P02649	APOE	EQVAEVR	0.453
Q13790	APOF	SLPTEDCENEK	0.473
P02749	APOH	ATVVYQGER	0.441
O14791	APOL1	LNILNNNYK	0.555
O95445	APOM	FLLYNR	0.39
P00966	ASSY	IDIVENR	0.487
Q76LX8	ATS13	LFINVAPHAR	0.542
P61769	B2MG	VNHVTLSQPK	0.573
P02730	B3AT	LSVPDGFK	0.432
Q8TDL5	BPIB1	ALGFEAAESSLTK	0.496

P43251	BTD	VDLITFDTPFAGR	0.476
Q06187	BTK	LVQLYGVCTK	0.417
P02745	C1QA	SLGFCDTTNK	0.568
P02746	C1QB	LEQGENVFLQATDK	0.629
P02747	C1QC	QTHQPPAPNSLIR	0.619
P00736	C1R	NIGEFCGK	0.527
Q9NZP8	C1RL	GSEAINAPGDNPAK	0.564
P09871	C1S	CEYQIR	0.604
P04003	C4BPA	LSCSYSHWSAPAPQCK	0.596
P54289	CA2D1	VLLDAGFTNELVQNYWSK	0.514
P00915	CAH1	GGPFSDSYR	0.486
P00918	CAH2	YGDFGK	0.454
P27797	CALR	FVLSSGK	0.411
P08185	CBG	HLVALSPK	0.498
P22681	CBL	GTEPIVVDPFDPR	0.461
Q96IY4	CBPB2	YPLYVLK	0.433
P15169	CBPN	VQNECPGITR	0.538
P30279	CCND2	ACQEQIEAVLLNSLQQYR	0.445
P08571	CD14	VLDLSCNR	0.574
O43866	CD5L	CYGPGVGR	0.619
P06731	CEAM5	TLTLFNVTR	0.481
P00450	CERU	EYTDASFTNR	0.393
P00751	CFAB	VSEADSSNADWVTK	0.454
P08603	CFAH	CVEISCK	0.576
P05156	CFAI	EANVACLDLGFQQGADTQR	0.53
P06276	CHLE	AEEILSR	0.525
P10909	CLUS	EIQNAVNGVK	0.521
Q96KN2	CNDP1	AIHLDLEEYR	0.46
P02452	CO1A1	VLCDDVICDETK	0.518
P06681	CO2	HAFILQDTK	0.458
P01024	CO3	DSCVGSLVVK	0.556
P0C0L4	CO4A	ANSFLGEK	0.468
P08572	CO4A2	IAVQPGTVGPQGR	0.55
P01031	CO5	IDTALIK	0.427
P13671	CO6	GFVVAGPSR	0.57
P10643	CO7	VLFYVDSEK	0.545
P07357	CO8A	AIDEDCSQYEPIPGSQK	0.474

P07358	CO8B	CEGFVCAQTGR	0.572
P07360	CO8G	AGQLSVK	0.503
P02748	CO9	TSNFNAAISLK	0.586
Q03692	COAA1	GTHVWVGLYK	0.41
Q9UMD9	COHA1	QAAYNADSGLK	0.482
P49747	COMP	DTDLDGFPDEK	0.524
P20815	CP3A5	DTINFLSK	0.516
P22792	CPN2	QLVCPVTR	0.536
P02741	CRP	ESDTSYVSLK	0.491
P02775	CXCL7	NIQSLEVIGK	0.359
P01034	CYTC	ALDFAVGEYNK	0.474
O95822	DCMC	LCAWYLYGEK	0.455
P09172	DOPO	VISTLEEPTPQCPTSQGR	0.577
Q14126	DSG2	ILDVNDNIPVVENK	0.448
P32926	DSG3	LAEISLGVDGEGK	0.328
Q16610	ECM1	FCEAEFSVK	0.605
P00533	EGFR	CNLLEGEPR	0.52
Q01780	EXOSX	SGPLPSAER	0.572
P00488	F13A	STVLTIPEIIIK	0.427
P05160	F13B	VLHGDLIDFVCK	0.537
P00742	FA10	QEDACQGDSGGPHVTR	0.593
P03951	FA11	SCALSNLACIR	0.572
P00748	FA12	CFEPQLLR	0.547
P12259	FA5	GEYEEHLGILGPIIR	0.574
P08709	FA7	LHQPVVLTDHVVPLCLPER	0.515
P00740	FA9	NCELDVTCNIK	0.563
P23142	FBLN1	CVDVDECAPPAEPCGK	0.646
Q12805	FBLN3	NPCQDPYILTPENR	0.553
P35556	FBN2	FNLSHLGSK	0.516
P22087	FBRL	NGGHFVISIK	0.575
P08637	FCG3A	AVVFLEPQWYR	0.478
Q9Y6R7	FCGBP	AIGYATAADCGR	0.602
O75636	FCN3	YGIDWASGR	0.474
P02765	FETUA	CNLLAEK	0.548
Q9UGM5	FETUB	SQASSCSLQSSDSVPVGLCK	0.443
Q03591	FHR1	TGESAEFVCK	0.482
Q02985	FHR3	AQTTVTCTEK	0.595

Q9BXR6	FHR5	TGDAVEFQCK	0.528
P02671	FIBA	VQHIQLLQK	0.46
P02675	FIBB	QGFGNVATNTDGK	0.527
P02679	FIBG	ASTPNGYDNGIIWATWK	0.465
P02751	FINC	VPGTSTSATLTGLTR	0.458
Q06787	FMR1	EPCCWWLAK	0.439
O95954	FTCD	SDLQVAAK	0.498
P06396	GELS	QTQVSVLPEGGETPLFK	0.549
Q92820	GGH	YLESAGAR	0.487
P22352	GPX3	NSCPPTSELLGTSDR	0.352
Q14520	HABP2	FCEIGSDDCYVGDGYSYR	0.612
P69905	HBA	VGAHAGEYGAEALER	0.509
P08397	HEM3	ELEHALEK	0.395
P02790	HEMO	SGAQATWTELPWPHEK	0.441
P05546	HEP2	TLEAQLTPR	0.419
Q04756	HGFA	LEACESLTR	0.602
P00738	HPT	VGYVSGWGR	0.492
P00739	HPTR	VVLHPNYHQVDIGLIK	0.518
P04196	HRG	QIGSVYR	0.546
P18065	IBP2	LIQGAPTIR	0.415
P17936	IBP3	YGQPLPGYTTK	0.543
P24593	IBP5	GVCLNEK	0.529
P05155	IC1	LLDSLPSDTR	0.556
P05362	ICAM1	VELAPLPSWQPVGK	0.56
P22304	IDS	QSTEQAIQLLEK	0.467
P01344	IGF2	GIVEECCFR	0.586
P01857	IGHG1	TPEVTCVVVDVSHEDPEVK	0.529
P01860	IGHG3	SCDTPPPCPR	0.538
P01871	IGHM	QVGSGVTTDQVQAEAK	0.598
P01834	IGKC	DSTYSLSSTLTLSK	0.563
P05113	IL5	ETLALLSTHR	0.443
P06213	INSR	VCHLLEGEK	0.539
P05154	IPSP	AAAATGTIFTFR	0.596
P08514	ITA2B	IVLLDVPVR	0.383
P19827	ITIH1	GSLVQASEANLQAAQDFVR	0.404
P19823	ITIH2	IYLQPGR	0.403
Q06033	ITIH3	EVSFDVELPK	0.536

Q14624	ITIH4	NVVFVIDK	0.386
O60674	JAK2	SDNIIFQFTK	0.535
P35527	K1C9	TLLDIDNTR	0.453
P04264	K2C1	SLVNLGGSK	0.404
P13647	K2C5	ISISTSGGSFR	0.469
P29622	KAIN	GFQHLLHTLNLPGHGLETR	0.58
P06732	KCRM	ELFDPIISDR	0.455
O75037	KI21B	AQEQGVAGPEFK	0.442
P03952	KLKB1	LVGITSWGEGCAR	0.442
P01042	KNG1	QVVAGLNFR	0.55
P05455	LA	IGCLLK	0.402
P11279	LAMP1	ALQATVGNSYK	0.435
P13473	LAMP2	IPLNDLFR	0.526
P18428	LBP	GLQYAAQEGLLALQSELLR	0.554
P04180	LCAT	SSGLVSNAPGVQIR	0.607
P07195	LDHB	SADTLWDIQK	0.366
P51884	LUM	LPSGLPVSLLTLYLDNNK	0.464
P14151	LYAM1	AEIEYLEK	0.53
Q86UE4	LYRIC	WNSVSPASAGK	0.488
P61626	LYSC	STDYGIFQINSR	0.479
P11226	MBL2	WLTFSLGK	0.545
O43772	MCAT	EGITGLYR	0.511
Q16674	MIA	GQVVYVFSK	0.464
P43246	MSH2	DIYQDLNR	0.478
P61916	NPC2	SEYPSIK	0.447
Q13093	PAFA	ASLAFLQK	0.426
P36955	PEDF	ALYYDLISSPDIHGTYK	0.538
P05164	PERM	VVLEGGIDPILR	0.421
P09619	PGFRB	LPGFHGLR	0.557
Q96PD5	PGRP2	EFTEAFLGCPAIHPR	0.527
P80108	PHLD	TLLLVGSPTWK	0.468
Q6UXB8	PI16	WDEELAAFAK	0.562
P02776	PLF4	ICLDLQAPLYK	0.351
P00747	PLMN	YEFLNGR	0.435
P55058	PLTP	AVEPQLQEEER	0.431
P27169	PON1	IHVYEK	0.518
P32119	PRDX2	TDEGIAYR	0.475

P04070	PROC	GDSPWQVVLLDSK	0.39
P07737	PROF1	STGGAPTFNVTVTK	0.403
P07225	PROS	VYFAGFPR	0.459
P20742	PZP	HQDGSYSTFGER	0.507
P00797	RENI	LFDASDSSSYK	0.515
P02753	RET4	QEELCLAR	0.542
P06702	S10A9	LGHPDTLNQGEFK	0.434
P0DJI8	SAA1	FFGHGAEDSLADQAANEWGR	0.558
P02743	SAMP	IVLGQEQDSYGGK	0.445
Q9BYB0	SHAN3	FEDHEIEGAHLPALTK	0.517
P04278	SHBG	TSSSFEVR	0.468
Q13103	SPP24	DYYVSTAVCR	0.373
P22105	TENX	LQGLIPGAR	0.558
P05452	TETN	CFLAFTQTK	0.612
P05543	THBG	NALALFVLPK	0.53
P00734	THRB	VTGWGNLK	0.443
P25942	TNR5	YCDPNLGLR	0.469
P02787	TRFE	CSTSSLLEACTFR	0.62
P02788	TRFL	CSTSPLLEACEFLR	0.498
P02766	TTHY	AADDTWEPFASGK	0.391
P07911	UROM	VLNLGPITR	0.394
Q86UX7	URP2	VVLAGGVAPALFR	0.36
Q6EMK4	VASN	LAGLGLQQLDEGLFSR	0.556
P35916	VGFR3	SGVDLADSNQK	0.476
P18206	VINC	QVATALQNLQTK	0.364
P02774	VTDB	YTFELSR	0.448
P04004	VTNC	CTEGFNVDK	0.616
P04275	VWF	VTVFPIGIGDR	0.447
P25311	ZA2G	CLAYDFYPGK	0.561

[a] Uniprot accession number, protein ID, and AUROC value for each corresponding peptide. Abbreviations: AUROC, area under the receiver operating characteristics; MDD, major depressive disorder; BD, bipolar disorder.

[b] Because MDD and BD were labeled 0 and 1, proteins and their corresponding peptides in the range of AUROC values (0.5-1) indicate that they can more elaborately classify BD as BD itself compared with MDD (light yellow). Proteins and their corresponding peptides in

the range of AUROC values (0-0.5) signify that they can more precisely classify MDD as MDD itself compared with BD (light blue). The AUROC values (0-0.5) are the same as those (0.5-1) when BD and MDD are designated 0 and 1, respectively.

7. Development of model for discriminating MDD from BD

A total of 210 candidate features were used to develop the model. Least absolute shrinkage and selection operator (LASSO) was used to decrease overfitting by simultaneous shrinkage of the coefficients and model selection. Features with poor discriminatory power were excluded from the model because their coefficients converged to 0. Conversely, features with a non-zero coefficient were selected for the model. LASSO regression was performed using the R package glmnet [139]. Ten-fold crossvalidation was used to determine the optimized value of the shrinkage parameter, lambda, which resulted in the most standardized model.

LASSO with ten-fold crossvalidation (100 repetitions) was applied to the training set. Using this method, uncertainty of the model selection was assessed by examining the fluctuation in model selection regarding small changes in the data that originated from the random division in the ten-fold crossvalidation. For each feature, the proportion of models of the 100 from which it was selected was calculated and defined as the proportion of feature selected, assigned a value from 0 to 1. The proportion of feature selected was used to examine the relative significance of the features. Unique models that were based on the combination of selected features were generated. In addition, the frequency and probability of each unique model were measured. Feature extraction and model averaging across all 100 models were performed to obtain a generalizable and reproducible model for discriminating between MDD and BD. Only features with a proportion of feature selected ≥ 0.9 were combined in the final model—ie, feature extraction. These features constituted the combined features (proteins) in the following analysis. Subsequently, the coefficients of the extracted features were averaged across 100 models—ie, model averaging, based on Akaike’s information criterion (AIC), the bias-corrected version of

AIC (AICc), and Akaike weight (w). These methods thoroughly referred to those of preceding studies [100, 105, 112].

8. Discriminatory performance of the model

The performance of the model was evaluated by examining its AUROC value (R package ROCR) [140]. For this model, the AUROC value is the probability that a randomly chosen individual patient with major depressive disorder (MDD) is ranked higher or lower than one with bipolar disorder (BD). Because the MDD and BD groups were designated 0 and 1, respectively, AUROC values larger than 0.5 reflect probabilities that the BD group can be classified as BD itself compared with the MDD group. Conversely, AUROC values less than 0.5 are probabilities that the MDD group can be categorized as MDD itself versus the BD group. In particular, these probabilities are the same as those for AUROC values above 0.5 when the MDD and BD groups are designated as 0 and 1. In the case of the model's discriminatory performance between the patient groups and HCs, this is the probability that a randomly selected individual in each patient group is ranked higher than HC. AUROC values for the model performance were classified as follows: 0.5–0.6 = fail; 0.6–0.7 = poor; 0.7–0.8 = fair; 0.8–0.9 = good; 0.9–1 = excellent [141]. The optimal cutoff point of the model for discriminating 90 MDD and 90 BD patients was determined per the Youden Index, as follows: J = max (Sensitivity + Specificity – 1) [142]. The sensitivity, specificity, and accuracy for distinguishing 90 MDD from 90 BD patients were calculated at the optimal cutoff point.

9. Differences in abundance of the combined features (proteins) in the model

Differences in protein abundance were examined using SPSS, version 25.0 (IBM, Armonk, NY, USA). An unpaired Student's *T*-test (for two groups) was performed between MDD and BD in the training set. One-way analysis of variance (ANOVA) (for three groups) was performed for the study population, consisting of 90 MDD, 90 BD, and 90 HCs. Subsequently, post hoc analysis for each specific comparison between the groups was performed using Tukey's honestly significant difference (HSD). *P*-value < 0.05 was considered to be statistically significant in these statistical tests, which were two-tailed.

10. Covariate analysis of the combined features (proteins) in the model

Correlations between demographic/clinical variables and features in the model were examined by Pearson's correlation between continuous variables and by point-biserial correlation between dichotomous variables. Subsequently, for features and demographic/clinical variables that had a mutually and statistically significant correlation, influences of covariates on the features were examined by a blend of analysis of variance and regression (ANCOVA) using SPSS, version 25.0 (IBM, Armonk, NY, USA). *P*-value of less than 0.05 was regarded as statistical significance.

11. Bioinformatics analysis of the combined features (proteins) in the model

The top protein network with the highest score was examined by Ingenuity Pathway Analysis (IPA, QIAGEN, Hilden, Germany), based on the features (proteins) that were included in the developed model, with matched gene names [143]. Subsequently, diseases and functions categories and canonical pathways that were associated with the top network were examined. The analytical algorithms in IPA use lists of proteins to predict protein networks, diseases/functions, and canonical pathways. Statistical significance in this analysis was determined by Fisher's exact test and *P*-value of less than 0.05 was thought of as statistically significant.

RESULTS

1. Demographics and clinical characteristics

In total, 90 patients with MDD, 90 patients with BD, and 90 HCs were included in this study. There was no significant difference between patient's groups regarding demographics. However, compared with the BD group, MDD patients had higher MADRS and HAM-A scores and lower YMRS scores. The BD group had a higher proportion of AP and MS use, and MDD patients had a higher proportion of AD use. Whereas there was no significant difference in duration from first onset between MDD and BD patients, duration from first medication differed significantly between these groups. The statistical significance was identical after the groups were divided into training and test sets, with the exception of the loss of statistical significance in the test set for HAM-A and BZD/HNT use and duration from first medication. Comparing patient groups with HCs, there were significant differences in current smoking status, current exercise status, blood collection time, and fasting time. In addition, all clinical assessments and medication use differed significantly. A summary of these characteristics is presented in Table 3.

Table 3. Demographics and clinical characteristics of the study subjects[a]

Characteristics		**N (%) or value**			***P*-value[b]**			**N (%) or value**				***P*-value[b]**	
								Training set		**Test set**		**Training set**	**Test set**
		MDD (n=90)	**BD (n=90)**	**HC (n=90)**	**MDD vs BD**	**MDD vs HC**	**BD vs HC**	**MDD (n=72)**	**BD (n=72)**	**MDD (n=18)**	**BD (n=18)**	**MDD vs BD**	**MDD vs BD**
Gender	Male N (%)	33 (36.7%)	24 (26.7%)	33 (36.7%)	0.20	1.00	0.20	26 (36.1%)	22 (30.6%)	7 (38.9%)	7 (38.9%)	0.60	1.00
Age	Mean (s.d)	37.6 (12.9)	34.5 (12.2)	35.1 (10.6)	0.10	0.15	0.76	37.4 (12.5)	34.7 (11.6)	38.5 (16.0)	34.1 (15.5)	0.19	0.40
BMI	Mean (s.d)	23.8 (4.5)	24.3 (4.1)	23.7 (2.5)	0.43	0.88	0.25	24.2 (4.5)	24.5 (4.0)	26.3 (5.9)	25.8 (4.0)	0.69	0.80
Current smoking status	Yes N (%)	33 (36.7%)	29 (32.2%)	5 (5.6%)	0.64	**< 0.001**	**< 0.001**	26 (36.1%)	23(31.9%)	8 (44.4%)	6 (33.3%)	0.73	0.73
Current exercise status	Yes N (%)	28 (31.1%)	38 (42.2%)	67 (74.4%)	0.16	**< 0.001**	**< 0.001**	22 (30.6%)	29 (40.3%)	5 (27.8%)	8 (44.4%)	0.30	0.49
Current alcohol use	Yes N (%)	37 (41.1%)	30 (33.3%)	38 (42.2%)	0.36	1.00	0.28	31 (43.1%)	24 (33.3%)	6 (33.3%)	6 (33.3%)	0.30	1.00
Blood collection time	AM N (%)	28 (31.1%)	23 (25.6%)	42 (46.7%)	0.51	**0.046**	**0.005**	25 (34.7%)	19 (26.4%)	3 (16.7%)	4 (22.2%)	0.37	1.00
Fasting time	≥ 8 hours N (%)	11 (12.2%)	10 (11.1%)	41 (45.6%)	1.00	**< 0.001**	**< 0.001**	11 (15.3%)	9 (12.5%)	0 (0%)	1 (5.6%)	0.81	1.00
Duration from first onset (years)	Mean (s.d)	7.0 (8.4)	7.9 (8.5)	N/A	0.48	N/A	N/A	7.1 (8.8)	8.4 (7.4)	8.7 (7.5)	13.7 (13.4)	0.32	0.19
Duration from first medication (years)	Mean (s.d)	3.8 (6.2)	7.4 (8.9)	N/A	**0.002**	N/A	N/A	3.4 (5.5)	6.5 (6.8)	5.5 (8.5)	10.9 (14.4)	**0.003**	0.18
Clinical assessments													
BPRS	Mean (s.d.)	27.4 (4.7)	28.3 (5.3)	20.9 (1.5)	0.26	**< 0.001**	**< 0.001**	27.6 (4.9)	28.4 (5.1)	26.6(3.8)	27.9(6.4)	0.39	0.45
MADRS	Mean (s.d.)	26.9 (10.7)	17.2 (9.9)	3.6 (4.1)	**< 0.001**	**< 0.001**	**< 0.001**	27.4 (9.8)	17.8 (10.5)	25.2 (13.9)	15.0 (7.4)	**< 0.001**	**0.01**
YMRS	Mean (s.d.)	2.1 (2.5)	6.1 (7.4)	1.2 (1.9)	**< 0.001**	**0.017**	**< 0.001**	2.3 (2.6)	5.5 (6.4)	1 (1.9)	8.11 (10.7)	**< 0.001**	**0.01**
HAM-A	Mean (s.d.)	15.7 (7.5)	9.7 (5.7)	2.1 (1.8)	**< 0.001**	**< 0.001**	**< 0.001**	16.2 (7.0)	9.6 (5.7)	14.0 (9.3)	10.1 (5.9)	**< 0.001**	0.15
Medications													
Antipsychotics (AP)	N (%)	41 (46%)	66 (73%)	0 (0%)	**< 0.001**	**< 0.001**	**< 0.001**	34 (47.2%)	55 (76.4%)	7 (38.9%)	11 (66.1%)	**0.002**	0.318
Mood stabilizer (MS)	N (%)	15 (17%)	63 (70%)	0 (0%)	**< 0.001**	**< 0.001**	**< 0.001**	11 (15.3%)	51 (70.8%)	4 (22.2%)	12 (66.7%)	**< 0.001**	**< 0.001**
Antidepressants (AD)	N (%)	74 (82%)	26 (29%)	0 (0%)	**< 0.001**	**< 0.001**	**< 0.001**	59 (81.9%)	23 (31.9%)	15 (83.3%)	3 (16.7%)	**< 0.001**	**< 0.001**
Benzodiazepines/hypnotics (BZD/HNT)	N (%)	63 (70%)	53 (59%)	0 (0%)	0.16	**< 0.001**	**< 0.001**	41 (56.9%)	47 (65.3%)	12 (66.7%)	14 (77.8%)	0.39	0.71

[a] Abbreviations: MDD, major depressive disorder; BD, bipolar disorder; HC, healthy control; BMI, body mass index; BPRS, Brief Psychiatric Rating Scale; MADRS, Montgomery-Asberg Depression Rating Scale; YMRS, Young Mania Rating Scale; HAM-A, Hamilton Anxiety Rating Scale; AP, antipsychotics; MS, mood stabilizer; AD, antidepressants; BZD/HNT, benzodiazepines/hypnotics; N/A, not applicable.

[b] **Statistically significant differences between groups (MDD, BD, and HC) were examined regarding demographics and clinical variables. Student's *T*-test was performed for continuous variables and Fisher's exact test was conducted for dichotomous variables. *P*-value < 0.05 was considered statistically significant, which is denoted by Bold font.**

2. Construction of model to distinguish MDD from BD

When all 210 candidate features were considered in developing the model, the number of selected features ranged from 8 to 25, representing a fluctuation in model selection (Figure 3A). Twenty-six features were selected at least once, and 9 were selected at least 90 times (Figure 3B and Table 4). The most frequent model that was generated comprised 11 features, with a model probability value of 0.70. A total of 10 unique models were generated, but the model frequency was not necessarily associated with the model probability for the data that were used (Table 5). Due to the fluctuation in the number of selected features and the fact that there is no unique model that was absolutely supported, feature extraction and model averaging were performed for all 100 models.

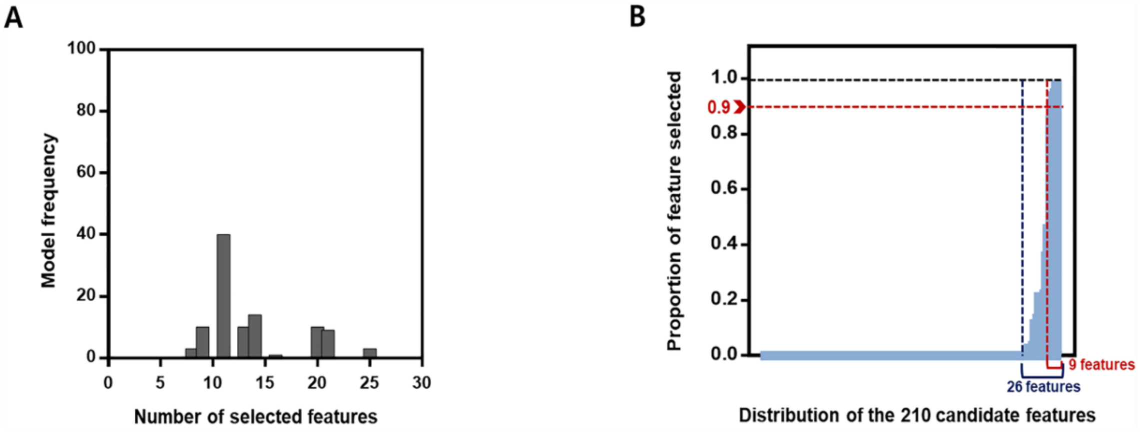


Figure 3. Feature selection and extraction across 100 models generated by repeated application of LASSO regression with ten-fold crossvalidation on the training set. A. Frequency of selected features for the 100 obtained models. Perturbed distribution between model frequency and number of selected features was showed. **B.** Proportion of features selected with respect to the 210 candidate features. A total of 26 features among 210 candidate features had more than 0 of the value of proportion of feature selected. Total 9 features had more than 0.9 of the value of proportion of feature selected value.

Table 4. Proportion of features selected for the 210 candidate features (210 proteins/210 peptides) used to determine the model for discriminating MDD from BD[a]

Feature	Proportion of feature selected
Protein_Peptide sequence	
A1AG1_SDVVYTDWK	0
A1AG2_EQLGEFYEALDCLCIPR	0
A1BG_CEGPIPDVTFELLR	0
A2AP_QEDDLANINQWVK	0
A2MG_VYDYYETDEFAIAEYNAPCSK	0
ADIPO_GDIGETGVPGAEGPR	0
AFAM_HFQNLGK	0
ALBU_LVNEVTEFAK	0
ALDOA_ALQASALK	0.87
ALS_LHSLHLEGSCLGR	0
AMBP_CVLFPYGGCQGNGNK	0
AMPN_AQIINDAFNLASAHK	0.12
ANAG_DFCGCHVAWSGSQLR	0
ANGT_LQAILGVPWK	0
ANT3_VWELSK	0
APOA_NPDAVAAPYCYTR	0
APOA1_QGLLPVLESFK	0
APOA2_SPELQAEAK	0
APOA4_GNTEGLQK	0
APOB_ITLPDFR	0
APOC2_ESLSSYWESAK	0
APOC3_DYWSTVK	0
APOD_VLNQELR	0
APOE_EQVAEVR	0
APOF_SLPTEDCENEK	0
APOH_ATVVYQGER	0
APOL1_LNILNNNYK	0
APOM_FLLYNR	0
ASSY_IDIVENR	0
ATS13_LFINVAPHAR	0
B2MG_VNHVTLSQPK	0
B3AT_LSVPDGFK	0
BPIB1_ALGFEAAESSLTK	0
BTD_VDLITFDTPFAGR	0
BTK_LVQLYGVCTK	0
C1QA_SLGFCDTTNK	0

C1QB_LEQGENVFLQATDK	**1**
C1QC_QTHQPPAPNSLIR	0
C1R_NIGEFCGK	0
C1RL_GSEAINAPGDNPAK	0
C1S_CEYQIR	0
C4BPA_LSCSYSHWSAPAPQCK	0
CA2D1_VLLDAGFTNELVQNYWSK	0
CAH1_GGPFSDSYR	0
CAH2_YGDFGK	0
CALR_FVLSSGK	0
CBG_HLVALSPK	0
CBL_GTEPIVVDPFDPR	0
CBPB2_YPLYVLK	0
CBPN_VQNECPGITR	0
CCND2_ACQEQIEAVLLNSLQQYR	0
CD14_VLDLSCNR	0
CD5L_CYGPGVGR	0
CEAM5_TLTLFNVTR	0
CERU_EYTDASFTNR	0.03
CFAB_VSEADSSNADWVTK	0
CFAH_CVEISCK	0
CFAI_EANVACLDLGFQQGADTQR	0
CHLE_AEEILSR	0
CLUS_EIQNAVNGVK	0
CNDP1_AIHLDLEEYR	0
CO1A1_VLCDDVICDETK	0
CO2_HAFILQDTK	0
CO3_DSCVGSLVVK	0
CO4A_ANSFLGEK	0
CO4A2_IAVQPGTVGPQGR	0
CO5_IDTALIK	0
CO6_GFVVAGPSR	0
CO7_VLFYVDSEK	0
CO8A_AIDEDCSQYEPIPGSQK	0
CO8B_CEGFVCAQTGR	0
CO8G_AGQLSVK	0
CO9_TSNFNAAISLK	0
COAA1_GTHVWVGLYK	0.22
COHA1_QAAYNADSGLK	0
COMP_DTDLDGFPDEK	0
CP3A5_DTINFLSK	0
CPN2_QLVCPVTR	0
CRP_ESDTSYVSLK	0

CXCL7_NIQSLEVIGK	0.04
CYTC_ALDFAVGEYNK	0
DCMC_LCAWYLYGEK	0
DOPO_VISTLEEPTPQCPTSQGR	0
DSG2_ILDVNDNIPVVENK	0
DSG3_LAEISLGVDGEGK	**1**
ECM1_FCEAEFSVK	0
EGFR_CNLLEGEPR	0
EXOSX_SGPLPSAER	0
F13A_STVLTIPEIIIK	0
F13B_VLHGDLIDFVCK	0
FA10_QEDACQGDSGGPHVTR	0.37
FA11_SCALSNLACIR	0
FA12_CFEPQLLR	0
FA5_GEYEEHLGILGPIIR	0
FA7_LHQPVVLTDHVVPLCLPER	0
FA9_NCELDVTCNIK	0.87
FBLN1_CVDVDECAPPAEPCGK	**1**
FBLN3_NPCQDPYILTPENR	0
FBN2_FNLSHLGSK	0
FBRL_NGGHFVISIK	0
FCG3A_AVVFLEPQWYR	0
FCGBP_AIGYATAADCGR	**1**
FCN3_YGIDWASGR	0
FETUA_CNLLAEK	0
FETUB_SQASSCSLQSSDSVPVGLCK	0
FHR1_TGESAEFVCK	0
FHR3_AQTTVTCTEK	**1**
FHR5_TGDAVEFQCK	0
FIBA_VQHIQLLQK	0
FIBB_QGFGNVATNTDGK	0
FIBG_ASTPNGYDNGIIWATWK	0
FINC_VPGTSTSATLTGLTR	0
FMR1_EPCCWWLAK	0
FTCD_SDLQVAAK	0
GELS_QTQVSVLPEGGETPLFK	0
GGH_YLESAGAR	0
GPX3_NSCPPTSELLGTSDR	**0.97**
HABP2_FCEIGSDDCYVGDGYSYR	0
HBA_VGAHAGEYGAEALER	0
HEM3_ELEHALEK	0
HEMO_SGAQATWTELPWPHEK	0
HEP2_TLEAQLTPR	0

HGFA_LEACESLTR	0.22
HPT_VGYVSGWGR	0
HPTR_VVLHPNYHQVDIGLIK	0
HRG_QIGSVYR	0
IBP2_LIQGAPTIR	0
IBP3_YGQPLPGYTTK	0
IBP5_GVCLNEK	0
IC1_LLDSLPSDTR	0
ICAM1_VELAPLPSWQPVGK	0
IDS_QSTEQAIQLLEK	0
IGF2_GIVEECCFR	0
IGHG1_TPEVTCVVVDVSHEDPEVK	0
IGHG3_SCDTPPPCPR	0
IGHM_QVGSGVTTDQVQAEAK	**1**
IGKC_DSTYSLSSTLTLSK	0
IL5_ETLALLSTHR	0
INSR_VCHLLEGEK	0.47
IPSP_AAAATGTIFTFR	0
ITA2B_IVLLDVPVR	0
ITIH1_GSLVQASEANLQAAQDFVR	0
ITIH2_IYLQPGR	**0.96**
ITIH3_EVSFDVELPK	0.22
ITIH4_NVVFVIDK	0.03
JAK2_SDNIIFQFTK	0
K1C9_TLLDIDNTR	0
K2C1_SLVNLGGSK	0
K2C5_ISISTSGGSFR	0
KAIN_GFQHLLHTLNLPGHGLETR	0.22
KCRM_ELFDPIISDR	0
KI21B_AQEQGVAGPEFK	0
KLKB1_LVGITSWGEGCAR	0
KNG1_QVVAGLNFR	0
LA_IGCLLK	0
LAMP1_ALQATVGNSYK	0
LAMP2_IPLNDLFR	0
LBP_GLQYAAQEGLLALQSELLR	0
LCAT_SSGLVSNAPGVQIR	0
LDHB_SADTLWDIQK	0.14
LUM_LPSGLPVSLLTLYLDNNK	0
LYAM1_AEIEYLEK	0
LYRIC_WNSVSPASAGK	0
LYSC_STDYGIFQINSR	0
MBL2_WLTFSLGK	0

MCAT_EGITGLYR	0
MIA_GQVVYVFSK	0
MSH2_DIYQDLNR	0
NPC2_SEYPSIK	0
PAFA_ASLAFLQK	0.12
PEDF_ALYYDLISSPDIHGTYK	0
PERM_VVLEGGIDPILR	0
PGFRB_LPGFHGLR	0
PGRP2_EFTEAFLGCPAIHPR	0
PHLD_TLLLVGSPTWK	0
PI16_WDEELAAFAK	0
PLF4_ICLDLQAPLYK	**1**
PLMN_YEFLNGR	0
PLTP_AVEPQLQEEER	0.03
PON1_IHVYEK	0
PRDX2_TDEGIAYR	0
PROC_GDSPWQVVLLDSK	0.47
PROF1_STGGAPTFNVTVTK	0
PROS_VYFAGFPR	0
PZP_HQDGSYSTFGER	0
RENI_LFDASDSSSYK	0
RET4_QEELCLAR	0
S10A9_LGHPDTLNQGEFK	0
SAA1_FFGHGAEDSLADQAANEWGR	0
SAMP_IVLGQEQDSYGGK	0
SHAN3_FEDHEIEGAHLPALTK	0
SHBG_TSSSFEVR	0
SPP24_DYYVSTAVCR	0
TENX_LQGLIPGAR	0
TETN_CFLAFTQTK	0
THBG_NALALFVLPK	0
THRB_VTGWGNLK	0
TNR5_YCDPNLGLR	0
TRFE_CSTSSLLEACTFR	0.23
TRFL_CSTSPLLEACEFLR	0
TTHY_AADDTWEPFASGK	0
UROM_VLNLGPITR	0
URP2_VVLAGGVAPALFR	0
VASN_LAGLGLQQLDEGLFSR	0
VGFR3_SGVDLADSNQK	0
VINC_QVATALQNLQTK	0
VTDB_YTFELSR	0
VTNC_CTEGFNVDK	0

VWF_VTVFPIGIGDR	0
ZA2G_CLAYDFYPGK	0

[a] The proportion of feature selected for the 210 candidate features are listed. The 9 selected features chosen for the developed model are shown in bold. Protein_peptide sequence is listed for each feature.

Table 5. Summary of unique models originating from combinations of selected features[a]

Model #	Unique model (combination of selected features)	Number of features combined	Model Frequency	Model probability
Model 1	ALDOA_ALQASALK + AMPN_AQIINDAFNLASAHK + C1QB_LEQGENVFLQATDK + COAA1_GTHVWVGLYK + DSG3_LAEISLGVDGEGK + FA10_QEDACQGDSGGPHVTR + FA9_NCELDVTCNIK + FBLN1_CVDVDECAPPAEPCGK+FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR + HGFA_LEACESLTR + IGHM_QVGSGVTTDQVQAEAK + INSR_VCHLLEGEK + ITIH2_IYLQPGR + ITIH3_EVSFDVELPK + KAIN_GFQHLLHTLNLPGHGLETR + PAFA_ASLAFLQK + PLF4_ICLDLQAPLYK + PROC_GDSPWQVVLLDSK + TRFE_CSTSSLLEACTFR + (intercept)	21	9	0.0010
Model 2	ALDOA_ALQASALK + C1QB_LEQGENVFLQATDK + DSG3_LAEISLGVDGEGK + FA9_NCELDVTCNIK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR + IGHM_QVGSGVTTDQVQAEAK + ITIH2_IYLQPGR + PLF4_ICLDLQAPLYK + (intercept)	11	40	0.7018
Model 3	ALDOA_ALQASALK + C1QB_LEQGENVFLQATDK + DSG3_LAEISLGVDGEGK + FA10_QEDACQGDSGGPHVTR + FA9_NCELDVTCNIK + FBLN1_CVDVDECAPPAEPCGK+FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR + IGHM_QVGSGVTTDQVQAEAK + INSR_VCHLLEGEK + ITIH2_IYLQPGR + PLF4_ICLDLQAPLYK + PROC_GDSPWQVVLLDSK + (intercept)	14	14	0.1229
Model 4	ALDOA_ALQASALK + AMPN_AQIINDAFNLASAHK + C1QB_LEQGENVFLQATDK + CERU_EYTDASFTNR + COAA1_GTHVWVGLYK +DSG3_LAEISLGVDGEGK + FA10_QEDACQGDSGGPHVTR + FA9_NCELDVTCNIK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR + HGFA_LEACESLTR + IGHM_QVGSGVTTDQVQAEAK + INSR_VCHLLEGEK + ITIH2_IYLQPGR + ITIH3_EVSFDVELPK + ITIH4_NVVFVIDK + KAIN_GFQHLLHTLNLPGHGLETR + LDHB_SADTLWDIQK + PAFA_ASLAFLQK + PLF4_ICLDLQAPLYK + PLTP_AVEPQLQEEER + PROC_GDSPWQVVLLDSK + TRFE_CSTSSLLEACTFR + (intercept)	25	3	8.13E-06
Model 5	C1QB_LEQGENVFLQATDK + DSG3_LAEISLGVDGEGK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR +IGHM_QVGSGVTTDQVQAEAK + ITIH2_IYLQPGR + PLF4_ICLDLQAPLYK + (intercept)	10	9	0.0761
Model 6	ALDOA_ALQASALK + C1QB_LEQGENVFLQATDK + DSG3_LAEISLGVDGEGK + FA9_NCELDVTCNIK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR +FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR + IGHM_QVGSGVTTDQVQAEAK + INSR_VCHLLEGEK + ITIH2_IYLQPGR + PLF4_ICLDLQAPLYK + PROC_GDSPWQVVLLDSK + (intercept)	13	10	0.0885
Model 7	C1QB_LEQGENVFLQATDK + CXCL7_NIQSLEVIGK + DSG3_LAEISLGVDGEGK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK +IGHM_QVGSGVTTDQVQAEAK + PLF4_ICLDLQAPLYK + (intercept)	8	3	0.0043
Model 8	ALDOA_ALQASALK + C1QB_LEQGENVFLQATDK + DSG3_LAEISLGVDGEGK + FA10_QEDACQGDSGGPHVTR + FA9_NCELDVTCNIK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR + IGHM_QVGSGVTTDQVQAEAK + INSR_VCHLLEGEK + ITIH2_IYLQPGR + LDHB_SADTLWDIQK +PLF4_ICLDLQAPLYK + PROC_GDSPWQVVLLDSK + TRFE_CSTSSLLEACTFR + (intercept)	16	1	0.0024
Model 9	ALDOA_ALQASALK + C1QB_LEQGENVFLQATDK + COAA1_GTHVWVGLYK + DSG3_LAEISLGVDGEGK + FA10_QEDACQGDSGGPHVTR + FA9_NCELDVTCNIK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK + GPX3_NSCPPTSELLGTSDR + HGFA_LEACESLTR +IGHM_QVGSGVTTDQVQAEAK + INSR_VCHLLEGEK + ITIH2_IYLQPGR + ITIH3_EVSFDVELPK + KAIN_GFQHLLHTLNLPGHGLETR + LDHB_SADTLWDIQK + PLF4_ICLDLQAPLYK + PROC_GDSPWQVVLLDSK + TRFE_CSTSSLLEACTFR + (intercept)	20	10	4.56E-04
Model 10	C1QB_LEQGENVFLQATDK + CXCL7_NIQSLEVIGK + DSG3_LAEISLGVDGEGK + FBLN1_CVDVDECAPPAEPCGK + FCGBP_AIGYATAADCGR + FHR3_AQTTVTCTEK +GPX3_NSCPPTSELLGTSDR + IGHM_QVGSGVTTDQVQAEAK + PLF4_ICLDLQAPLYK + (intercept)	9	1	0.0025

[a] Ten unique models were generated. Number of features, frequency, and model probability for each unique model are listed. Protein_peptide sequence of the components is listed for each combination.

The generating average model consisted of 9 features, which had a proportion of feature selected ≥ 0.9 (Figure 4A). The 9 selected features were as follows: complement C1q subcomponent subunit B (C1QB), desmoglein-3 (DSG3), fibulin-1 (FBLN1), IgG Fc-binding protein (FCGBP), complement factor H-related protein 3 (FHR3), glutathione peroxidase 3 (GPX3), immunoglobulin heavy constant mu (IGHM), inter-alpha-trypsin inhibitor heavy chain H2 (ITIH2), and platelet factor 4 (PLF4). The average coefficients for the 9 features across the 100 models are presented in Figure 4B and Table 6. Five features—C1QB, FBLN1, FCGBP, FHR3, and IGHM—had average coefficients greater than 0, which were higher in BD versus MDD patients. In contrast, 4 features—DSG3, GPX3, ITIH2, and PLF4—had average coefficients of less than 0, which were higher in MDD versus BD patients. The 9-feature average model showed good discriminatory performance between MDD and BD when applied to the training set (AUC = 0.84) (Figure 5A) and when extrapolated to the test set (AUC = 0.81) (Figure 5B). In addition, the model showed good discriminatory power when applied to the combined set (AUC = 0.83), resulting in 72% sensitivity, 82% specificity, and 77% accuracy at the optimal cutoff (Youden Index) of 0.54 (Figure 5C). Furthermore, the discriminatory performance was similar when the model was applied to the MRM-MS data of only young BD and MDD patients (age <35 years) (AUC = 0.84) (data not shown).

The average model was applied to the data on MDD and BD without current hypomanic/manic/mixed symptoms, performing well (AUC = 0.83) (Figure 6A), and to only drug-free patients (11 MDD and 10 BD), showing excellent performance (AUC = 0.96) (Figure 6B). When the model was applied to the data on the patient groups and HCs, the AUC value was 0.87 (MDD vs HC) and 0.86 (BD vs HC) (Figure 6C-D).

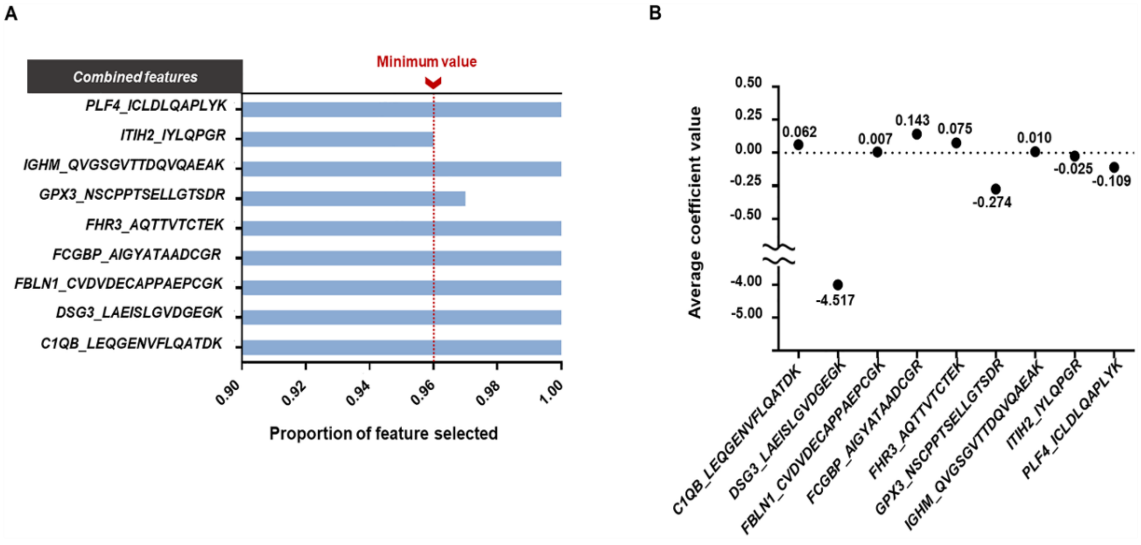
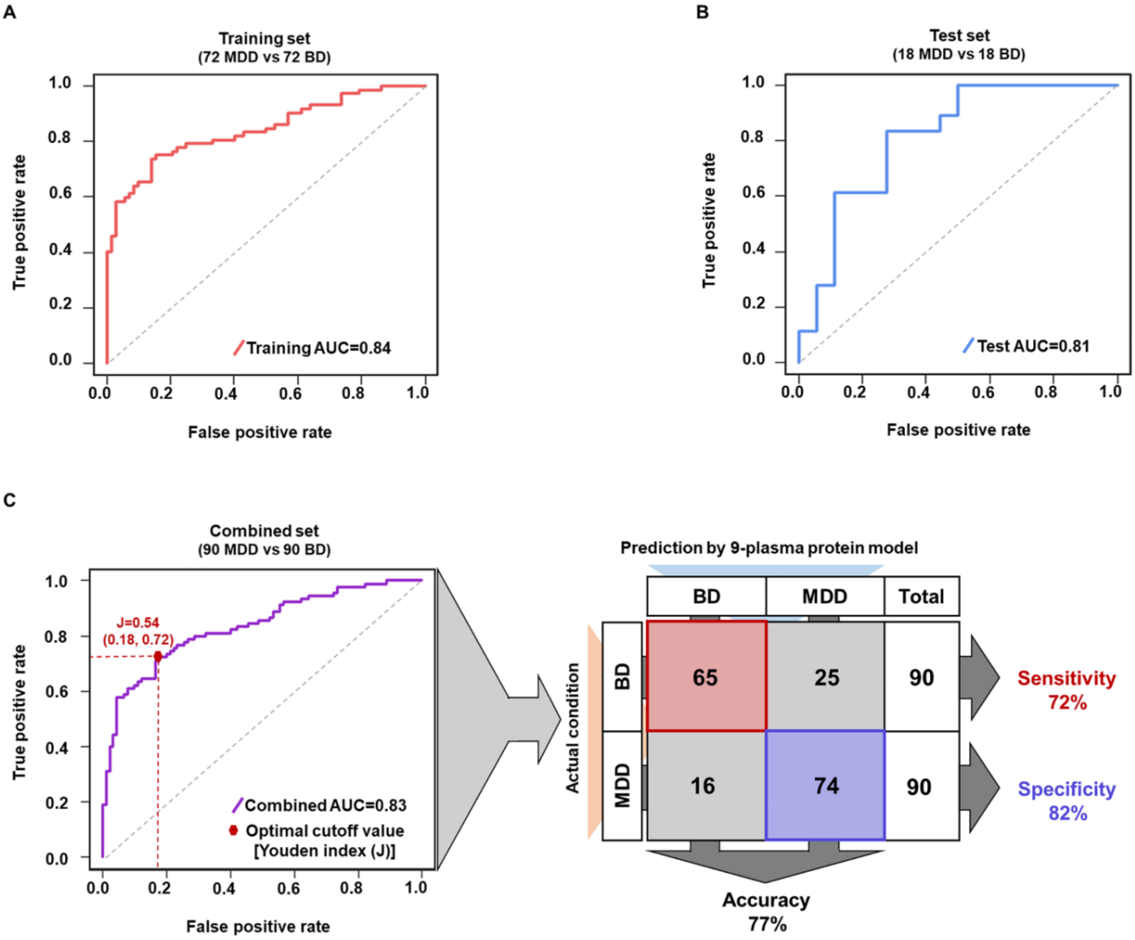


Figure 4. The nine features of the developed model. Proportion of the features that were selected and the average coefficients for the nine features are shown. **A.** The proportion of each feature that was selected across the 100 models. The minimum value is indicated by the dotted line (red). Features are designated by the corresponding protein_peptide sequence. **B.** Average coefficient value for each feature across the 100 models. C1QB, complement C1q subcomponent subunit B; DSG3, desmoglein-3; FBLN1, fibulin-1; FCGBP, IgG Fc-binding protein; FHR3, complement factor H-related protein 3; GPX3, glutathione peroxidase 3; IGHM, immunoglobulin heavy constant mu; ITIH2, inter-alpha-trypsin inhibitor heavy chain H2; PLF4, platelet factor 4.

Table 6. Confidence interval of weighted average coefficient for the nine features of the developed model

Combined features	**Weighted average coefficient**	**Standard deviation**	**Lower bound**	**Upper bound**	**95% confidence interval (CI)**
C1QB_LEQGENVFLQATDK	0.0623	0.0090	0.0606	0.0641	0.0035
DSG3_LAEISLGVDGEGK	-4.5171	0.3162	-4.5791	-4.4551	0.1240
FBLN1_CVDVDECAPPAEPCGK	0.0070	0.0015	0.0067	0.0073	0.0006
FCGBP_AIGYATAADCGR	0.1425	0.0627	0.1302	0.1548	0.0246
FHR3_AQTTVTCTEK	0.0751	0.0190	0.0714	0.0788	0.0075
GPX3_NSCPPTSELLGTSDR	-0.2740	0.1811	-0.3095	-0.2385	0.0710
IGHM_QVGSGVTTDQVQAEAK	0.0097	0.0022	0.0093	0.0102	0.0008
ITIH2_IYLQPGR	-0.0247	0.0219	-0.0290	-0.0204	0.0086
PLF4_ICLDLQAPLYK	-0.1086	0.0111	-0.1108	-0.1064	0.0043



	BD	MDD	Total
BD	65	25	90
MDD	16	74	90

Figure 5. Performance of model in discriminating between MDD and BD based on AUROC curves. The developed model consisted of nine proteins. **A.** The performance of the model in the training set (72 MDD patients vs 72 BD patients) (Table 1) and **B.** test set (18 MDD patients vs 18 BD patients) (Table 1). **C.** Performance of the model (left panel) and its corresponding confusion matrix (right panel) in the combined set (90 MDD patients vs 90 BD patients) (Table 1). The optimal cutoff [Youden Index (J)] is presented in red font and as a dotted line in the AUROC curve (left panel). Sensitivity, specificity, and accuracy corresponding to the optimal cutoff are presented in the confusion matrix (right panel). AUC, area under the curve; MDD, major depressive disorder; BD, bipolar disorder; AUROC, area under receiver operating characteristics; J, Youden index.

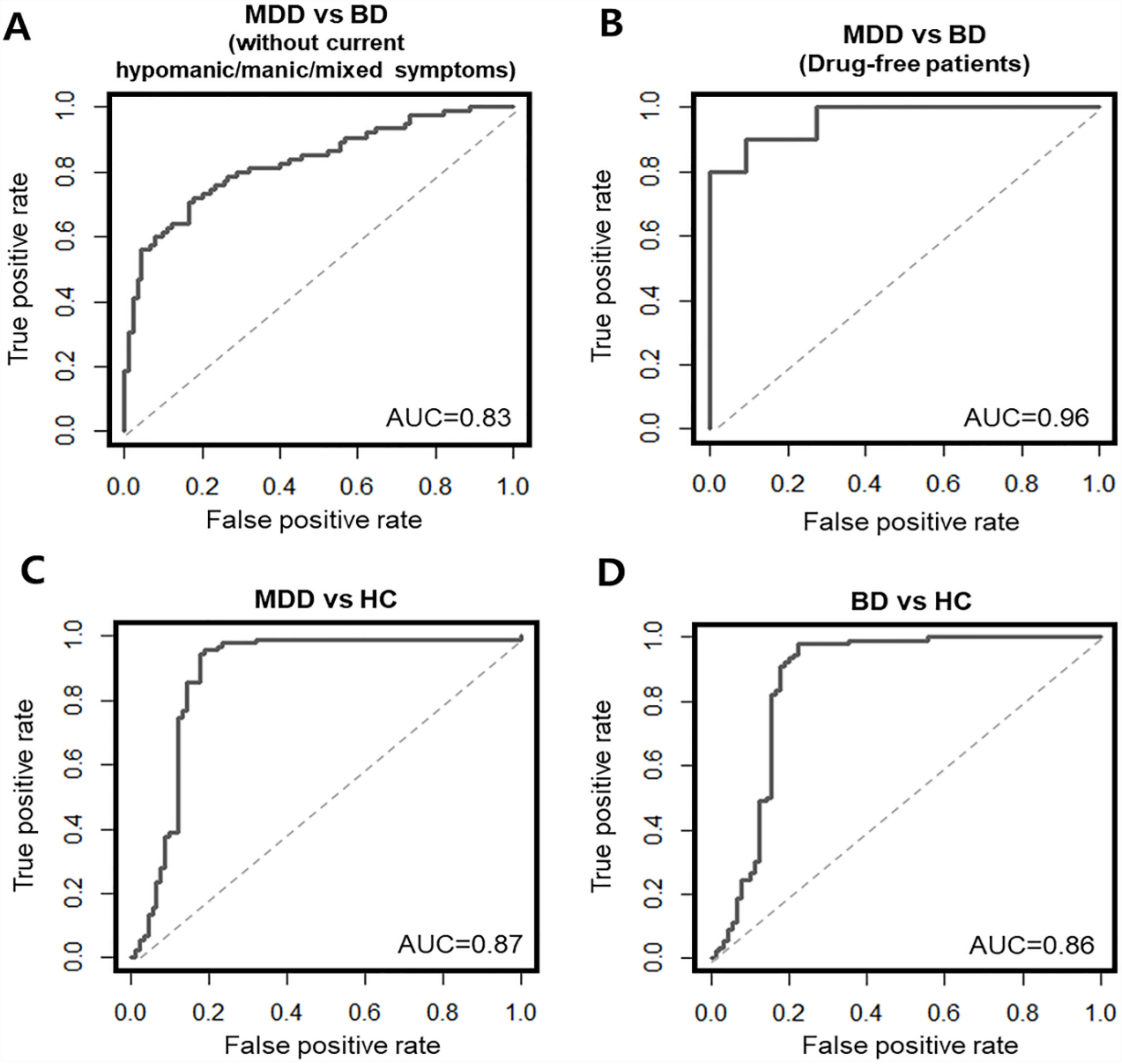


Figure 6. AUROC curves representing model performance in discriminating between MDD and BD (without current hypomanic/manic/mixed symptoms), between MDD and BD (drug-free patients), and between patient groups and HC. The developed model for discriminating MDD from BD was applied to data on patient groups and HCs in all study subjects. **A.** The model's performance in the patient groups (90 MDD patients vs 75 BD patients without current hypomanic/manic/mixed symptoms). **B.** The model's performance in drug-free MDD and BD patients (11 MDD vs 10 BD). **C.** The model's performance in the MDD group and HC (90 MDD patients vs 90 HCs). **D.** The model's performance in the BD group and HC (90 BD patients vs 90 HCs). AUC, area under the curve; MDD, major depressive disorder; BD, bipolar disorder; HC, healthy control; AUROC, area under receiver operating characteristics.

3. Nine features in the model

The levels of protein abundances for the 9 features in the training set are shown in Figure 7. All 9 features differed significantly between MDD and BD. C1QB, FBLN1 FCGBP, FHR3 and IGHM were upregulated in BD, whereas DSG3, GPX3, ITIH2, and PLF4 were upregulated in MDD. C1QB showed the most statistically significant difference (*P*-value = 5.27E-04) among features that were upregulated in BD. Conversely, DSG3 had the most significant difference (*P*-value = 1.43E-04) among those that were upregulated in BD (Table 6). The expression patterns of the 9 features (ie, fold-change between MDD and BD) corresponded with the direction of the average coefficients (Figure 4 and Table 6).

The levels of the 9 features were also measured in all study subjects (90 MDD, 90 BD, and 90 HCs) (Figure 8 and Table 7). All 9 features differed significantly between MDD and BD, and the distributions of levels for each feature were similar to those in the training set. Only DSG3 differed significantly between MDD, BD, and HCs. DSG3 was upregulated in MDD but downregulated in BD versus HCs. Whereas PLF4 was upregulated in MDD compared with HCs, C1QB and FBLN1 were downregulated in MDD versus HCs. FCGBP was upregulated and GPX3 was downregulated in BD compared with HC. The remaining features—FHR3, IGHM, and ITIH2—did not differ significantly between patient groups and HCs.

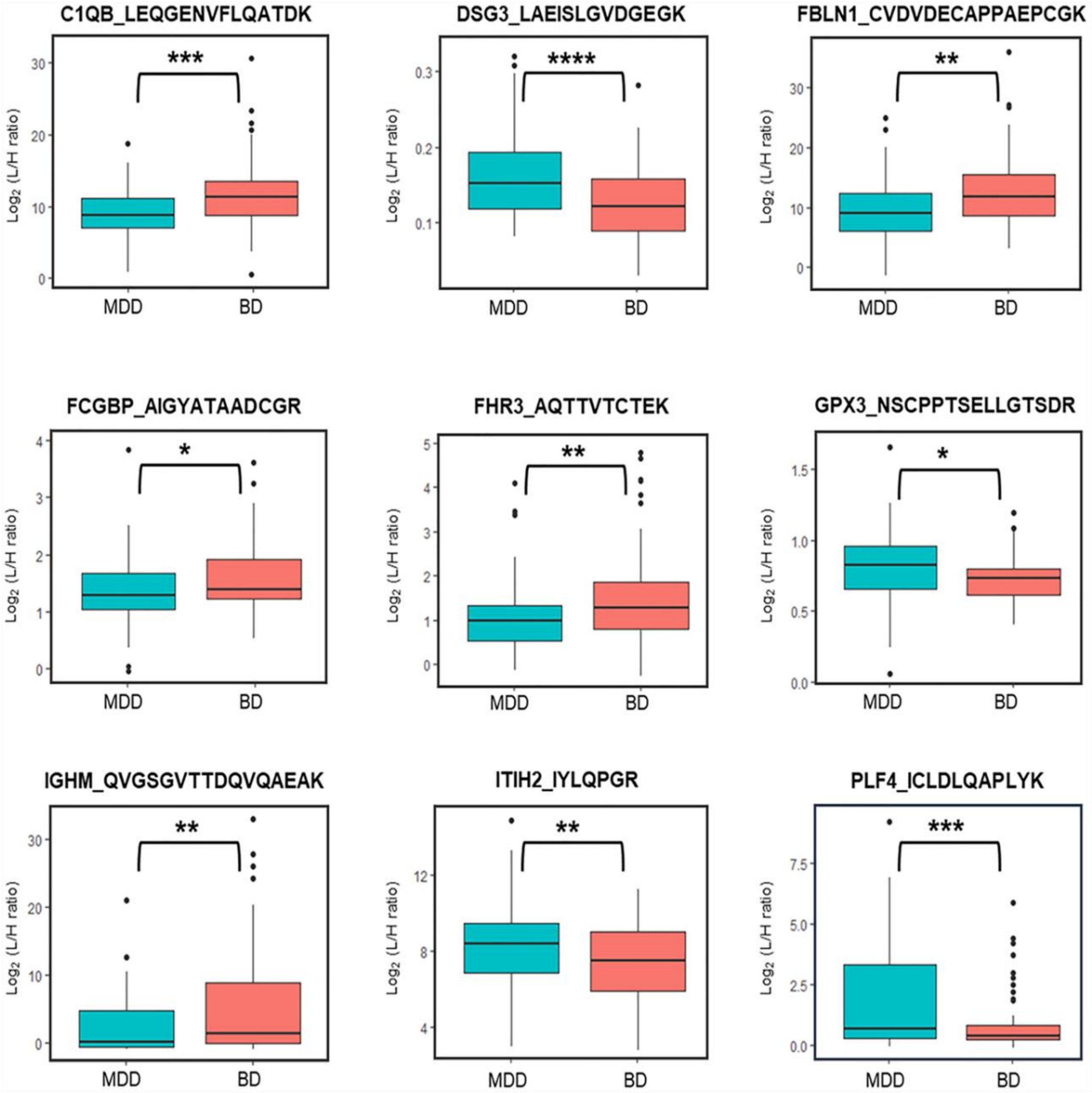


Figure 7. Protein levels of the nine features in the developed model. The distribution of protein levels for MDD and BD patients in the training set. Protein levels are shown as $\log_2$-transformed peak area ratio (light/heavy ratio). An unpaired Student's *T*-test was performed to examine statistically significant differences. L/H ratio, light/heavy ratio; MDD, major depressive disorder; BD, bipolar disorder; *, *P*<0.05; **, *P*<0.01; ***, *P*<0.001; ****, *P*<0.0005.

Table 7. Differences in protein abundance of the nine selected features between MDD and BD in the training set[a]

Features	Protein name	Gene names	Fold-change (MDD/BD)	Student's t-test statistics	P-value[b]
Protein_peptide sequence					
C1QB_LEQGENVFLQATDK	Complement C1q subcomponent subunit B	C1QB	0.785	3.548	**5.27.E-04**
DSG3_LAEISLGVDGEGK	Desmoglein-3	DSG3	1.291	-3.909	**1.43.E-04**
FBLN1_CVDVDECAPPAEPCGK	Fibulin-1	FBLN1	0.776	2.903	**4.29.E-03**
FCGBP_AIGYATAADCGR	IgGFc-binding protein	FCGBP	0.837	2.516	**1.30.E-02**
FHR3_AQTTVTCTEK	Complement factor H-related protein 3	FHR3	0.718	2.714	**7.48.E-03**
GPX3_NSCPPTSELLGTSDR	Glutathione peroxidase 3	GPX3	1.120	-2.487	**1.41.E-02**
IGHM_QVGSGVTTDQVQAEAK	Immunoglobulin heavy constant mu	IGHM	0.415	2.967	**3.53.E-03**
ITIH2_IYLQPGR	Inter-alpha-trypsin inhibitor heavy chain H2	ITIH2	1.115	-2.640	**9.21.E-03**
PLF4_ICLDLQAPLYK	Platelet factor 4	PF4	2.382	-3.692	**3.17.E-04**

[a] **Abbreviations: MDD, major depressive disorder BD, bipolar disorder.**

[b] **Statistically significant differences across the 9 selected features were analyzed by student's *T*-test. Bold font denotes statistical difference at *P*-value < 0.05. Protein_peptide sequence is listed for each feature.**

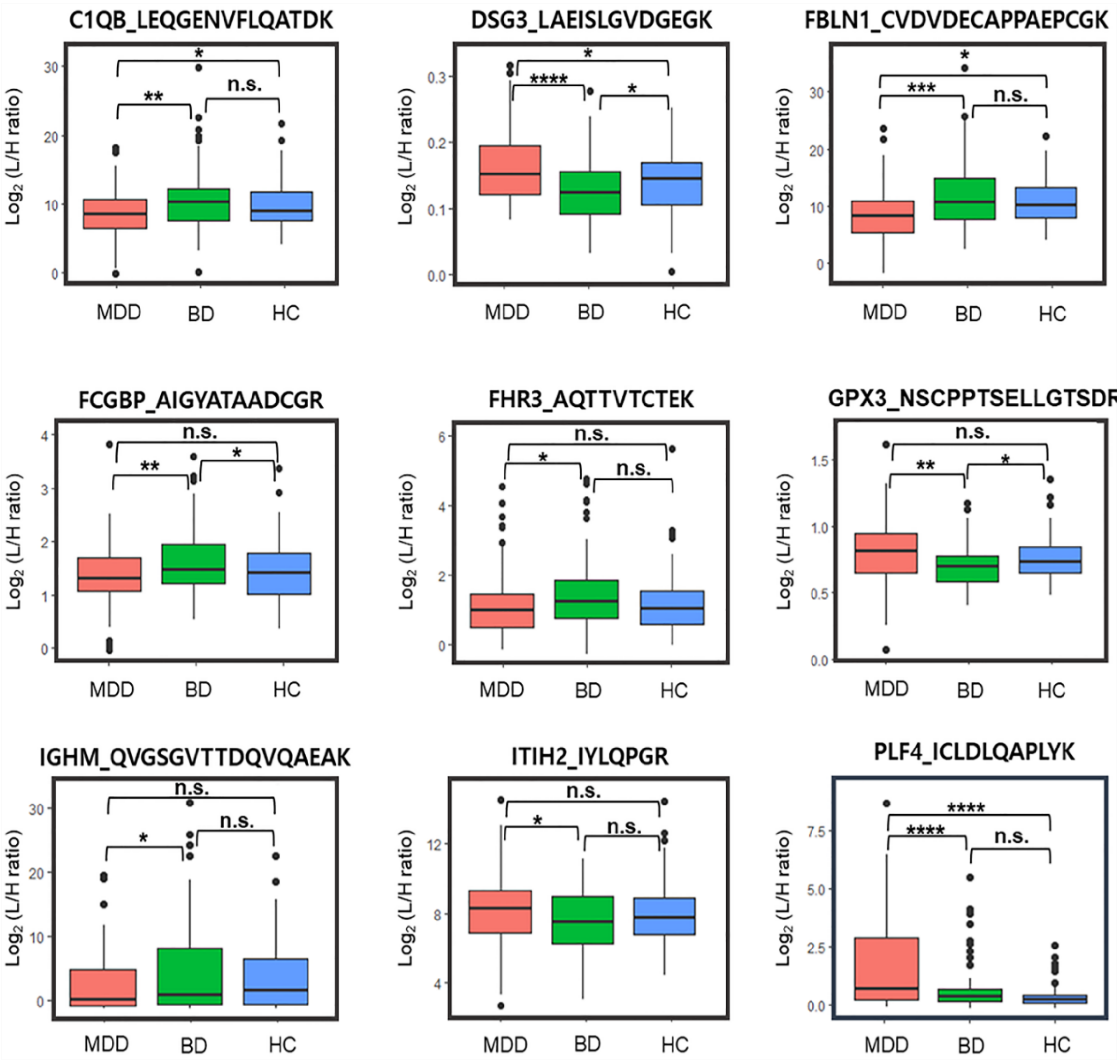


Figure 8. Box-and-whisker plots representing protein abundance of the nine features in all subjects of this study. The distributions of protein abundance between the MDD, BD, and HC groups are shown in all study subjects. Protein abundance is represented by the log$_2$-transformed peak area ratio (Light/Heavy ratio). One-way ANOVA (for three groups) was performed to examine statistically significant differences between the 3 groups. Subsequently, post hoc analysis was performed by Tukey's HSD. L/H ratio, Light/Heavy ratio; HC, healthy control; MDD, major depressive disorder; and BD, bipolar disorder; HSD, honestly significant difference; ANOVA, analysis of variance; *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0005; n.s., no significance.

Table 8. Differences in protein abundance of the nine selected features between MDD, BD, and HC in the study population (90 MDD, 90 BD, and 90 HC)[a]

Features	Protein names	Gene names	MDD vs BD vs HC		MDD vs BD			MDD vs HC			BD vs HC		
			F-statistics	P-value[b] (ANOVA)	Fold-change (MDD/BD)	Mean difference	P-value[b] (Tukey's HSD)	Fold-change (MDD/HC)	Mean difference	P-value[b] (Tukey's HSD)	Fold-change (BD/HC)	Mean difference	P-value[b] (Tukey's HSD)
Protein_peptide sequence													
C1QB_LEQGENVFLQATDK	Complement C1q subcomponent subunit B	C1QB	5.741	**3.62.E-03**	0.822	-1.965	**2.50.E-03**	0.886	-1.474	**1.12.E-02**	1.077	0.792	3.65.E-01
DSG3_LAEISLGVDGEGK	Desmoglein-3	DSG3	10.326	**4.79.E-05**	1.291	0.037	**2.54.E-05**	1.139	0.020	**3.94.E-02**	0.882	-0.030	**3.20.E-02**
FBLN1_CVDVDECAPPAEPCGK	Fibulin-1	FBLN1	7.660	**5.82.E-04**	0.763	-2.881	**5.55.E-04**	0.816	-2.097	**1.72.E-02**	1.069	0.784	5.58.E-01
FCGBP_AIGYATAADCGR	IgGFc-binding protein	FCGBP	5.010	**7.31.E-03**	0.837	-0.276	**9.32.E-03**	0.967	-0.048	8.65.E-01	1.155	0.228	**3.96.E-02**
FHR3_AQTTVTCTEK	Complement factor H-related protein 3	FHR3	3.228	**4.22.E-02**	0.820	0.231	**3.89.E-02**	0.978	-0.028	9.79.E-01	1.192	0.238	2.16.E-01
GPX3_NSCPPTSELLGTSDR	Glutathione peroxidase 3	GPX3	6.389	**1.95.E-03**	1.147	0.106	**1.24.E-03**	1.061	0.048	2.50.E-01	0.925	-0.079	**2.07.E-02**
IGHM_QVGSGVTTDQVQAEAK	Immunoglobulin heavy constant mu	IGHM	4.027	**1.89.E-02**	0.511	-2.597	**1.36.E-02**	0.667	-1.357	3.01.E-01	1.304	1.240	3.66.E-01
ITIH2_IYLQPGR	Inter-alpha-trypsin inhibitor heavy chain H2	ITIH2	3.055	**4.88.E-02**	1.092	0.700	**3.82.E-02**	1.037	0.297	5.50.E-01	0.950	-0.403	3.33.E-01
PLF4_ICLDLQAPLYK	Platelet factor 4	PF4	27.651	**1.22.E-11**	2.589	1.171	**2.26.E-07**	4.619	1.495	**5.14.E-09**	1.784	0.324	2.78.E-01

[a] **Protein_peptide sequence is listed for each feature. Abbreviations: MDD, major depressive disorder; BD, bipolar disorder; HC, healthy control; ANOVA, analysis of variation; HSD, honestly significant difference.**

[b] **Statistically significant differences across the 9 selected features were analyzed by ANOVA. Post hoc analysis was performed by Tukey's HSD. Bold font denotes statistical significance at *P*-value < 0.05.**

4. Covariate analysis of the nine features

Correlations between the 9 selected features in the model and clinical/demographic variables were examined (Table 8). DSG3 was associated significantly with MADRS and HAM-A scores, GPX3 correlated with HAM-A scores and AP use, IGHM was linked to MS and AD use, and FHR3 correlated significantly with AD use. These 4 features and 5 clinical variables were examined by ANCOVA to evaluate the influence of the clinical variables as covariates on the features. Whereas no significant differences were found regarding the covariates, such differences were seen between groups (Table 9), indicating that the levels of the 9 features were associated with differences between groups rather than clinical state or medication use.

Table 9. Correlations between the nine selected features and demographic/clinical variables in the training set[a]

MDD & BD (n=144)		*Gender*	*Age*	*BMI*	*Current smoking status*	*Current exercise status*	*Current alcohol use*	*Blood collection time*	*Fasting time*	*Duration from first onset*	*Duration from first medication*	*BPRS*	*MADRS*	*YMRS*	*HAM-A*	*AP*	*MS*	*AD*	*BZD/HNT*
Protein_peptide sequence																			
C1QB_LEQGENVFLQATDK	Correlation value (r)[b]	0.139	0.122	0.035	-0.041	0.083	-0.073	0.057	-0.029	0.073	0.131	-0.036	0.071	0.071	-0.112	-0.011	0.125	-0.007	-0.044
	P-value[c]	0.097	0.145	0.678	0.625	0.324	0.383	0.497	0.733	0.387	0.120	0.668	0.424	0.395	0.182	0.896	0.136	0.933	0.600
DSG3_LAEISLGVDGEGK	Correlation value (r)[b]	0.067	0.126	-0.015	-0.064	0.128	-0.044	0.060	-0.018	-0.127	-0.114	-0.086	**0.169**	-0.157	**-0.165**	-0.121	0.013	0.005	0.052
	P-value[c]	0.428	0.158	0.859	0.445	0.125	0.601	0.476	0.828	0.131	0.152	0.303	**0.043**	0.061	**0.052**	0.149	0.124	0.952	0.534
FBLN1_CVDVDECAPPAEPCGK	Correlation value (r)[b]	0.101	-0.047	-0.136	-0.033	0.051	-0.071	-0.034	0.100	0.055	0.047	-0.033	-0.122	0.089	0.103	0.010	0.042	-0.132	-0.031
	P-value[c]	0.229	0.576	0.102	0.699	0.541	0.401	0.688	0.232	0.513	0.573	0.695	0.253	0.288	0.164	0.904	0.619	0.119	0.712
FCGBP_AIGYATAADCGR	Correlation value (r)[b]	0.065	-0.067	-0.138	-0.051	0.102	-0.099	-0.007	0.026	-0.028	0.036	-0.016	-0.002	-0.084	-0.139	0.035	0.132	-0.157	-0.068
	P-value[c]	0.441	0.425	0.100	0.544	0.223	0.237	0.936	0.758	0.740	0.673	0.852	0.982	0.317	0.096	0.681	0.115	0.060	0.417
FHR3_AQTTVTCTEK	Correlation value (r)[b]	0.042	0.101	0.126	0.010	-0.029	-0.100	-0.001	-0.049	-0.006	0.124	0.139	-0.088	0.127	-0.037	0.077	0.088	**-0.182**	0.131
	P-value[c]	0.614	0.226	0.133	0.908	0.733	0.233	0.987	0.559	0.942	0.142	0.097	0.293	0.129	0.658	0.358	0.296	**0.028**	0.117
GPX3_NSCPPTSELLGTSDR	Correlation value (r)[b]	0.081	-0.101	-0.159	-0.035	-0.046	0.063	0.090	0.110	0.013	-0.107	-0.055	0.110	-0.047	**-0.154**	**-0.233**	-0.092	-0.010	-0.047
	P-value[c]	0.333	0.226	0.057	0.674	0.581	0.455	0.281	0.188	0.876	0.204	0.514	0.188	0.572	**0.049**	**0.017**	0.353	0.904	0.574
IGHM_QVGSGVTTDQVQAEAK	Correlation value (r)[b]	0.156	-0.149	-0.124	-0.061	0.125	-0.127	0.011	0.128	0.052	0.067	-0.022	-0.082	0.026	-0.127	0.005	**0.202**	**-0.253**	-0.044
	P-value[c]	0.062	0.074	0.173	0.465	0.134	0.128	0.897	0.127	0.535	0.424	0.793	0.330	0.756	0.128	0.954	**0.014**	**0.002**	0.598
ITIH2_IYLQPGR	Correlation value (r)[b]	-0.006	0.024	0.056	0.002	0.081	-0.062	-0.013	0.040	-0.124	-0.104	0.041	0.013	-0.053	0.027	-0.136	-0.154	0.163	-0.059
	P-value[c]	0.946	0.715	0.502	0.981	0.337	0.457	0.875	0.637	0.138	0.202	0.622	0.879	0.524	0.753	0.081	0.068	0.050	0.482
PLF4_ICLDLQAPLYK	Correlation value (r)[b]	0.067	0.092	0.077	0.044	-0.103	-0.004	0.088	-0.148	0.161	-0.043	-0.127	0.128	0.011	-0.013	-0.148	-0.125	0.049	-0.146
	P-value[c]	0.428	0.274	0.360	0.604	0.219	0.958	0.296	0.077	0.054	0.609	0.124	0.127	0.872	0.881	0.220	0.178	0.561	0.058

[a] Protein_peptide sequence is listed for each feature. Abbreviations: MDD, major depressive disorder; BD, bipolar disorder; BMI, body mass index; AP, antipsychotics; MS, mood stabilizer; AD, antidepressants; BZD/HNT, benzodiazepines/hypnotics; BPRS, Brief Psychiatric Rating Scale; MADRS, Montgomery-Asberg Depression Rating Scale; YMRS, Young Mania Rating Scale; HAM-A Hamilton Anxiety Rating Scale.

[b] Correlations between the 9 selected features and continuous variables were examined based on Pearson's correlation. Correlation values between the 9 selected features and dichotomous variables were analyzed by point-biserial correlation.

[c] Bold font denotes statistical significance at *P*-value < 0.05.

Table 10. Covariate analysis of the four features that showed statistically significant correlation with clinical variables in the training set[a]

Features	Covariates	P-value[b] (ANCOVA)
Protein_peptide seqeunce		
DSG3_LAEISLGVDGEGK		
	Group (MDD vs BD)	**0.004**
	MADRS	0.812
	HAM-A	0.280
FHR3_AQTTVTCTEK		
	Group (MDD vs BD)	**0.025**
	AD	0.327
GPX3_NSCPPTSELLGTSDR		
	Group (MDD vs BD)	**0.031**
	AP	0.854
	HAM-A	0.102
IGHM_QVGSGVTTDQVQAEAK		
	Group (MDD vs BD)	**0.026**
	MS	0.313
	AD	0.100

[a] Protein_peptide sequence is listed for each feature. Abbreviations: MDD, major depressive disorder; BD, bipolar disorder; MADRS, Montgomery-Asberg.

[b] Analysis of covariance (ANCOVA) was performed to assess the potential influence of covariates that correlated significantly with the 4 features (DSG3_LAEISLGVDGEGK, FHR3_AQTTVTCTEK, GPX3_NSCPPTSELLGTSDR, and IGHM_QVGSGVTTDQVQAEAK). Bold font denotes statistical significance at *P*-value < 0.05.

5. Bioinformatics analysis of the nine features

The nine features that were included in the final model were subjected to network analysis. Seven (FBLN1, C1QB, FCGBP, ITIH2, DSG3, GPX3, and PLF4) were included in the top network (score = 17), which consisted of 35 molecules (Figure 9). Diseases and functions that were associated with the top network included cell-to-cell signaling and interaction (29 molecules, *P*-value = 5.49E-17 to 1.2E-6), hematological system development and function (27 molecules, *P*-value = 1.19E-15 to 1.12E-2), immune cell trafficking (23 molecules, *P*-value = 1.19E-15 to 2.11E-7), and psychological disorders (2 molecules, *P*-value = 2.28E-3 to 4.49E-2). In addition, the network was associated with canonical pathways, such as LXR/RXR activation, FXR/RXR activation, neuro-inflammation signaling pathway, acute phase response signaling, NF-kB signaling, production of nitric oxide and reactive oxygen species in macrophages, NRF2-mediated oxidative stress response, synaptic long-term potentiation, synaptic long-term depression, and CREB signaling in neurons.

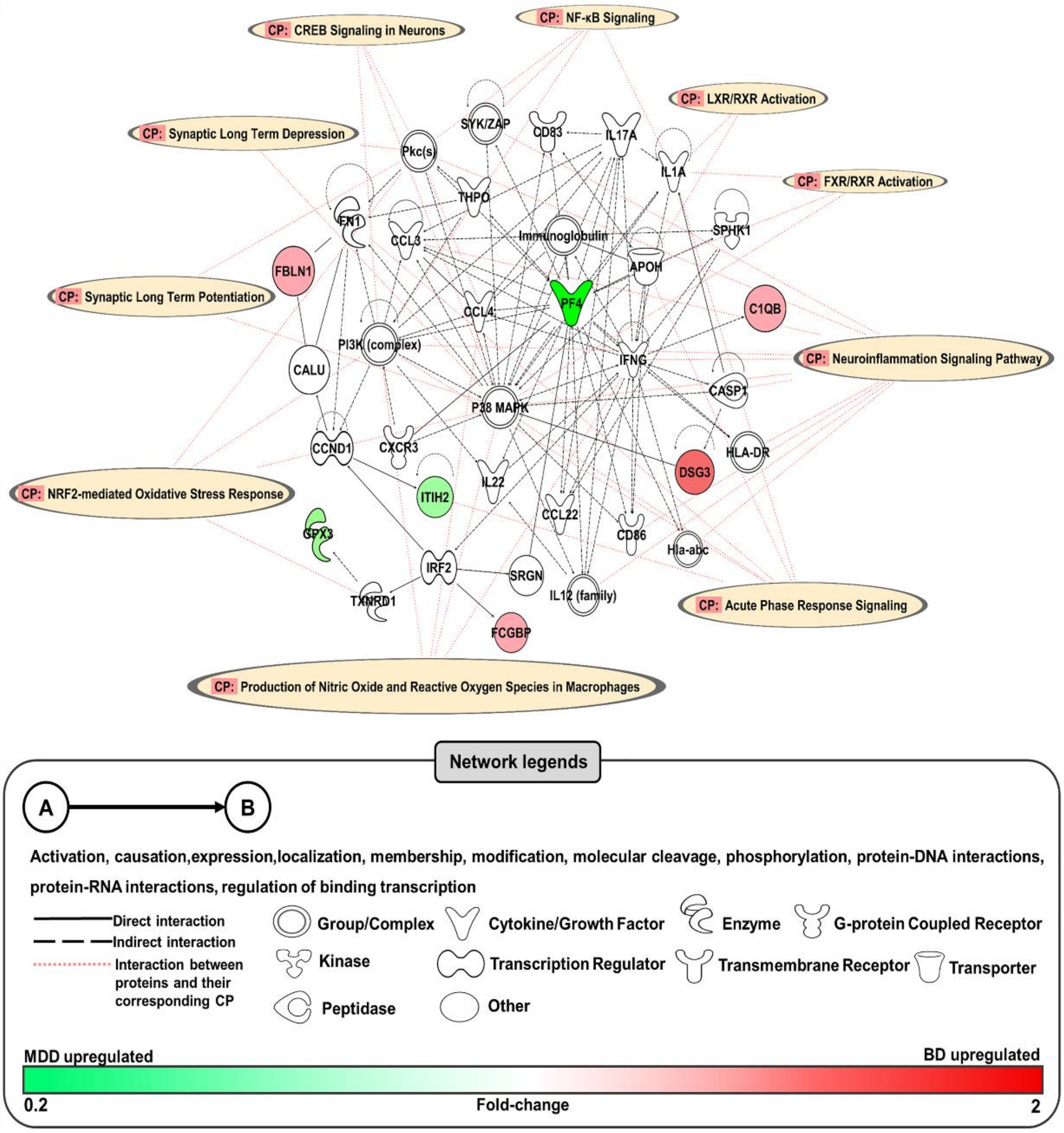

CP: CREB Signaling in Neurons
CP: NF-κB Signaling
CP: LXR/RXR Activation
CP: Synaptic Long Term Depression
CP: FXR/RXR Activation
CP: Synaptic Long Term Potentiation
CP: Neuroinflammation Signaling Pathway
CP: NRF2-mediated Oxidative Stress Response
CP: Acute Phase Response Signaling
CP: Production of Nitric Oxide and Reactive Oxygen Species in Macrophages
SYK/ZAP
CD83
IL17A
Pkc(s)
IL1A
THPO
FN1
CCL3
Immunoglobulin
SPHK1
FBLN1
APOH
PF4
C1QB
CCL4
PI3K (complex)
IFNG
CALU
CASP1
P38 MAPK
CCND1
CXCR3
HLA-DR
DSG3
IL22
ITIH2
CCL22
GPX3
CD86
Hla-abc
IRF2
SRGN
TXNRD1
IL12 (family)
FCGBP
Network legends
A
B
Activation, causation,expression,localization, membership, modification, molecular cleavage, phosphorylation, protein-DNA interactions, protein-RNA interactions, regulation of binding transcription
Direct interaction
Indirect interaction
Interaction between proteins and their corresponding CP
Group/Complex
Cytokine/Growth Factor
Enzyme
G-protein Coupled Receptor
Kinase
Transcription Regulator
Transmembrane Receptor
Transporter
Peptidase
Other
MDD upregulated
BD upregulated
0.2
Fold-change
2

Figure 9. The top protein network and associated canonical pathways generated by IPA for the nine selected features (proteins). Seven [DSG3, FBLN1, FCGBP, C1QB, GPX3, ITIH2, and PF4 (PLF4)] of the 9 selected features (proteins) were included in the top protein network. Direct and indirect interactions are represented by solid and dashed lines, respectively. Shapes signify the molecular classes of proteins defined in the legend. Canonical pathways associated with proteins in the network are represented by dotted lines (light pink). Differences in protein expression of the 7 features between MDD and BD are represented by fold-change. MDD, major depressive disorder; BD, bipolar disorder; CP, canonical pathway; IPA, Ingenuity Pathway Analysis.

DISCUSSION

In this study, I developed a model for discriminating MDD from BD using an MRM-MS-based quantitative targeted proteomics approach. Our model performed well in both the training and test sets. There was no difference in demographics between MDD and BD, including gender, age, BMI, smoking, exercise, alcohol, blood collection time, and fasting, all of which are recommended in biomarker studies [90-92]. However, because the overall symptom severity and medication use differed, I also tested our model on patients without current hypomanic/manic/mixed symptoms and those who were drug-free, which demonstrated that its performance did not decrease. Furthermore, by ANCOVA, the features remained related to mood disorder type when significant covariates of symptom severity and medication use were controlled for. Thus, our model has potential clinical applicability and implications, enabling objective discrimination between MDD and BD patients. In addition to relying on subjective clinical interviews, the model could serve as a reference during decision-making regarding the diagnosis or treatment of patients. Furthermore, the model has potential in differentiating patient groups and HCs.

In light of our high-dimensional data, which were drawn from a small sample relative to the number of features [144], the reproducibility of the performance of our model could be limited by overfitting. Overfitting can occur when the model selects noise, as well as important signals in the data, resulting in fluctuations in model selection—ie, when no single model is supported absolutely by the data [145]. I considered the effects of overfitting when developing a generalizable model. Consequently, I demonstrated that feature extraction and model averaging, based on LASSO regression with repeated ten-fold crossvalidation, yielded a generalizable model, taking into account the fluctuation of the model selection. Recently, the applicability of these methods was demonstrated in previous studies of psychiatric disorders [100, 105, 112].

I identified 9 features as being important targets of MDD and BD. The level of DSG3 differed significantly between all 3 groups, increasing in MDD and declining in BD compared with HCs. DSG3 is a calcium-dependent adhesion protein in epithelial cells [146], and its association with mood disorders is novel. However, in animal models, other types of DSG proteins are known to be expressed in the corpus callosum, and there is evidence that subtypes of DSG are upregulated in oligodendrocytes after chronic stress exposure [147]. Further study is needed to determine its function in mood disorders.

The level of PLF4 was higher, and those of C1QB and FBLN were decreased in MDD versus BD and HCs. PLF4 is released from platelet alpha granules during platelet activation and is known to be upregulated in the plasma of MDD patients with coronary artery diseases versus those with coronary artery diseases without depression [148, 149]. In addition, considering that serotonin is the main monoamine considered in MDD and given that platelet serotonin receptors are prone to increase in depression [150], PLF4 could be upregulated specifically in mood disorders, especially MDD. C1Q initiates the classical complement cascade and is essential for synaptic elimination [151]. The level of C1Q in peripheral blood was reported to be elevated in MDD [152] and BD [153] versus HC. Regarding C1QB, polymorphisms in the C1QB are associated with SZ in the Armenian population [154]. In addition, C1QC, which was included in the list of candidate features but was not selected in our model, was upregulated in peripheral blood in MDD [155] and manic BD [156]. C1QB was probably chosen over C1QC by LASSO due to multi-collinearity of both proteins that share biological functions and structure [157-159], and as the discrimination ability of C1QB was higher than C1QC. This implies that combining other potential proteins by manual selection of previously determined candidate features might need consideration. FBLN1 is an extracellular glycoprotein that is involved in cell adhesion and motility along fibers in the extracellular matrix [160]. The SNP located in FBLN1 has been associated with hyperthymic temperament in a GWAS of BD [161], which is also a risk factor for BD itself.

FCGBP levels were higher, and GPX3 was decreased in BD, compared with MDD and HC. GPX3 is a selenoprotein with peroxidase activity [162]. Its level in the CSF is increased in MDD, BD, and SZ compared with HC, but there is no significant difference between MDD and BD [163]. It is also differentially expressed in bipolar disorder brains versus HCs, based on the Stanley Medical Research Institute Online Genomics database [164]. Finally, a variant of this gene correlates with adolescent BD compared with HCs [165]. FCGBP is a well-known protein that is associated with the maintenance of mucosal structures [166]. In a transcriptome sequencing study of postmortem dorsal striatum brains, FCGBP was upregulated in BD versus HCs [167].

The levels of FHR3 (upregulated in BD), IGHM (upregulated in BD), and ITIH2 (upregulated in MDD) were significant only between MDD and BD and, as such, were only associated with the differentiation between MDD and BD. IGHM is a membrane-bound or secreted glycoprotein that is produced by B lymphocytes, which are associated with primary defense mechanisms [168]. IGHM was the only protein to show differential expression in blood between MDD and BD consistently from previous studies. However, its expression pattern was the opposite from that in our study [94]. In addition, its levels in lymphoblastoid cell lines are upregulated in schizophrenia (SZ) compared with HCs [169]. ITIH proteins are serine protease inhibitors and have significant functions as anti-inflammatory molecules [170]. The serum level of ITIH2 is decreased in MDD versus HCs [171]. However, there are several reports on the association of ITIH1 and ITIH4 with mood disorders [155, 156, 171]. FHR3 is related to complement factor H, which is a major alternative complement pathway regulator [172], and has been linked to SZ [173].

Although previous studies[13-16] have proposed potential biomarker candidates (ie, C3, C4BPA, CFI, B2RAN2, ENG, RAB7A, ROCK2, XPO7, PDGF-BB, and TSP-1) to discriminate MDD and BD, their clinical significance and relevance must be validated in a large cohort. Specifically, considering that these candidates were discovered based on proteomic profiling studies that used different techniques (MALDI-TOF/TOF MS, LC-MS/MS, and immunoassay), their statistical significance and expression patterns between MDD and BD must be validated in large cohort studies, as do the 9 plasma proteins in our study. Although the 7 candidates (C3, C4BPA, CFI, B2RAN2, ENG, RAB7A, ROCK2, and XPO7) differed significantly between MDD and BD, their accuracies were not provided [94-96]. Conversely, 2 combined candidates (PDGF-BB and TSP-1) in the immunoassay-based panel had an accuracy of 67% in discriminating MDD and BD [97]. However, the 9 proteins of the previous studies were not included in our MRM-MS analysis of individual samples, having been excluded during the selection of MS-detectable targets, failing to satisfy the criteria, despite being integrated into our initial target list for MDD and BD. Thus, their discriminatory power could not be validated in our study.

Compared with candidate biomarkers that have been proposed in previous studies [94-97], there was only 1 replicated protein, IGHM, and its expression pattern was opposite to that in a previous study [94]. These discrepancies are likely due to different study designs with disparate techniques to quantify proteins and heterogeneous inclusion and exclusion criteria for patients, which might have led to heterogeneous groups. Specifically, regarding study design and techniques to quantify proteins, Chen et al. (2015) [94] compared proteomic profiles between MDD and BD using pooled samples for each group—using pooled samples of 15 MDD and 15 BD-II plasma samples each. Conversely, in our study, 270 individual samples (90 MDD, 90 BD, and 90 HCs) were used to quantify the targets. The inconsistency of alterations in IGHM between MDD and BD could have resulted from the disparate number of samples and characteristics of pooled and individual samples. In addition, Chen et al. (2015) [94] performed proteomic analyses, based on 2-dimensional gel electrophoresis (2-DE), coupled with MALDI-TOF/TOF MS. Specifically, after the proteins were separated by 2DE, which was repeated in triplicate, 25 distinct spots were selected using PDQuest, and 25 DEPs were identified by MALDI-TOF/TOF MS. However, in our study, no separation of proteins was performed, and our targets were quantified directly by MRM-MS in a single run. Subsequently, the MRM-MS data were processed in Skyline to yield quantitative levels of the targets. I propose that the various MS-based quantitation methods for proteomic profiling or targeted proteomics and the inconsistency in data-processing methods between tools influenced the opposite expression pattern of IGHM. Furthermore, with regard to the heterogeneous criteria for patients, drug-free MDD and BD patients were selected in Chen et al. (2015) [94]. But, in our study, most patients were medicated. Although IGHM was only related to the differential diagnosis by ANCOVA, it correlated with the use of MS and AD use in the univariate analysis. Thus, medication might have affected the inconsistency of the IGHM expression patterns. Disease subtype and episodes should also be considered. Although Chen et al. (2015) was based solely on BD-II depressive patients, our study was based on BD-I, BD-II, BD-NOS, and various episodes. Thus, the heterogeneity of BD in our study could have influenced these results.

In summary, several proteins (ie, C1QB, DSG3 and FHR3) were discovered to differentiate MDD from BD, in addition to known proteins that were identified in relation to MDD or BD. However, due to the heterogeneity of these diseases, it is unlikely that a single plasma protein differentiates MDD from BD and explains the disorder [90-92]. In addition, several proteins were associated with other psychiatric disorders, because different psychiatric disorders share genetic susceptibilities [104] and biological pathways [91]. A recent study reported that the same proteins were identified as blood biomarkers in 1 or more psychiatric disorders between MDD, BD, and SZ [92]. Thus, it is likely that the combination of certain key proteins differentiates MDD from BD.

Our bioinformatics analysis of the 9 features revealed the following. The first network, which comprised interactions between our features and other molecules, was associated with several biological functions (cell-to-cell signaling and interaction, hematological system development and function, and immune cell trafficking) and psychiatric diseases (psychological disorder). With regard to canonical pathways that were associated with the network, previous research demonstrated that LXR/RXR activation, FXR/RXR activation, and acute phase response signaling were enriched in MDD and BD [92]. In particular, LXR/RXR activation was significant in a previous study that compared MDD and BD [96]. Immunity and inflammation-related pathways—NF-kB signaling and neuro-inflammation signaling pathways—have also been examined in MDD and BD [113, 114]. Although the 25 differentially expressed proteins (DEPs) in Chen et al. (2015) [94] and the 14 DEPs in Rhee et al. (2020) [96] differed except for IGHM, the analysis of biological functions in previous studies also revealed that immune and inflammatory pathways were related to the DEPs. Notably, there were other pathways that lacked known associations with discriminatory proteins between MDD and BD, perhaps revealing collateral evidence for the pathogenesis of MDD and BD.

Several studies have reported that neuro-related signaling pathways—CREB signaling in neurons, synaptic long-term depression, and synaptic long-term potentiation—influence the pathogenesis of MDD and BD [175-179]. Furthermore, other studies reported that oxidative and nitrosative stress-related pathways are involved in the pathogenesis of mood disorders and might be biological mechanisms for therapeutic strategies of depressive disorders [180-182]. In conclusion, the activation or inhibition status of neuro-related, oxidative and nitrosative stress-related, and immunity/inflammation-related pathways, which are well known in mood disorders, might differ between MDD and BD, leading to disparate protein expression patterns in plasma. However, because these pathways were deduced from peripheral blood proteins, not the central nervous system, caution is necessary when interpreting the results.

There were several limitations of our study. First, the sample size was a major limitation, due to the difficulty in collecting the appropriate patient and healthy control samples. In addition, our model performance should be evaluated using an independent validation set in future studies—I examined its performance in an individual test set. Second, there could have been significant confounders that influenced our results. We classified medication use broadly, and specific dosages/durations of medication were not controlled for. Although I performed several analyses to determine whether certain covariates influenced the results, other covariates might have affected the discriminatory ability. Third, because it was a cross-sectional study, the interpretation for causality is limited. Longitudinal studies would enable observations of the diagnostic conversion from MDD to BD, and serial measures of proteins in an individual would allow differentiation between proteins that are associated with the trait and state of both disorders.

Lastly, although I tried to select important proteins regarding MDD and BD during the initial target integration and determined quantifiable protein targets based on our criteria, other targets might have been overlooked.

However, our study has several strengths. It was the first report to differentiate MDD from BD by MRM-MS. Through our high-throughput MRM-MS method, multiple protein targets in MDD and BD were able to be quantified stably and reproducibly. In addition, these potential targets could be analyzed simultaneously in 270 individual samples, which was larger than in other proteomic studies [94-97]. Second, to decrease overfitting and develop a generalizable model, feature extraction and model averaging were used, resulting in good performance (AUC > 0.8), which was similar in the training and test sets. Third, the features in our model had biologically important relationships with MDD and BD. Finally, when our model was applied to several conditions, its discriminatory performance did not decrease. I propose that our model has applicability in data from various conditions.

In conclusion, I examined the viability of discriminating MDD and BD patients using a targeted proteomic approach (MRM-MS). I developed a 9-feature generalizable model for distinguishing MDD from BD using feature extraction and model averaging. Our results suggest that these disorders can be differentiated using our model. Furthermore, I propose that the 9 plasma proteins that were used as features have biologically important associations with these disorders. Although our model performed well, further studies need to be conducted in a large cohort that consists of drug-free MDD and BD patients to verify whether its performance can be replicated. This proof-of-concept study also demonstrates the potential of the proteomic-based model for discriminating mood disorders.

GENERAL CONCLUSION

Liquid chromatography (LC)-mass spectrometry (MS)-based proteomics has evolved tremendously over the past few years. Therefore, expectations for discovery and development of biomarkers in specific diseases and disorders using proteomics have expanded. Recently high-throughput LC-MS-based proteomic profiling and targeted proteomics have been implemented in the research of various diseases and disorders including breast cancer and mood disorders for discovery and development of protein biomarkers. Nonetheless, efficient biomarkers have not yet been discovered and developed for various lethal diseases and disorders. First, sample preparation methods for the LC-MS-based proteomic analysis of clinical and pathological specimens have not been established, robustly. In some cases, protein biomarker researches have not been performed at all because sample preparation methods of particular specimens are exceedingly difficult. Second, the untargeted proteomic method based on LC-high resolution-MS for increase of the depth of protein identification and targeted proteomic method based on LC-MRM-MS for stable multiplexed protein quantification have not been actively exploited. Lastly, collaborative studies between proteomics and clinics have been partially successful. In order to challenge these limitations, I have established proteomic analytical strategies for clinical and pathological specimens using appropriate LC-MS in protein biomarker discovery and development.

In this study, biomarkers of specific diseases and disorders were discovered and developed by employing state-of-the-art untargeted and targeted proteomics technologies to respective clinical and pathological samples in order to overcome these limitations. FFPE tissue slides and blood plasma were used for untargeted and targeted proteomic analysis of distant metastatic breast cancer and mood disorders, respectively. Particularly, proteomic analysis of central nervous system-related specimens such as human brain tissue or cerebrospinal fluid is the best way to develop protein biomarkers reflecting the molecular mechanisms of mood disorders. However, there are restrictions on the academic use of these specimens (especially brain) in Korea, making it strenuous to use in mood disorder researches. Therefore, as an alternative, the peripheral blood plasma samples, which are non-invasive and accessible in a number of patients, were used in this study. Taking advantage of the blood plasma, a large number of samples of patients and healthy controls were analyzed, and diagnostic protein biomarkers for discriminating mood disorders and biological interactions between these biomarkers and both disorders were developed and investigated, respectively.

The methods for protein biomarker discovery and development of these clinical and pathologic specimens have been reported to be particularly necessary or have not yet been established. Therefore, I have established proper analytical procedures relying on the type of specimens and analyzed them using untargeted and targeted LC-MS-based proteomic technologies. As a result, in Chapter I, a novel protein candidate biomarker was discovered for the prediction of distant metastatic breast cancer. Through investigation of the expression level of TUBB2A in stage1-3 breast cancer patients, clinicians are likely to predict distant metastatic breast cancer by monitoring the progression of breast cancer from stage1-3 to stage. However, for clinical use of the discovered biomarkers, substantial validation in larger samples is required by using non-MS platforms such as immunohistochemistry and ELISA. Whether TUBB2A expression is consistent between different quantitation platforms including LC-MS should be examined. In-depth functional analyses for TUBB2A are required to reveal molecular mechanism of regulating distant metastatic breast cancer. Nonetheless, the constructed distant metastatic breast cancer FFPE proteome, which is the largest data, will be advantageous as pivotal proteomic data for other breast cancer researchers. In Chapter II, nine novel plasma protein biomarkers for discriminating mood disorders were developed, and 9-plasma protein-based diagnostic model was also constructed. These findings demonstrated the potential of protein biomarkers in diagnosis of mood disorders. However, there were several limitations. The sample size is a major limitation due to the difficulty in collecting appropriate patients and HC samples. Secondly, there were potential confounders that could have affected the performance of the model. Specifically, we categorized medication use dichotomously, and specific dosages and durations of medication were not controlled for. This was not enough to restrict medication effects on expression levels of plasma proteins, thus demanding analysis of plasma samples of first-episode and drug-naive patients. Third, the interpretation of causality is limited because the study was cross-sectional. Thus, longitudinal studies are required to observe diagnostic alterations of MDD and BD. Fourth, additional experimental validation of the proteins should be performed to examine whether the plasma proteins are correlated with the central nervous system. Lastly, the potential of our LC-MS-based proteomic approaches should be validated using conventional immunoassays such as ELISA in order to evaluate consistency between different analytical platforms.

Through a series of LC-MS-based proteomic analyses, these studies emphasize the role of proteomics in the translational medicine. In addition, these analyses demonstrated that LC-MS-based proteomics remains one of the most robust and powerful research fields to discover and develop biomarkers for specific diseases and disorders. However, obviously, there remains still a tremendous demand for protein biomarkers available for various diseases and disorders. To resolve this issue, proteomic methods and technologies appropriate for clinical and pathologic specimens should be developed and applied via ceaseless proteomic research. Providing that robust analytical procedures of proteomic technique for almost all clinical and pathological specimens are established, researchers will be aided in discovering and developing protein biomarkers.

Although I focused on proteomics technologies and protein biomarkers in this dissertation, integration of multi-omics data including proteomics, genomics, transcriptomics, and metabolomics is expected to provide useful insight into biomarker discovery and development for specific diseases and disorders. Provided that multi-omics techniques are optimized in clinical practice and integration of multi-omics and clinical data is achieved, it will contribute to facilitating discovery and development of robust biomarkers. Therefore, the multi-omics approach should be considered in further study to rationalize the results of biomarkers in this study.

REFERENCES

1. Yanovich G, Agmon H, Harel M, Sonnenblick A, Peretz T, Geiger T: Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification. *Cancer Res* 2018, 78:6001-6010.

2. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A: Breast cancer statistics, 2017, racial disparity in mortality by state. *CA Cancer J Clin* 2017, 67:439-448.

3. Anastasiadi Z, Lianos GD, Ignatiadou E, Harissis HV, Mitsis M: Breast cancer in young women: an overview. *Updates Surg* 2017, 69:313-317.

4. Fredholm H, Eaker S, Frisell J, Holmberg L, Fredriksson I, Lindman H: Breast cancer in young women: poor survival despite intensive treatment. *PLoS One* 2009, 4:e7695.

5. Hess KR, Varadhachary GR, Taylor SH, Wei W, Raber MN, Lenzi R, Abbruzzese JL: Metastatic patterns in adenocarcinoma. *Cancer* 2006, 106:1624-1633.

6. Chang J, Clark GM, Allred DC, Mohsin S, Chamness G, Elledge RM: Survival of patients with metastatic breast carcinoma: importance of prognostic markers of the primary tumor. *Cancer* 2003, 97:545-553.

7. Coleman RE, Rubens RD: The clinical course of bone metastases from breast cancer. *Br J Cancer* 1987, 55:61-66.

8. Lobbezoo DJ, van Kampen RJ, Voogd AC, Dercksen MW, van den Berkmortel F, Smilde TJ, van de Wouw AJ, Peters FP, van Riel JM, Peters NA, et al: Prognosis of metastatic breast cancer: are there differences between patients with de novo and recurrent metastatic breast cancer? *Br J Cancer* 2015, 112:1445-1451.

9. Horton J: Follow-up of breast cancer patients. *Cancer* 1984, 53:790-797.

10. Kennecke H, Yerushalmi R, Woods R, Cheang MC, Voduc D, Speers CH, Nielsen TO, Gelmon K: Metastatic behavior of breast cancer subtypes. *J Clin Oncol* 2010, 28:3271-3277.

11. Chia S, Norris B, Speers C, Cheang M, Gilks B, Gown AM, Huntsman D, Olivotto IA, Nielsen TO, Gelmon K: Human epidermal growth factor receptor 2 overexpression as a prognostic factor in a large tissue microarray series of node-negative breast cancers. *J Clin Oncol* 2008, 26:5697-5704.

12. Alanko A, Heinonen E, Scheinin T, Tolppanen EM, Vihko R: Significance of estrogen and progesterone receptors, disease-free interval, and site of first metastasis on survival of breast cancer patients. *Cancer* 1985, 56:1696-1700.

13. Kate RJ, Nadig R: Stage-specific predictive models for breast cancer survivability. *Int J Med Inform* 2017, 97:304-311.

14. Schnitt SJ: Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Mod Pathol* 2010, 23 Suppl 2:S60-64.

15. Minn AJ, Gupta GP, Padua D, Bos P, Nguyen DX, Nuyten D, Kreike B, Zhang Y, Wang Y, Ishwaran H, et al: Lung metastasis genes couple breast tumor size and metastatic spread. *Proc Natl Acad Sci U S A* 2007, 104:6740-6745.

16. Bos PD, Zhang XH, Nadal C, Shu W, Gomis RR, Nguyen DX, Minn AJ, van de Vijver MJ, Gerald WL, Foekens JA, Massague J: Genes that mediate breast cancer metastasis to the brain. *Nature* 2009, 459:1005-1009.

17. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, Viale A, Olshen AB, Gerald WL, Massague J: Genes that mediate breast cancer metastasis to lung. *Nature* 2005, 436:518-524.

18. Kang Y, Siegel PM, Shu W, Drobnjak M, Kakonen SM, Cordon-Cardo C, Guise TA, Massague J: A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* 2003, 3:537-549.

19. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, et al: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005, 365:671-679.

20. Bellahcene A, Bachelier R, Detry C, Lidereau R, Clezardin P, Castronovo V:

Transcriptome analysis reveals an osteoblast-like phenotype for human osteotropic breast cancer cells. *Breast Cancer Res Treat* 2007, 101:135-148.

21. Garcia M, Millat-Carus R, Bertucci F, Finetti P, Birnbaum D, Bidaut G: Interactome-transcriptome integration for predicting distant metastasis in breast cancer. *Bioinformatics* 2012, 28:672-678.

22. Geiger T, Cox J, Mann M: Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet* 2010, 6:e1001090.

23. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al: Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014, 513:382-387.

24. Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Algenas C, Lundeberg J, Mann M, Uhlen M: Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol* 2010, 6:450.

25. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: Global quantification of mammalian gene expression control. *Nature* 2011, 473:337-342.

26. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, Mann M: Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 2011, 7:548.

27. Jezequel P, Guette C, Lasla H, Gouraud W, Boissard A, Guerin-Charbonnel C, Campone M: iTRAQ-Based Quantitative Proteomic Analysis Strengthens Transcriptomic Subtyping of Triple-Negative Breast Cancer Tumors. *Proteomics* 2019, 19:e1800484.

28. Johansson HJ, Socciarelli F, Vacanti NM, Haugen MH, Zhu Y, Siavelis I, Fernandez-Woodbridge A, Aure MR, Sennblad B, Vesterlund M, et al: Breast cancer quantitative proteome and proteogenomic landscape. *Nat Commun* 2019, 10:1600.

29. Bouchal P, Schubert OT, Faktor J, Capkova L, Imrichova H, Zoufalova K, Paralova V, Hrstka R, Liu Y, Ebhardt HA, et al: Breast Cancer Classification Based on Proteotypes

Obtained by SWATH Mass Spectrometry. *Cell Rep* 2019, 28:832-843 e837.

30. Shin J, Dan K, Han D, Kim JW, Kim KK, Koh Y, Shin DY, Hong J, Yoon SS, Park S, et al: Plasma-based protein biomarkers can predict the risk of acute graft-versus-host disease and non-relapse mortality in patients undergoing allogeneic hematopoietic stem cell transplantation. *Blood Cells Mol Dis* 2019, 74:5-12.

31. Kim YS, Han D, Kim J, Kim DW, Kim YM, Mo JH, Choi HG, Park JW, Shin HW: In-Depth, Proteomic Analysis of Nasal Secretions from Patients With Chronic Rhinosinusitis and Nasal Polyps. *Allergy Asthma Immunol Res* 2019, 11:691-708.

32. Duangkumpha K, Stoll T, Phetcharaburanin J, Yongvanit P, Thanan R, Techasen A, Namwat N, Khuntikeo N, Chamadol N, Roytrakul S, et al: Discovery and Qualification of Serum Protein Biomarker Candidates for Cholangiocarcinoma Diagnosis. *J Proteome Res* 2019, 18:3305-3316.

33. Lee H, Kim K, Woo J, Park J, Kim H, Lee KE, Kim H, Kim Y, Moon KC, Kim JY, 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, 17:1788-1802.

34. Jin J, Son M, Kim H, Kim H, Kong SH, Kim HK, Kim Y, Han D: Comparative proteomic analysis of human malignant ascitic fluids for the development of gastric cancer biomarkers. *Clin Biochem* 2018, 56:55-61.

35. Do M, Han D, Wang JI, Kim H, Kwon W, Han Y, Jang JY, Kim Y: Quantitative proteomic analysis of pancreatic cyst fluid proteins associated with malignancy in intraductal papillary mucinous neoplasms. *Clin Proteomics* 2018, 15:17.

36. Park J, Han D, Do M, Woo J, Wang JI, Han Y, Kwon W, Kim SW, Jang JY, Kim Y: Proteome characterization of human pancreatic cyst fluid from intraductal papillary mucinous neoplasm by liquid chromatography/tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2017, 31:1761-1772.

37. Geyer PE, Kulak NA, Pichler G, Holdt LM, Teupser D, Mann M: Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* 2016, 2:185-195.

38. Aebersold R, Mann M: Mass-spectrometric exploration of proteome structure and function. *Nature* 2016, 537:347-355.

39. Choudhary C, Mann M: Decoding signalling networks by mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol* 2010, 11:427-439.

40. Murphy JP, Stepanova E, Everley RA, Paulo JA, Gygi SP: Comprehensive Temporal Protein Dynamics during the Diauxic Shift in Saccharomyces cerevisiae. *Mol Cell Proteomics* 2015, 14:2454-2465.

41. Kim DK, Park J, Han D, Yang J, Kim A, Woo J, Kim Y, Mook-Jung I: Molecular and functional signatures in a novel Alzheimer's disease mouse model assessed by quantitative proteomics. *Mol Neurodegener* 2018, 13:2.

42. Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, Naake T, Gatto L, Viner R, Martinez Arias A, Lilley KS: A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun* 2016, 7:8992.

43. Weekes MP, Tomasec P, Huttlin EL, Fielding CA, Nusinow D, Stanton RJ, Wang EC, Aicheler R, Murrell I, Wilkinson GW, et al: Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell* 2014, 157:1460-1472.

44. Jin MS, Lee H, Woo J, Choi S, Do MS, Kim K, Song MJ, Kim Y, Park IA, Han D, Ryu HS: Integrated Multi-Omic Analyses Support Distinguishing Secretory Carcinoma of the Breast from Basal-Like Triple-Negative Breast Cancer. *Proteomics Clin Appl* 2018, 12:e1700125.

45. Han D, Moon S, Kim Y, Kim J, Jin J, Kim Y: In-depth proteomic analysis of mouse microglia using a combination of FASP and StageTip-based, high pH, reversed-phase fractionation. *Proteomics* 2013, 13:2984-2988.

46. Han D, Jin J, Woo J, Min H, Kim Y: Proteomic analysis of mouse astrocytes and their secretome by a combination of FASP and StageTip-based, high pH, reversed-phase

fractionation. *Proteomics* 2014, 14:1604-1609.

47. Kulak NA, Pichler G, Paron I, Nagaraj N, Mann M: Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* 2014, 11:319-324.

48. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, et al: ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014, 32:223-226.

49. Wisniewski JR, Zougman A, Nagaraj N, Mann M: Universal sample preparation method for proteome analysis. *Nat Methods* 2009, 6:359-362.

50. Brinton LT, Brentnall TA, Smith JA, Kelly KA: Metastatic biomarker discovery through proteomics. *Cancer Genomics Proteomics* 2012, 9:345-355.

51. Yi X, Luk JM, Lee NP, Peng J, Leng X, Guan XY, Lau GK, Beretta L, Fan ST: Association of mortalin (HSPA9) with liver cancer metastasis and prediction for early tumor recurrence. *Mol Cell Proteomics* 2008, 7:315-325.

52. Piao HL, Yuan Y, Wang M, Sun Y, Liang H, Ma L: alpha-catenin acts as a tumour suppressor in E-cadherin-negative basal-like breast cancer by inhibiting NF-kappaB signalling. *Nat Cell Biol* 2014, 16:245-254.

53. Liu R, Lu S, Deng Y, Yang S, He S, Cai J, Qiang F, Chen C, Zhang W, Zhao S, et al: PSMB4 expression associates with epithelial ovarian cancer growth and poor prognosis. *Arch Gynecol Obstet* 2016, 293:1297-1307.

54. Wang H, He Z, Xia L, Zhang W, Xu L, Yue X, Ru X, Xu Y: PSMB4 overexpression enhances the cell growth and viability of breast cancer cells leading to a poor prognosis. *Oncol Rep* 2018, 40:2343-2352.

55. Conacci-Sorrell M, Zhurinsky J, Ben-Ze'ev A: The cadherin-catenin adhesion system in signaling and cancer. *J Clin Invest* 2002, 109:987-991.

56. Leaderer D, Hoffman AE, Zheng T, Fu A, Weidhaas J, Paranjape T, Zhu Y: Genetic and epigenetic association studies suggest a role of microRNA biogenesis gene exportin-5

(XPO5) in breast tumorigenesis. *Int J Mol Epidemiol Genet* 2011, 2:9-18.

57. Diz AP, Truebano M, Skibinski DO: The consequences of sample pooling in proteomics: an empirical study. *Electrophoresis* 2009, 30:2967-2975.

58. Bauer KR, Brown M, Cress RD, Parise CA, Caggiano V: Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry. *Cancer* 2007, 109:1721-1728.

59. Weigelt B, Peterse JL, van 't Veer LJ: Breast cancer metastasis: markers and models. *Nat Rev Cancer* 2005, 5:591-602.

60. Ridley AJ, Schwartz MA, Burridge K, Firtel RA, Ginsberg MH, Borisy G, Parsons JT, Horwitz AR: Cell migration: integrating signals from front to back. *Science* 2003, 302:1704-1709.

61. Altschuler SJ, Angenent SB, Wang Y, Wu LF: On the spontaneous emergence of cell polarity. *Nature* 2008, 454:886-889.

62. Wu J, Mlodzik M: A quest for the mechanism regulating global planar cell polarity of tissues. *Trends Cell Biol* 2009, 19:295-305.

63. Gilmore TD: Introduction to NF-kappaB: players, pathways, perspectives. *Oncogene* 2006, 25:6680-6684.

64. Perkins ND: Integrating cell-signalling pathways with NF-kappaB and IKK function. *Nat Rev Mol Cell Biol* 2007, 8:49-62.

65. Ezumi Y, Uchiyama T, Takayama H: Molecular cloning, genomic structure, chromosomal localization, and alternative splice forms of the platelet collagen receptor glycoprotein VI. *Biochem Biophys Res Commun* 2000, 277:27-36.

66. Jandrot-Perrus M, Busfield S, Lagrue AH, Xiong X, Debili N, Chickering T, Le Couedic JP, Goodearl A, Dussault B, Fraser C, et al: Cloning, characterization, and functional studies of human and mouse glycoprotein VI: a platelet-specific collagen receptor from the immunoglobulin superfamily. *Blood* 2000, 96:1798-1807.

67. Pavon-Eternod M, Gomes S, Geslain R, Dai Q, Rosner MR, Pan T: tRNA over-expression in breast cancer and functional consequences. *Nucleic Acids Res* 2009, 37:7268-7280.

68. Kallergi G, Agelaki S, Kalykaki A, Stournaras C, Mavroudis D, Georgoulias V: Phosphorylated EGFR and PI3K/Akt signaling kinases are expressed in circulating tumor cells of breast cancer patients. *Breast Cancer Res* 2008, 10:R80.

69. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, et al: The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 2019, 47:D442-D450.

70. Perez-Riverol Y, Xu QW, Wang R, Uszkoreit J, Griss J, Sanchez A, Reisinger F, Csordas A, Ternent T, Del-Toro N, et al: PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets. *Mol Cell Proteomics* 2016, 15:305-317.

71. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, Campbell DS, Bernal-Llinares M, Okuda S, Kawano S, et al: The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 2017, 45:D1100-D1106.

72. Mulvihill MM, Benjamin DI, Ji X, Le Scolan E, Louie SM, Shieh A, Green M, Narasimhalu T, Morris PJ, Luo K, Nomura DK: Metabolic profiling reveals PAFAH1B3 as a critical driver of breast cancer pathogenicity. *Chem Biol* 2014, 21:831-840.

73. Dai X, Cheng H, Bai Z, Li J: Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. *J Cancer* 2017, 8:3131-3141.

74. Sommers CL, Byers SW, Thompson EW, Torri JA, Gelmann EP: Differentiation state and invasiveness of human breast cancer cell lines. *Breast Cancer Res Treat* 1994, 31:325-335.

75. Bae SN, Arand G, Azzam H, Pavasant P, Torri J, Frandsen TL, Thompson EW: Molecular

and cellular analysis of basement membrane invasion by human breast cancer cells in Matrigel-based in vitro assays. *Breast Cancer Res Treat* 1993, 24:241-255.

76. Hughes L, Malone C, Chumsri S, Burger AM, McDonnell S: Characterisation of breast cancer cell lines and establishment of a novel isogenic subclone to study migration, invasion and tumourigenicity. *Clin Exp Metastasis* 2008, 25:549-557.

77. Quail DF, Maciel TJ, Rogers K, Postovit LM: A unique 3D in vitro cellular invasion assay. *J Biomol Screen* 2012, 17:1088-1095.

78. Ziperstein MJ, Guzman A, Kaufman LJ: Breast Cancer Cell Line Aggregate Morphology Does Not Predict Invasive Capacity. *PLoS One* 2015, 10:e0139523.

79. Ribeiro AS, Albergaria A, Sousa B, Correia AL, Bracke M, Seruca R, Schmitt FC, Paredes J: Extracellular cleavage and shedding of P-cadherin: a mechanism underlying the invasive behaviour of breast cancer cells. *Oncogene* 2010, 29:392-402.

80. Li J, Wei J, Mei Z, Yin Y, Li Y, Lu M, Jin S: Suppressing role of miR-520a-3p in breast cancer through CCND1 and CD44. *Am J Transl Res* 2017, 9:146-154.

81. Chiu HW, Lin HY, Tseng IJ, Lin YF: OTUD7B upregulation predicts a poor response to paclitaxel in patients with triple-negative breast cancer. *Oncotarget* 2018, 9:553-565.

82. Shin C, Kim Y, Park S, Yoon S, Ko YH, Kim YK, Kim SH, Jeon SW, Han C: Prevalence and Associated Factors of Depression in General Population of Korea: Results from the Korea National Health and Nutrition Examination Survey, 2014. *J Korean Med Sci* 2017, 32(11):1861-1869.

83. Grande I, Berk M, Birmaher B, Vieta E: Bipolar disorder. *Lancet* 2016, 387(10027):1561-1572.

84. Lim D, Lee WK, Park H: Disability-adjusted Life Years (DALYs) for Mental and Substance Use Disorders in the Korean Burden of Disease Study 2012. *J Korean Med Sci* 2016, 31 Suppl 2:S191-S199.

85. Ghaemi SN, Boiman EE, Goodwin FK: Diagnosing bipolar disorder and the effect of antidepressants: a naturalistic study. *J Clin Psychiatry* 2000, 61(10):804-808; quiz 809.

86. Hirschfeld RM, Lewis L, Vornik LA: Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *J Clin Psychiatry* 2003, 64(2):161-174.

87. Angst J, Sellaro R, Stassen HH, Gamma A: Diagnostic conversion from depression to bipolar disorders: results of a long-term prospective study of hospital admissions. *J Affect Disord* 2005, 84(2-3):149-157.

88. Fornaro M, Anastasia A, Novello S, Fusco A, Solmi M, Monaco F, Veronese N, De Berardis D, de Bartolomeis A: Incidence, prevalence and clinical correlates of antidepressant-emergent mania in bipolar depression: a systematic review and meta-analysis. *Bipolar disorders* 2018, 20(3):195-227.

89. Mattes JA: Antidepressant-induced rapid cycling: another perspective. *Ann Clin Psychiatry* 2006, 18(3):195-199.

90. Preece RL, Han SYS, Bahn S: Proteomic approaches to identify blood-based biomarkers for depression and bipolar disorders. *Expert review of proteomics* 2018, 15(4):325-340.

91. Domenici E, Muglia P: The search for peripheral disease markers in psychiatry by genomic and proteomic approaches. *Expert opinion on medical diagnostics* 2007, 1(2):235-251.

92. Comes AL, Papiol S, Mueller T, Geyer PE, Mann M, Schulze TG: Proteomics for blood biomarker exploration of severe mental illness: pitfalls of the past and potential for the future. *Transl Psychiatry* 2018, 8(1):160.

93. Haenisch F, Cooper JD, Reif A, Kittel-Schneider S, Steiner J, Leweke FM, Rothermundt M, van Beveren NJM, Crespo-Facorro B, Niebuhr DW *et al*: Towards a blood-based diagnostic panel for bipolar disorder. *Brain Behav Immun* 2016, 52:49-57.

94. Chen J, Huang C, Song Y, Shi H, Wu D, Yang Y, Rao C, Liao L, Wu Y, Tang J *et al*: Comparative proteomic analysis of plasma from bipolar depression and depressive disorder: identification of proteins associated with immune regulatory. *Protein Cell* 2015, 6(12):908-911.

95. Ren J, Zhao G, Sun X, Liu H, Jiang P, Chen J, Wu Z, Peng D, Fang Y, Zhang C: Identification of plasma biomarkers for distinguishing bipolar depression from major depressive disorder by iTRAQ-coupled LC-MS/MS and bioinformatics analysis. *Psychoneuroendocrinology* 2017, 86:17-24.

96. Rhee SJ, Han D, Lee Y, Kim H, Lee J, Lee K, Shin H, Kim H, Lee TY, Kim M *et al*: Comparison of serum protein profiles between major depressive disorder and bipolar disorder. *BMC Psychiatry* 2020, 20(1):145.

97. Kittel-Schneider S, Hahn T, Haenisch F, McNeill R, Reif A, Bahn S: Proteomic Profiling as a Diagnostic Biomarker for Discriminating Between Bipolar and Unipolar Depression. *Front Psychiatry* 2020, 11:189.

98. Aebersold R, Mann M: Mass-spectrometric exploration of proteome structure and function. *Nature* 2016, 537(7620):347-355.

99. Choudhary C, Mann M: Decoding signalling networks by mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol* 2010, 11(6):427-439.

100. Han SYS, Cooper JD, Ozcan S, Rustogi N, Penninx B, Bahn S: Integrating proteomic, sociodemographic and clinical data to predict future depression diagnosis in subthreshold symptomatic individuals. *Transl Psychiatry* 2019, 9(1):277.

101. Kim EY, Lee MY, Kim SH, Ha K, Kim KP, Ahn YM: Diagnosis of major depressive disorder by combining multimodal information from heart rate dynamics and serum proteomics using machine-learning algorithm. *Prog Neuropsychopharmacol Biol Psychiatry* 2017, 76:65-71.

102. Whiteaker JR, Lin C, Kennedy J, Hou L, Trute M, Sokal I, Yan P, Schoenherr RM, Zhao L, Voytovich UJ *et al*: A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nat Biotechnol* 2011, 29(7):625-634.

103. Frantzi M, Bhat A, Latosinska A: Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clin Transl Med* 2014, 3(1):7.

104. Kennedy JJ, Abbatiello SE, Kim K, Yan P, Whiteaker JR, Lin C, Kim JS, Zhang Y, Wang X, Ivey RG *et al*: Demonstrating the feasibility of large-scale development of standardized assays to quantify human proteins. *Nat Methods* 2014, 11(2):149-155.

105. Cooper JD, Han SYS, Tomasik J, Ozcan S, Rustogi N, van Beveren NJM, Leweke FM, Bahn S: Multimodel inference for biomarker development: an application to schizophrenia. *Transl Psychiatry* 2019, 9(1):83.

106. Kim Y, Kang M, Han D, Kim H, Lee K, Kim SW, Kim Y, Park T, Jang JY, Kim Y: Biomarker Development for Intraductal Papillary Mucinous Neoplasms Using Multiple Reaction Monitoring Mass Spectrometry. *J Proteome Res* 2016, 15(1):100-113.

107. Kim H, Sohn A, Yeo I, Yu SJ, Yoon JH, Kim Y: Clinical Assay for AFP-L3 by Using Multiple Reaction Monitoring-Mass Spectrometry for Diagnosing Hepatocellular Carcinoma. *Clin Chem* 2018, 64(8):1230-1238.

108. Do M, Kim H, Yeo I, Lee J, Park IA, Ryu HS, Kim Y: Clinical Application of Multiple Reaction Monitoring-Mass Spectrometry to Human Epidermal Growth Factor Receptor 2 Measurements as a Potential Diagnostic Tool for Breast Cancer Therapy. *Clin Chem* 2020, 66(10):1339-1348.

109. Yu SJ, Kim H, Min H, Sohn A, Cho YY, Yoo JJ, Lee DH, Cho EJ, Lee JH, Gim J *et al*: Targeted Proteomics Predicts a Sustained Complete-Response after Transarterial Chemoembolization and Clinical Outcomes in Patients with Hepatocellular Carcinoma: A Prospective Cohort Study. *J Proteome Res* 2017, 16(3):1239-1248.

110. Kim H, Yu SJ, Yeo I, Cho YY, Lee DH, Cho Y, Cho EJ, Lee JH, Kim YJ, Lee S *et al*: Prediction of Response to Sorafenib in Hepatocellular Carcinoma: A Putative Marker Panel by Multiple Reaction Monitoring-Mass Spectrometry (MRM-MS). *Mol Cell Proteomics* 2017, 16(7):1312-1323.

111. Kim H, Park J, Kim Y, Sohn A, Yeo I, Jong Yu S, Yoon JH, Park T, Kim Y: Serum fibronectin distinguishes the early stages of hepatocellular carcinoma. *Sci Rep* 2017, 7(1):9449.

112. Han SYS, Tomasik J, Rustogi N, Lago SG, Barton-Owen G, Eljasz P, Cooper JD, Ozcan S, Olmert T, Farrag LP *et al*: Diagnostic prediction model development using data from dried blood spot proteomics and a digital mental health assessment to identify major depressive disorder among individuals presenting with low mood. *Brain Behav Immun* 2020, 90:184-195.

113. Jabbi M, Arasappan D, Eickhoff SB, Strakowski SM, Nemeroff CB, Hofmann HA: Neuro-transcriptomic signatures for mood disorder morbidity and suicide mortality. *J Psychiatr Res* 2020, 127:62-74.

114. Najjar S, Pearlman DM, Alper K, Najjar A, Devinsky O: Neuroinflammation and psychiatric illness. *J Neuroinflammation* 2013, 10:43.

115. Garay-Baquero DJ, White CH, Walker NF, Tebruegge M, Schiff HF, Ugarte-Gil C, Morris-Jones S, Marshall BG, Manousopoulou A, Adamson J *et al*: Comprehensive plasma proteomic profiling reveals biomarkers for active tuberculosis. *JCI Insight* 2020, 5(18).

116. Vora N, Kalagiri R, Mallett LH, Oh JH, Wajid U, Munir S, Colon N, Raju VN, Beeram MR, Uddin MN: Proteomics and Metabolomics in Pregnancy-An Overview. *Obstet Gynecol Surv* 2019, 74(2):111-125.

117. Kim Y, Kang UB, Kim S, Lee HB, Moon HG, Han W, Noh DY: A Validation Study of a Multiple Reaction Monitoring-Based Proteomic Assay to Diagnose Breast Cancer. *J Breast Cancer* 2019, 22(4):579-586.

118. Dong W, Qiu C, Gong D, Jiang X, Liu W, Liu W, Zhang L, Zhang W: Proteomics and bioinformatics approaches for the identification of plasma biomarkers to detect Parkinson's disease. *Exp Ther Med* 2019, 18(4):2833-2842.

119. Ryan KM, Glaviano A, O'Donovan SM, Kolshus E, Dunne R, Kavanagh A, Jelovac A, Noone M, Tucker GM, Dunn MJ *et al*: Electroconvulsive therapy modulates plasma pigment epithelium-derived factor in depression: a proteomics study. *Transl Psychiatry* 2017, 7(3):e1073.

120. Noorbakhsh F, Aminian A, Power C: Application of "Omics" Technologies for Diagnosis and Pathogenesis of Neurological Infections. *Curr Neurol Neurosci Rep* 2015, 15(9):58.

121. Jayanthi S, Buie S, Moore S, Herning RI, Better W, Wilson NM, Contoreggi C, Cadet JL: Heavy marijuana users show increased serum apolipoprotein C-III levels: evidence from proteomic analyses. *Mol Psychiatry* 2010, 15(1):101-112.

122. Organization WH: Global Recommendations on Physical Activity for Health: World Health Organization; 2010.

123. Hafkenscheid A: Psychometric evaluation of a standardized and expanded Brief Psychiatric Rating Scale. *Acta psychiatrica Scandinavica* 1991, 84(3):294-300.

124. Young RC, Biggs JT, Ziegler VE, Meyer DA: A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry : the journal of mental science* 1978, 133:429-435.

125. Montgomery SA, Asberg M: A new depression scale designed to be sensitive to change. *The British journal of psychiatry : the journal of mental science* 1979, 134:382-389.

126. Hamilton M: The assessment of anxiety states by rating. *The British journal of medical psychology* 1959, 32(1):50-55.

127. Sussman N, Mullen J, Paulsson B, Vågerö M: Rates of remission/euthymia with quetiapine in combination with lithium/divalproex for the treatment of acute mania. *J Affect Disord* 2007, 100 Suppl 1:S55-63.

128. Alsaif M, Guest PC, Schwarz E, Reif A, Kittel-Schneider S, Spain M, Rahmoune H, Bahn S: Analysis of serum and plasma identifies differences in molecular coverage, measurement variability, and candidate biomarker selection. *Proteomics Clin Appl* 2012, 6(5-6):297-303.

129. Bot M, Chan MK, Jansen R, Lamers F, Vogelzangs N, Steiner J, Leweke FM, Rothermundt M, Cooper J, Bahn S *et al*: Serum proteomic profiling of major depressive disorder. *Transl Psychiatry* 2015, 5:e599.

130. Frye MA, Nassan M, Jenkins GD, Kung S, Veldic M, Palmer BA, Feeder SE, Tye SJ,

Choi DS, Biernacka JM: Feasibility of investigating differential proteomic expression in depression: implications for biomarker development in mood disorders. *Transl Psychiatry* 2015, 5:e689.

131. Haenisch F, Alsaif M, Guest PC, Rahmoune H, Dickerson F, Yolken R, Bahn S: Multiplex immunoassay analysis of plasma shows prominent upregulation of growth factor activity pathways linked to GSK3beta signaling in bipolar patients. *J Affect Disord* 2014, 156:139-143.

132. Herberth M, Koethe D, Levin Y, Schwarz E, Krzyszton ND, Schoeffmann S, Ruh H, Rahmoune H, Kranaster L, Schoenborn T *et al*: Peripheral profiling analysis for bipolar disorder reveals markers associated with reduced cell survival. *Proteomics* 2011, 11(1):94-105.

133. Lee MY, Kim EY, Kim SH, Cho KC, Ha K, Kim KP, Ahn YM: Discovery of serum protein biomarkers in drug-free patients with major depressive disorder. *Prog Neuropsychopharmacol Biol Psychiatry* 2016, 69:60-68.

134. Stelzhammer V, Haenisch F, Chan MK, Cooper JD, Steiner J, Steeb H, Martins-de-Souza D, Rahmoune H, Guest PC, Bahn S: Proteomic changes in serum of first onset, antidepressant drug-naive major depression patients. *Int J Neuropsychopharmacol* 2014, 17(10):1599-1608.

135. Xu HB, Zhang RF, Luo D, Zhou Y, Wang Y, Fang L, Li WJ, Mu J, Zhang L, Zhang Y *et al*: Comparative proteomic analysis of plasma from major depressive patients: identification of proteins associated with lipid metabolism and immunoregulation. *Int J Neuropsychopharmacol* 2012, 15(10):1413-1425.

136. Yuan N, Chen Y, Xia Y, Dai J, Liu C: Inflammation-related biomarkers in major psychiatric disorders: a cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. *Transl Psychiatry* 2019, 9(1):233.

137. Frye MA, Nassan M, Jenkins GD, Kung S, Veldic M, Palmer BA, Feeder SE, Tye SJ, Choi DS, Biernacka JM: Feasibility of investigating differential proteomic expression in

depression: implications for biomarker development in mood disorders. *Translational psychiatry* 2015, 5(12):e689.

138. Abbatiello SE, Mani DR, Keshishian H, Carr SA: Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. *Clin Chem* 2010, 56(2):291-305.

139. Friedman J, Hastie T, Tibshirani R: Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010, 33(1):1-22.

140. Sing T, Sander O, Beerenwinkel N, Lengauer T: ROCR: visualizing classifier performance in R. *Bioinformatics* 2005, 21(20):3940-3941.

141. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982, 143(1):29-36.

142. Youden WJ: Index for rating diagnostic tests. *Cancer* 1950, 3(1):32-35.

143. Kramer A, Green J, Pollard J, Jr., Tugendreich S: Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* 2014, 30(4):523-530.

144. Rutledge RB, Chekroud AM, Huys QJ: Machine learning and big data in psychiatry: toward clinical applications. *Curr Opin Neurobiol* 2019, 55:152-159.

145. Hawkins DM: The problem of overfitting. *J Chem Inf Comput Sci* 2004, 44(1):1-12.

146. Tsang SM, Liu L, Teh MT, Wheeler A, Grose R, Hart IR, Garrod DR, Fortune F, Wan H: Desmoglein 3, via an interaction with E-cadherin, is associated with activation of Src. *PloS one* 2010, 5(12):e14211.

147. Miyata S, Yoshikawa K, Taniguchi M, Ishikawa T, Tanaka T, Shimizu S, Tohyama M: Sgk1 regulates desmoglein 1 expression levels in oligodendrocytes in the mouse corpus callosum after chronic stress exposure. *Biochemical and biophysical research communications* 2015, 464(1):76-82.

148. Laghrissi-Thode F, Wagner WR, Pollock BG, Johnson PC, Finkel MS: Elevated platelet factor 4 and beta-thromboglobulin plasma levels in depressed patients with ischemic heart disease. *Biological psychiatry* 1997, 42(4):290-295.

149. Kuijpers PM, Hamulyak K, Strik JJ, Wellens HJ, Honig A: Beta-thromboglobulin and platelet factor 4 levels in post-myocardial infarction patients with major depression. *Psychiatry research* 2002, 109(2):207-210.

150. Mendelson SD: The current status of the platelet 5-HT(2A) receptor in depression. *J Affect Disord* 2000, 57(1-3):13-24.

151. Stevens B, Allen NJ, Vazquez LE, Howell GR, Christopherson KS, Nouri N, Micheva KD, Mehalow AK, Huberman AD, Stafford B *et al*: The classical complement cascade mediates CNS synapse elimination. *Cell* 2007, 131(6):1164-1178.

152. Yao Q, Li Y: Increased serum levels of complement C1q in major depressive disorder. *Journal of psychosomatic research* 2020, 133:110105.

153. Akcan U, Karabulut S, İsmail Küçükali C, Çakır S, Tüzün E: Bipolar disorder patients display reduced serum complement levels and elevated peripheral blood complement expression levels. *Acta neuropsychiatrica* 2018, 30(2):70-78.

154. Zakharyan R, Khoyetsyan A, Arakelyan A, Boyajyan A, Gevorgyan A, Stahelova A, Mrazek F, Petrek M: Association of C1QB gene polymorphism with schizophrenia in Armenian population. *BMC medical genetics* 2011, 12:126.

155. Lee J, Joo EJ, Lim HJ, Park JM, Lee KY, Park A, Seok A, Lee H, Kang HG: Proteomic analysis of serum from patients with major depressive disorder to compare their depressive and remission statuses. *Psychiatry Investig* 2015, 12(2):249-259.

156. Song YR, Wu B, Yang YT, Chen J, Zhang LJ, Zhang ZW, Shi HY, Huang CL, Pan JX, Xie P: Specific alterations in plasma proteins during depressed, manic, and euthymic states of bipolar disorder. *Braz J Med Biol Res* 2015, 48(11):973-982.

157. Kishore U, Reid KB: C1q: structure, function, and receptors. *Immunopharmacology* 2000, 49(1-2):159-170.

158. Botto M, Walport MJ: C1q, autoimmunity and apoptosis. *Immunobiology* 2002, 205(4-5):395-406.

159. Sellar GC, Blake DJ, Reid KB: Characterization and organization of the genes encoding

the A-, B- and C-chains of human complement subcomponent C1q. The complete derived amino acid sequence of human C1q. *Biochem J* 1991, 274 (Pt 2):481-490.

160. Twal WO, Czirok A, Hegedus B, Knaak C, Chintalapudi MR, Okagawa H, Sugi Y, Argraves WS: Fibulin-1 suppression of fibronectin-regulated cell adhesion and motility. *Journal of cell science* 2001, 114(Pt 24):4587-4598.

161. Greenwood TA, Akiskal HS, Akiskal KK, Kelsoe JR: Genome-wide association study of temperament in bipolar disorder reveals significant associations with three novel Loci. *Biological psychiatry* 2012, 72(4):303-310.

162. Brigelius-Flohé R, Maiorino M: Glutathione peroxidases. *Biochimica et biophysica acta* 2013, 1830(5):3289-3303.

163. Maccarrone G, Ditzen C, Yassouridis A, Rewerts C, Uhr M, Uhlen M, Holsboer F, Turck CW: Psychiatric patient stratification using biosignatures based on cerebrospinal fluid protein expression clusters. *J Psychiatr Res* 2013, 47(11):1572-1580.

164. Fullerton JM, Tiwari Y, Agahi G, Heath A, Berk M, Mitchell PB, Schofield PR: Assessing oxidative pathway genes as risk factors for bipolar disorder. *Bipolar disorders* 2010, 12(5):550-556.

165. Dimick MK, Cazes J, Fiksenbaum LM, Zai CC, Tampakeras M, Freeman N, Youngstrom EA, Kennedy JL, Goldstein BI: Proof-of-concept study of a multi-gene risk score in adolescent bipolar disorder. *J Affect Disord* 2020, 262:211-222.

166. Harada N, Iijima S, Kobayashi K, Yoshida T, Brown WR, Hibi T, Oshima A, Morikawa M: Human IgGFc binding protein (FcgammaBP) in colonic epithelial cells exhibits mucin-like structure. *J Biol Chem* 1997, 272(24):15232-15241.

167. Pacifico R, Davis RL: Transcriptome sequencing implicates dorsal striatum-specific gene network, immune response and energy metabolism pathways in bipolar disorder. *Mol Psychiatry* 2017, 22(3):441-449.

168. Tisch R, Roifman CM, Hozumi N: Functional differences between immunoglobulins M and D expressed on the surface of an immature B-cell line. *Proc Natl Acad Sci U S A*

1988, 85(18):6914-6918.

169. Yoshimi A, Yamada S, Kunimoto S, Aleksic B, Hirakawa A, Ohashi M, Matsumoto Y, Hada K, Itoh N, Arioka Y *et al*: Proteomic analysis of lymphoblastoid cell lines from schizophrenic patients. *Transl Psychiatry* 2019, 9(1):126.

170. Zhuo L, Hascall VC, Kimata K: Inter-alpha-trypsin inhibitor, a covalent protein-glycosaminoglycan-protein complex. *J Biol Chem* 2004, 279(37):38079-38082.

171. Wang Q, Su X, Jiang X, Dong X, Fan Y, Zhang J, Yu C, Gao W, Shi S, Jiang J *et al*: iTRAQ technology-based identification of human peripheral serum proteins associated with depression. *Neuroscience* 2016, 330:291-325.

172. Alic L, Papac-Milicevic N, Czamara D, Rudnick RB, Ozsvar-Kozma M, Hartmann A, Gurbisz M, Hoermann G, Haslinger-Hutter S, Zipfel PF *et al*: A genome-wide association study identifies key modulators of complement factor H binding to malondialdehyde-epitopes. *Proc Natl Acad Sci U S A* 2020, 117(18):9942-9951.

173. Jaros JA, Martins-de-Souza D, Rahmoune H, Rothermundt M, Leweke FM, Guest PC, Bahn S: Protein phosphorylation patterns in serum from schizophrenia patients and healthy controls. *J Proteomics* 2012, 76 Spec No.:43-55.

174. Cross-Disorder Group of the Psychiatric Genomics C: Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013, 381(9875):1371-1379.

175. Jia P, Kao CF, Kuo PH, Zhao Z: A comprehensive network and pathway analysis of candidate genes in major depressive disorder. *BMC Syst Biol* 2011, 5 Suppl 3:S12.

176. Ammala AJ, Urrila AS, Lahtinen A, Santangeli O, Hakkarainen A, Kantojarvi K, Castaneda AE, Lundbom N, Marttunen M, Paunio T: Epigenetic dysregulation of genes related to synaptic long-term depression among adolescents with depressive disorder and sleep symptoms. *Sleep Med* 2019, 61:95-103.

177. Gaspar L, van de Werken M, Johansson AS, Moriggi E, Owe-Larsson B, Kocks JW, Lundkvist GB, Gordijn MC, Brown SA: Human cellular differences in cAMP--CREB

signaling correlate with light-dependent melatonin suppression and bipolar disorder. *Eur J Neurosci* 2014, 40(1):2206-2215.

178. Zak N, Moberget T, Boen E, Boye B, Waage TR, Dietrichs E, Harkestad N, Malt UF, Westlye LT, Andreassen OA *et al*: Longitudinal and cross-sectional investigations of long-term potentiation-like cortical plasticity in bipolar disorder type II and healthy individuals. *Transl Psychiatry* 2018, 8(1):103.

179. Grande I, Fries GR, Kunz M, Kapczinski F: The role of BDNF as a mediator of neuroplasticity in bipolar disorder. *Psychiatry Investig* 2010, 7(4):243-250.

180. Filipovic D, Todorovic N, Bernardi RE, Gass P: Oxidative and nitrosative stress pathways in the brain of socially isolated adult male rats demonstrating depressive- and anxiety-like symptoms. *Brain Struct Funct* 2017, 222(1):1-20.

181. Maurya PK, Noto C, Rizzo LB, Rios AC, Nunes SO, Barbosa DS, Sethi S, Zeni M, Mansur RB, Maes M *et al*: The role of oxidative and nitrosative stress in accelerated aging and major depressive disorder. *Prog Neuropsychopharmacol Biol Psychiatry* 2016, 65:134-144.

182. Maes M, Landucci Bonifacio K, Morelli NR, Vargas HO, Barbosa DS, Carvalho AF, Nunes SOV: Major Differences in Neurooxidative and Neuronitrosative Stress Pathways Between Major Depressive Disorder and Types I and II Bipolar Disorder. *Mol Neurobiol* 2019, 56(1):141-156.

ABSTRACT IN KOREAN

국문초록

서론: 액체 크로마토그래피 및 질량 분석법 기반 단백체 접근법이 특정 질병 및 장애와 관련된 바이오 마커를 발굴하고 개발하기 위해 적용되었다. 액체 크로마토그래피 고해상도 질량 분석법을 기반으로하는 비표적 단백체학은 수천 개의 단백질의 식별과 정량을 동시에 가능하게 하여 소량의 샘플에서 수백 개의 차등 발현 단백질을 생성한다. 액체 크로마토그래피 다중반응겹지 질량 분석법을 포함한 표적 단백체학은 높은 민감도, 정확도 및 재현성을 기반으로 표적 단백질을 정량하는데 사용된다. 임상 단백체학 연구에서 포르말린 고정 파라핀 포매조직절편 (FFPE), 혈액 및 기타 체액과 같은 임상 코호트에서 수집된 병리 및 임상 검체가 분석되었다. 임상 단백체학 분석의 경우 액체 크로마토그래피 및 질량 분석법 기반 접근법은 생체표지자의 발굴 및 개발과 높은 처리량과 높은 민감도로 임상 진단에 기여하는 강력한 기술이다. 또한, 액체 크로마토그래피 및 질량 분석법에 기반한 단백체학 연구는 특정 질병과 장애의 생물학적 및 분자적 특징에 대한 이해에 기여할 것이다.

방법: 1 장에서는 필터 보조 검체 준비 (FASP), 연속 질량 꼬리 표지, 높은 산도 분획 및 액체 크로마토그래피-고분해능-질량분석법을 결합하여 원격 전이성 유방암 및 비원격 전이성 유방암의 포르말린 고정 파라핀 포매조직절편(FFPE)을 사용하여 심층 단백질 프로파일링 데이터를 획득하기 위한 통합 비표적 단백질 접근법이 적용되었다. 통계 분석은 차등 발현 단백질을 결정하고 원격 전이성 유방암을 예측하기 위한 후보 생체표지자를 발굴하기 위해 수행되었다. 원격 전이성 유방암의 분자 특성을 조사하기 위해 차등 발현 단백질 사용하여 유전자 온톨로지, 질병 및 기능, 표준 경로와 관련하여 생물정보학 분석이 수행되었다.

또한 후보 생체표지자의 원격 전이 가능성을 검증하기 위해 실시간 중합효소 연쇄 반응과 침입/이주 분석을 수행했다. 제 2 장에서는 크로마토그래피–다중반응검지–질량분석법에 기반한 표적 단백질 접근 방식을 적용하여 임상 코호트의 혈장 검체에서 주요 우울 장애 및 양극성 장애와 관련된 단백질 후보 생체표지자를 정량했다. 기술적 편차를 줄이기 위해 크로마토그래피–다중반응검지–질량분석법 데이터의 배치 효과 보정이 수행되었다. 이후 발현 양에 차이를 보이는 후보 단백질 생체표지자를 결정하기 위해 통계분석이 수행되었고, 특징 추출, 교차 검증 및 가중 모델 평균화를 결합한 최소 절대 수축 및 선택 연산자에 기반한 머신 러닝 접근법이 주요 우울 장애 와 양극성 장애를 구별하기 위한 잠재적 진단 모델을 개발하기 위해 수행되었다. 또한, 모델에 포함된 단백질과 기분 장애 사이의 생물학적 관계를 조사하기 위해 생물정보학 기반 네트워크 분석을 수행하였다.

결과: 1 장에서 포르말린 고정 파라핀 포매조직절편–연속 질량 꼬리 표지 풀링 샘플 세트와 포르말린 고정 파라핀 포매조직절편–연속 질량 꼬리 표지 개별 샘플 세트로부터 각각 원격 전이 및 비원격 전이 그룹을 비교한 총 9,441 개 및 8,746 개의 단백질이 동정 되었다. 또한, 저침습성 및 고침습성 세포주를 비교한 유방암 세포주–연속 질량 꼬리 표지 샘플 세트에서 총 7,823 개의 단백질이 동정 되었다. 후보 생체표지자의 단계별 결정 기준에 따라 2 개의 단백질(LTF, TUBB2A)을 유방암 원격전이 예측을 위한 후보 생체표지자로 결정하였다. 14 개 유방암 세포주의 RT–PCR 데이터의 LTF 와 TUBB2A 발현 패턴을 크로마토그래피–질량분석 데이터의 발현 패턴과 비교했을 때, TUBB2A 만이 두 데이터 사이에서 일관된 발현 패턴을 보였다. 그 결과, TUBB2A 는 이후 원격 전이 활성이 검증되는 새로운 생체표지자 후보로 선정되었다. 또한 생물정보학적 결과를 통해 원격 전이의 전반적인 분자적 특징을 도출하였으며, 유방암 아형 간 원격 전이성 유방암의 분자 기능 차이를 입증하였다. 제 2 장에서는 270 명의 혈장 샘플[90 명의 주요 우울 장애, 90 명의 양극성 장애, 90 명의 정상 대조군]에서 주요 우울 장애 및 양극성 장애 에 관한 671 펩타이드에 해당하는 총 210 개의

단백질표적을 크로마토그래피-다중반응검지-질량분석법을 사용하여 안정적으로 정량 하였다. 훈련 세트(72 명의 주요 우울 장애 및 72 명의 양극성 장애)에서는 9 개의 혈장 단백질로 구성된 일반화 가능한 모델이 개발되었다. 모델은 테스트 세트(18 명의 주요 우울 장애 및 18 명의 양극성 장애)에서 평가되었다. 이 모델은 훈련 (곡선 아래의 면적 = 0.84)과 테스트 세트(곡선 아래의 면적 = 0.81)에서 MDD 를 BD 와 구별하고 현재 고조증/저조증/혼합 증상 (90 명의 주요 우울 장애 및 75 명의 양극성 장애)(곡선 아래의 면적 = 0.83)에서 우수한 성능(곡선 아래의 면적 > 0.8)을 보였다. 그 후, 이 모델은 약물 투여 경험이 없는 주요 우울 장애와 양극성 장애 환자 (11 명의 주요 우울 장애 및 10 명의 양극성 장애)(곡선 아래의 면적 = 0.96)에서 우수한 성능을 보였고, 주요 우울 장애 대 정상 대조군(곡선 아래의 면적 = 0.87) 및 양극성 장애 대 정상 대조군 (곡선 아래의 면적 = 0.86)에서 우수한 성능을 보였다. 또한, 9 개의 단백질은 신경, 산화/질소 스트레스, 면역/염증 관련 생물학적 기능과 관련이 있었다.

결론: 제 1 장에서, 본 연구는 포르말린 고정 파라핀 포매조직절편 조직을 사용하여 가장 큰 원격 전이성 유방암 단백체를 처음으로 구축하였다. 깊이 있는 단백체 데이터를 통해 새로운 생체표지자 후보와 원격 전이성 유방암의 단백체 특성을 발견할 수 있었다. 다양한 유방암 아형에서 원격 전이성 유방암의 뚜렷한 분자적 특징도 확립되었다. 우리의 단백체 데이터는 원격 전이성 유방암 연구에 귀중한 자원을 제공한다. 제 2 장에서는 표적 단백체학 접근방식을 사용하여 주요 우울 장애 및 양극성 장애 환자를 구별 가능성을 제안했다. 우리는 주요 우울 장애와 양극성 장애를 구별하기 위해 9 개 혈장 단백질로 구성된 일반화 가능한 모델을 개발했다. 우리의 결과는 이러한 장애가 개발된 모델을 사용하여 구별 및 진단 할 수 있음을 시사한다. 또한, 우리는 9 개의 혈장 단백질이 우울 장애와 양극성 장애와 생물학적으로 중요한 연관성을 가질 것을 제안한다.

주요어: 비표적 그리고 표적 단백체학; 액체 크로마토그래피-고해상도 질량분석법; 액체 크로마토그래피-다중반응검지-질량분석법; 생체표지자; 원격 전이성 유방암; 정서질환

학번: 2017-25626

* 본 내용은 Clinical Proteomics 학술지에 출판 완료된 내용임. Identification of TUBB2A by quantitative proteomic analysis as a novel biomarker for the prediction of distant metastatic breast cancer/Clinical Proteomics/ 10.1186/s12014-020-09280-z.

* 본 내용은 Journal of Proteome Research 학술지에 출판 완료된 내용임. Quantitative Proteomic Approach for Discriminating Major Depressive Disorder and Bipolar Disorder by Multiple Reaction Monitoring-Mass Spectrometry/Journal of Proteome Research/10.1021/acs.jproteome.1c00058.