수의학박사학위논문

프로테오지노믹스 기법을 이용한 폐암 바이오마커 연구

Studies on Lung Cancer Biomarkers
by Proteogenomic Analysis

2019 년 2 월

서울대학교 대학원

수의학과 수의생명과학 전공

(수의생화학)

김 용 인

프로테오지노믹스 기법을 이용한 폐암 바이오마커 연구

Studies on Lung Cancer Biomarkers by Proteogenomic Analysis

지도교수 조제열

이 논문을 수의학박사학위논문으로 제출함

2018 년 11 월

서울대학교 대학원

수의학과 수의생명과학 전공 (수의생화학)

김용인

김용인의 박사학위논문을 인준함

2018 년 12 월

위원장 이 항 (인)

부위원장 조 제열 (인)

위원 백윤기 (인)

위원 이 진환 (인)

위원 이 소영 (인)

Studies on Lung Cancer Biomarkers by Proteogenomic Analysis

Under the supervision of Professor Je-Yoel Cho

DISSERTATION

Presented in Partial Fulfillment of the Requirement for
the Degree of DOCTOR OF PHILOSOPHY

By

Yong-In Kim

Major in Veterinary Biomedical Sciences
(Veterinary Biochemistry)
Department of Veterinary Medicine
The Graduate School
Seoul National University

February 2019

ABSTRACT

Studies on Lung Cancer Biomarkers by Proteogenomic Analysis

Yong-In Kim

Major in Veterinary Biomedical Sciences
Department of Veterinary Medicine
The Graduate School
Seoul National University

Biomarkers have been in high demand for disease diagnosis and therapeutics. Traditional hypothesis-based research has been challenging due to massive screening studies. Together with the emergence of omics technologies, currently, the paradigm for disease research has been moving toward evidence-based large-scale discovery studies. Proteins, as key effector molecules, can serve as ideal biomarkers for various diseases because they catalyze every biological function. Proteomics, which is represented by mass spectrometry (MS) technologies, stands as a solution for disease diagnosis and drug target discovery.

CHAPTER I includes a portion of a report from of the human proteome project (HPP) related to chromosome 9 (Chr 9). To identify missing proteins (MPs) and their potential features in regard to proteogenomic view, both LC-MS/MS analysis and next-generation RNA sequencing (RNA-seq)-based tools were used for the clinical samples including adjacent non-tumor tissues. When the Chr 9 working group of the Chromosome-Centric Human Proteome Project (C-HPP) began this project, there were 170 remaining MPs encoded by Chr 9 (neXtProt 2013.09.26 rel.); currently, 133 MPs remain unidentified at present (neXtProt 2015.04.28 rel.). Proteome analysis in this study identified 19 missing proteins across all chromosomes and one MP (SPATA31A4) from Chr 9. RNA-seq analysis enable detection of RNA expression of 4 nonsynonymous (NS) SNPs (in CDH17, HIST1H1T, SAPCD2, and ZNF695) and 3 synonymous SNPs (in CDH17, CST1, and HNF1A) in all 5 tumor tissues but not in any of the adjacent normal tissues. By constructing a cancer patient sample-specific protein database based on individual RNA-seq data, and by searching the proteomics data from the same sample, 7 missense mutations in 5 genes (LTF, HDLBP, TF, HBD, and HLA-DRB5) were identified. Two of these mutations were found in tumor tissues but not in the paired normal tissues. Additionally, this study discovered peptides that were derived from the expression of a pseudogene (EEF1A1P5) in both tumor and normal tissues. In summary, this proteogenomic study of human primary lung tumor tissues enabled detection of additional missing proteins and revealed novel missense mutations and synonymous SNP signatures, some of which are predicted to be specific to lung cancer.

CHAPTER II describes a study of the combination marker model using multiple reaction monitoring (MRM) quantitative data. Misdiagnosis of lung cancer remains a serious problem due to the difficulty of distinguishing lung cancer from other respiratory lung diseases. As a result, the development of serum-based differential diagnostic biomarkers is in high demand. In this study, 198 serum samples from non-cancer lung disease and lung cancer patients were analyzed using nLC-QqQ-MS to examine the diagnostic efficacy of seven lung cancer biomarker candidates. When the candidates were assessed individually, only SERPINEA4 showed statistically significant changes in the sera of cancer patient compared to those of control samples. The MRM results and clinical information were analyzed using logistic regression analysis to a select model for the best 'meta-marker', or combination of biomarkers for the differential diagnosis. Additionally, under consideration of statistical interaction, variables having a low significance as a single factor but statistically influencing the meta-marker model were selected. Using this probabilistic classification, the best meta-marker was determined to comprise two proteins SERPINA4 and

PON1, with an age factor. This meta-marker showed an enhanced differential diagnostic capability (AUC=0.915) to distinguish the lung cancer from lung disease patient groups. These results suggest that a statistical model can determine optimal meta-markers, which may have better specificity and sensitivity than a single biomarker and may thus improve the differential diagnosis of lung cancer and lung disease patients.

Keywords : Lung cancer; LC-MS/MS; Proteogenomics; Proteomics;
Multiple reaction monitoring (MRM)

Student Number : 2013-21536

CONTENTS

Page

CHAPTER I

CHAPTER II

LIST OF FIGURES

BACKGROUND

CHAPTER I

CHAPTER II

LIST OF TABLES

BACKGROUND

CHAPTER I

CHAPTER II

ABBREVIATIONS

2-DE	Two-Dimensional Gel Electrophoresis
ADC	Lung Adenocarcinoma
AUC	Area Under the Curve
C-HPP	Chromosome-Centric Human Proteome Project
CE	Collision Energy
CI	Confidence Interval
CNS	Central Nervous System.
CT	Computed Tomography
Curr	Current-Smoker
ED	Extensive Disease
ELISA	Enzyme-Linked ImmunoSorbent Assay
ESI	Electrospray Ionization
Ex	Ex-Smoker

FA	Formic Acid
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase of exon per Million fragments mapped
HPLC	High Performance Liquid Chromatography
HPP	Human Proteome Project
IAA	Iodoacetamide
iTRAQ	Isobaric Tag for Relative and Absolute Quantification
LC	Liquid Chromatography
LD	Limited Disease
MALDI	Matrix-Assisted Laser Desorption Ionization
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MW	Molecular Weight
ND	Nodule
NSCLC	Non-Small Cell Lung Cancer

Non	Non-Smoke
PN	Pneumonia
PTM	Post-Translational Modification
QqQ	Triple Quadrupole
SAAV	Single Amino Acid Variant
SCLC	Small Cell Lung Cancer
SCX	Strong Cation Exchange
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SE	Standard Error
SID	Stable Isotope Dilution
SILAC	Stable Isotope Label with Amino Acid in Cell Culture
SIS-peptide	Stable Isotope-labeled Synthetic peptide
Sig.	Significance
SQCC	Lung Squamous Cell Carcinoma
TB	Tuberculosis
TMT	Tandem Mass Tag

BACKGROUND

1. LUNG CANCER

Cancer is a major life-threatening disease in Korea (Table B-1). In 2015, the total number of deaths from cancer was 76,855, accounting for 27.9% of all deaths (Jung et al., 2018). The number of cancer incidences and deaths is expected to increase with the aging population and westernized lifestyles (Jung et al., 2015). Additionally, the economic burden of cancer in Korea has increased approximately 1.8-fold, from $11,424 to $20,858 million, between 2000 and 2010 (Lee et al., 2015).

Lung cancer has demonstrated the highest mortality among all types of solid cancers in both the US and Korea (Figure B-1 and Table B-2). According to the cancer statistics reported in 2002 and 2012, lung cancer was the most common cancer since 1985 and by 2002, and in 2012, lung cancer was not the most frequently diagnosed cancer, but it was the second leading cause of cancer related deaths (Jemal et al., 2011, Parkin et al., 2005). Regarding the cancer statistics in 2017 estimated in the US, the 5-year survival rate of patients who were diagnosed with lung cancer showed the least increase along with liver and pancreatic cancers (Siegel et al., 2017), likely due to the low diagnostic rate in the early stages of cancer.

Compared with other cancers, lung cancer has little subjective symptoms and there are few screening methods for lung cancer diagnosis. The diagnosis of lung cancer is still largely depend on imaging techniques; such as X-rays and computed tomography (CT) scans. Histopathological examination is conducted on the biopsied mass when a suspicious mass is detected during imaging screening. The recent development of low-dose, fast-spiral CT and advances in imaging machines have improved the accuracy of chest CT (Midthun and Jett, 2008). Despite the improvement of imaging diagnostic methods, several problems, such as the cost, high misdiagnosis rates, and poor early diagnosis rates, persist.

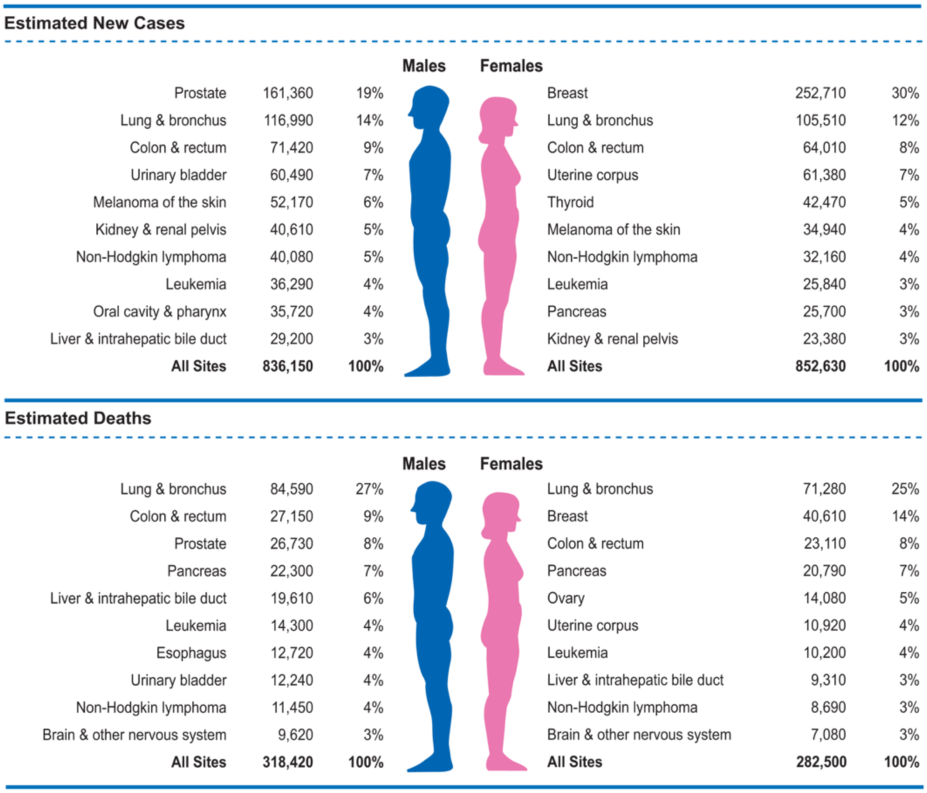
Table B-1. The top 10 leading causes of death in Korea, 2015

Rank	Cause of death	No. of deaths (%)	Age-standardized death rate per 100,000*
	All causes	275,895 (100)	289.3
1	Cancer	76,855 (27.9)	82
2	Heart disease	28,326 (10.3)	27.9
3	Cerebrovascular disease	24,455 (8.9)	23.4
4	Pneumonia	14,718 (5.3)	13.5
5	Intentional self-harm (suicide)	13,513 (4.9)	18.3
6	Diabetes mellitus	10,558 (3.8)	10.2
7	Chronic lower respiratory diseases	7,538 (2.7)	6.8
8	Disease of liver	6,847 (2.5)	8.1
9	Transport accidents	5,539 (2.0)	7.6
10	Hypertensive diseases	5,050 (1.8)	4.6
	Others	82,496 (29.9)	87.1

Source: Mortality Data, 2015, Statistics Korea.

*Age-adjusted using the Segi's world standard population.

Adapted from Jung et al., 2018



Estimated New Cases

Males			Females		
Prostate	161,360	19%	Breast	252,710	30%
Lung & bronchus	116,990	14%	Lung & bronchus	105,510	12%
Colon & rectum	71,420	9%	Colon & rectum	64,010	8%
Urinary bladder	60,490	7%	Uterine corpus	61,380	7%
Melanoma of the skin	52,170	6%	Thyroid	42,470	5%
Kidney & renal pelvis	40,610	5%	Melanoma of the skin	34,940	4%
Non-Hodgkin lymphoma	40,080	5%	Non-Hodgkin lymphoma	32,160	4%
Leukemia	36,290	4%	Leukemia	25,840	3%
Oral cavity & pharynx	35,720	4%	Pancreas	25,700	3%
Liver & intrahepatic bile duct	29,200	3%	Kidney & renal pelvis	23,380	3%
All Sites	**836,150**	**100%**	**All Sites**	**852,630**	**100%**

Estimated Deaths

Males			Females		
Lung & bronchus	84,590	27%	Lung & bronchus	71,280	25%
Colon & rectum	27,150	9%	Breast	40,610	14%
Prostate	26,730	8%	Colon & rectum	23,110	8%
Pancreas	22,300	7%	Pancreas	20,790	7%
Liver & intrahepatic bile duct	19,610	6%	Ovary	14,080	5%
Leukemia	14,300	4%	Uterine corpus	10,920	4%
Esophagus	12,720	4%	Leukemia	10,200	4%
Urinary bladder	12,240	4%	Liver & intrahepatic bile duct	9,310	3%
Non-Hodgkin lymphoma	11,450	4%	Non-Hodgkin lymphoma	8,690	3%
Brain & other nervous system	9,620	3%	Brain & other nervous system	7,080	3%
All Sites	**318,420**	**100%**	**All Sites**	**282,500**	**100%**

Figure B-1. Ten Leading Cancer Types for the Estimated New Cancer Cases and Deaths by Sex, United States, 2017.

Estimates are rounded to the nearest 10 and cases exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder.

Adapted from Siegel et al., 2017

Table B-2. Cancer incidence, deaths and prevalence by sex in Korea, 2015

Rank by Deaths	Site/Type	Deaths			New cases		
		Both sexes	Men	Women	Both sexes	Men	Women
	All sites	76,855	47,678	29,177	214,701	113,335	101,366
1	Lung	17,399	12,677	4,722	24,267	17,015	7,252
2	Liver	11,311	8,382	2,929	15,757	11,732	4,025
3	Stomach	8,526	5,507	3,019	29,207	19,545	9,662
4	Colon and rectum	8,301	4,698	3,603	26,790	15,911	10,879
5	Pancreas	5,439	2,908	2,531	6,342	3,359	2,983
6	Gallbladder*	4,211	2,135	2,076	6,251	3,220	3,031
7	Other and ill defined	3,931	2,101	1,830	16,309	8,300	8,009
8	Breast	2,354	16	2,338	19,219	77	19,142
9	Non-Hodgkin lymphoma	1,771	1,026	745	4,396	2,519	1,877
10	Leukemia	1,720	1,003	717	3,242	1,830	1,412
11	Prostate	1,700	1,700	-	10,212	10,212	-
12	Esophagus	1,531	1,401	130	2,420	2,201	219
13	Bladder	1,299	960	339	4,033	3,245	788
14	Brain and CNS	1,266	674	592	1,776	958	818

15	Lip, oral cavity, and pharynx	1,170	884	286	3,309	2,390	919
16	Ovary	1,055	-	2,443	2,443	-	2,443
17	Cervix uteri	967	-	3,582	3,582	-	3,582
18	Kidney	952	672	280	4,555	3,134	1,421
19	Multiple myeloma	889	466	423	1,455	762	693
20	Larynx	344	319	25	1,146	1,079	67
21	Thyroid	341	106	235	25,029	5,386	19,643
22	Corpus uteri	319	-	2,404	2,404	-	2,404
23	Hodgkin lymphoma	49	33	16	271	174	97
24	Testis	10	10	-	286	286	-

*Include the gallbladder and other/unspecified parts of the biliary tract.

Adapted from Jung et al., 2018

2. BIOMARKERS

Biomarkers are a means of measurement that can define the normal and abnormal status of an individual. Biomarkers have been in high demand, because the discovery of predictive biomarkers will save time and money, and leading to better diagnoses and disease cure. Biomarkers include any types of hall mark of physiological states, such as physiologic profiles, images, genes, or proteins (Dalton and Friend, 2006). Particularly, proteins as key effector molecule have been regarded as ideal biomarkers for various diseases because they catalyze every biological function (Gygi and Aebersold, 2000a).

The advantages of proteins as a class of biomarkers include their enormous diversity, dynamic turnover and secretion into blood and bodily fluids. There is an estimated number of 20,300 genes (Legrain et al., 2011), 40,000 unique metabolites (Wishart et al., 2012), ~100,000 mRNA transcripts, and up to 1.8 millions of different proteoforms, if posttranslational modifications (PTMs) are considered (Jensen, 2004). Such enormous diversity in proteoforms increases the chances to identify a marker, or a panel of markers, for each disease state. Since protein sequences may also reflect some genomic variations, a single instrumentation platform of mass spectrometry can measure not only changes in protein abundance but also genomic and transcriptomic variations, such as mutant proteins (Drabovich et al., 2015).

However, traditional hypothesis-based research has been challenging due to massive screening studies for biomarker development. Together with emerging of proteomics, which is represented by mass spectrometry (MS) technologies, presently, the paradigm for biomarker development has been moving toward evidence-based, large-scale discovery studies. The major advantage of hypothesis-free MS-based proteomics is that no assumptions need to be made regarding the possible nature and number of potential biomarkers, in stark contrast to single-protein measurements in classical biomarker research.

Conceptually, MS-based proteomics combines all possible hypothesis-driven biomarker studies for each disease into one and furthermore defines the relationship of potential biomarkers to each other (Geyer et al., 2017). In practice, the challenges of plasma proteomics have, thus far, prevented in-depth and quantitative studies on large cohorts. Instead, a stepwise strategy for biomarker discovery has been advocated, with several phases in which the number of individuals increases from a few to many, whereas the number of proteins decreases from hundreds or thousands to just a few (Rifai et al., 2006, Geyer et al., 2017) (Figure B-2). The bottom-up proteomics strategy is regarded as a typical workflow for hypothesis-free biomarker discovery (Aebersold and Mann, 2016, Altelaar and Heck, 2012). Targeted proteomics for candidate verification is a second phase of the stepwise strategy. A few proteins (typically < 10) with differential expression in the discovery phase are tested in a larger and ideally independent cohort. Because immunoassays are often not available, targeted MS methods can be employed. The most widespread of these is multiple reaction monitoring (MRM). The sensitivity of proteomics can be improved to the low ng/ml or even high pg/ml ranges by more extensive sample preprocessing with depletion or fractionation both in the discovery and verification phases (Burgess et al., 2014, Kim et al., 2015, Nie et al., 2017). The final phase in the triangular strategy is the validation with immunoassays, a field that has matured over decades. For maximum specificity, sandwich assays are typically preferred (Geyer et al., 2017).

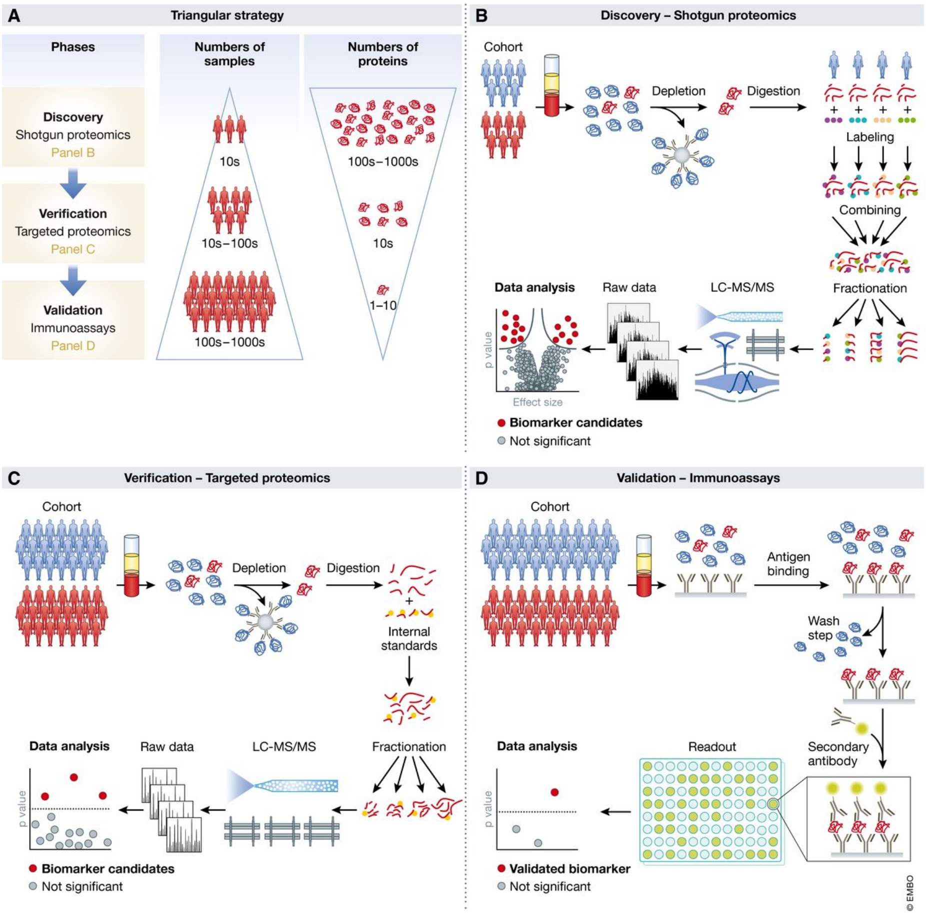


Figure B-2. Current paradigms in plasma biomarker research

(A) A relatively small number of cases and controls are analyzed by hypothesis-free discovery proteomics in great depth, ideally leading to the quantification of thousands of proteins (top layer in the panel). This may yield tens of candidates with differential expression that are screened by targeted proteomics methods in cohorts of moderate size (middle layer). Finally, for one or a few of the remaining candidates, immunoassays are developed, which are then validates in large cohorts and applied in the clinic (bottom layer). (B) Workflow for hypothesis-free discovery proteomics. (C) Targeted proteomics for candidate verification. (D) Development of immunoassays for clinical validation and application.

Adapted from Geyer et al., 2017

3. MASS SPECTROMETRY-BASED PROTEOMICS

Edman degradation, which is used to sequence a protein, relies on the identification of amino acids that have been chemically cleaved in a stepwise fashion from the amino terminus of the protein and requires much expertise(Steen and Mann, 2004). In 1996, Mann and colleagues showed that MS could identify gel-separated proteins using a much smaller quantity of the sample than was required by Edman degradation and can fragment the peptides in seconds instead of hours or days (Wilm et al., 1996). Currently, MS-based proteomics has proliferated, and many biologists have access to a service to which they can submit a sample and are handed back a list of proteins that have been identified by MS.

To measure biomolecules, which can be peptides or proteins, by mass spectrometry, analytes are ionized via electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) (Figure B-3d), and their mass is measured by following their specific trajectories in vacuum system. Ionized molecules are recorded as values on the m/z scale, which has units of mass per charge (Steen and Mann, 2004).

Having determined the m/z values and intensities of all the peaks in the spectrum, the mass spectrometer then proceeds to obtain sequence information about these biomolecules. This process is called MS/MS because it couples two stages of MS. In tandem MS, a particular biomolecule ion is isolated, energy is imparted by collisions with an inert gas, and this energy causes the analyte to break apart. A mass spectrum of the resulting fragments is then generated (Figure B-3e).

In general proteomics, the mass spectrometer does not measure proteins, but peptides. First, peptides can be easy to handle and are stable to introduce MS. Second, the sensitivity of MS for peptides is much better than that for proteins, and the protein might be processed and modified such that the combinatorial effect makes determining the masses of the numerous resulting isoforms impossible. Third, the sequence of a peptide is easy to predict, unlike that of a mature protein is not. Finally, mass spectrometry is most efficient at obtaining sequence information from peptides that are up to ~20 residues long, rather than from whole proteins peptides (Steen and Mann, 2004). The mostly highly sequence-specific proteases are used to convert proteins to peptides, such as trypsin.

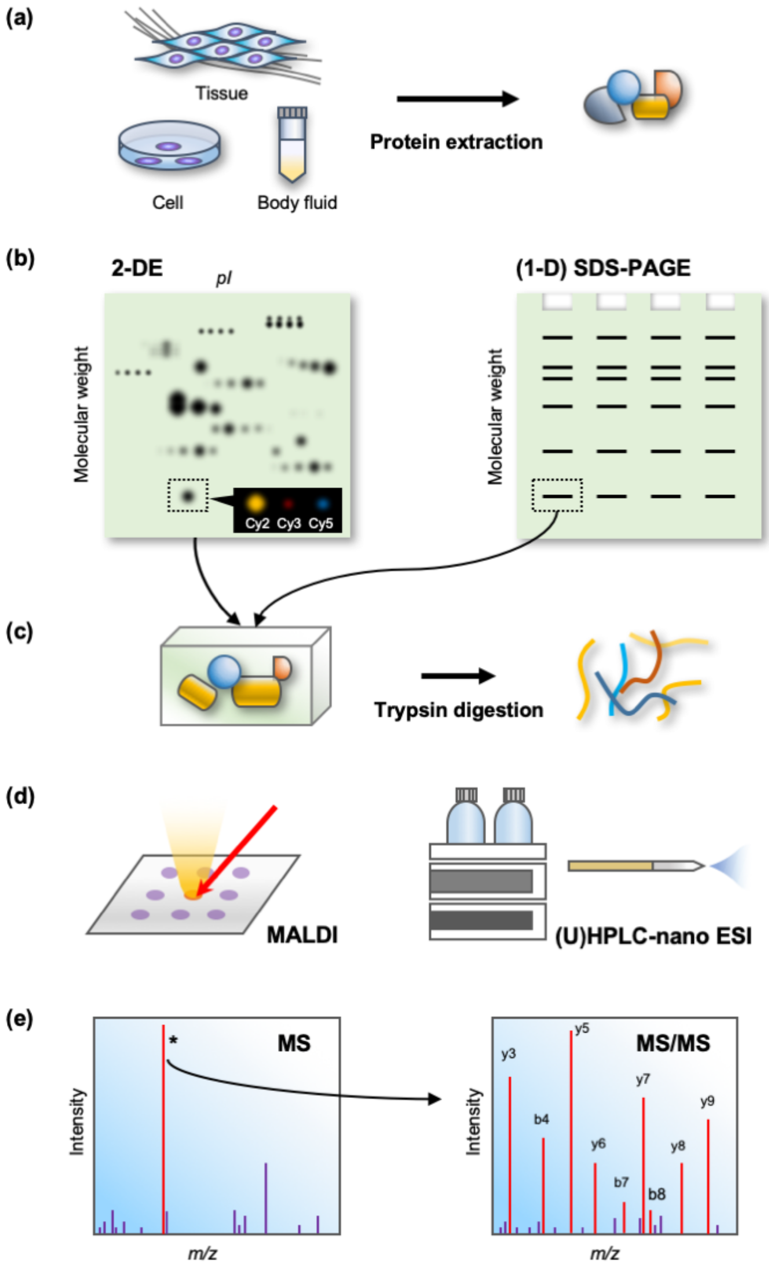


Figure B-3. General workflow of gel-based proteomics.

(a) Proteins are extracted from bio specimen. **(b)** Extracted protein mixture from biosamples separated by 2-DE or SDS-PAGE. In most case, proteins are quantified on a gel. Using the quantitative difference from DIGE, target spots can be selected from 2-DE. **(c)** Excised gel pieces are trypsinized and resulting peptides are collected. **(d)** Peptides are ionized via MALDI or nano ESI and are inducted to MS. **(e)** Peptide is measured in MS spectrum, followed by selected and isolated, subsequently fragmented to get the sequence information from MS/MS spectrum.

After the acquisition of peptide sequence information (MS/MS), peptide identification software tools analyze the fragmentation spectra by de novo or database search engines(Sinitcyn et al., 2018). De novo peptide sequencing uses only information from the input spectrum and characteristics of the fragmentation method. Mass differences between certain peak pairs correspond to amino acid masses, which are interpreted as consecutive ions in one of the expected fragment series.

However, the most popular approach is database search. This search method is easier than *de novo* sequencing because incompletely fragmented MS/MS spectra still have sufficient information to match it uniquely to a peptide sequence in the database (Steen and Mann, 2004). The database is generated from all protein sequences that are known or thought to be produced according to the instructions in the genome sequence of an organism. For a given measured fragmentation spectrum, the search engine calculates a match score against all theoretical fragmentation spectra within a specified peptide mass tolerance.

Since the highest scoring spectra might be false positive, most workflows control the false discovery rate using a target-decoy approach (Elias and Gygi, 2007). In this approach, fragmentation spectra are searched not only against the target database but also against a decoy database, which is designed to produce a false-positive result. Comparing the score distributions of the target and decoy spectra, posterior error probabilities could be calculated and false discovery rate (FDR) could be controlled (Figure B-4b) (Sinitcyn et al., 2018).

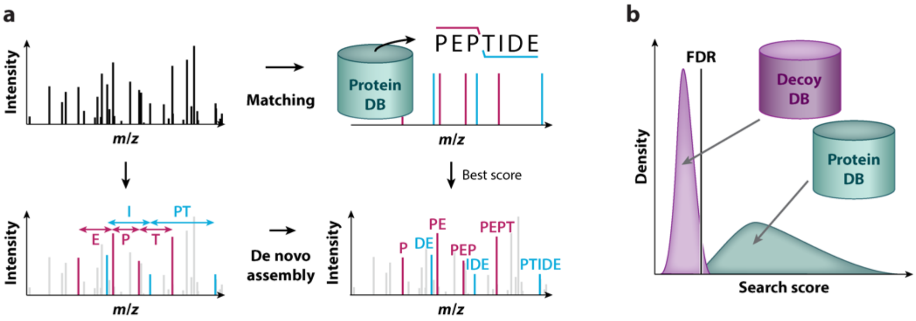


Figure B-4. Overview of peptide identification methods.

(a) In the peptide database search engine approach, measured MS/MS spectra are scored against a list of theoretical spectra from an *in silico* digest of protein sequences. De novo peptide identification allows reading the peptide sequence partially or completely out of the MS/MS spectrum. **(b)** In the target-decoy approach, true and decoy protein sequences are offered to estimate the false discovery rate (FDR).

Adapted from Sinitcyn et al., 2018

In the early era of proteomics, gel electrophoresis was a dominant technique of sample preparation for MS analysis. O'Farrell introduced modern 2-DE in 1975 that combined the separation according to the charges by isoelectric focusing under denaturing conditions with the fractionation corresponding to the sizes by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (Westermeier, 2016, O'Farrell, 1975). This separation technique provides a suitable resolution of protein mixture and has become the standard methods for proteome analysis (Wasinger et al., 1995). Highly separated proteins through 2-DE are visualized by gel staining (Chevalier, 2010) followed by the preparation of gel spots, which include the protein-of-interest, excision and in-gel digestion using an endopeptidase (e.g., trypsin, lys-C, glu-C, and chymotrypsin). The resulting peptides generated by trypsin were highly capable of ionization and were measurable within the MS dynamic range (Gygi et al., 2000, Hustoft et al., 2012, Shevchenko et al., 2006). The peptides were measured by MS via MALDI or ESI. Tandem mass spectra derived from the fragmented peptide ion in the mass spectrometry were unique identifiers for amino acid sequence information (Figure B-3).

It is undeniable that gel-based proteomics represent one of the most powerful protein separation and qualitative methods. SDS-PAGE has a nice capability to separate proteomes; however, some proteins are not applicable, such as highly acidic or basic proteins, extremely high/low molecular weight (MW) proteins, and membrane proteins (Gygi and Aebersold, 2000b). Additionally, the gel has a loading capacity limit of approximately 100 μg of proteins onto a single gel; and if using DIGE to compare three samples, the amount of protein loaded per sample is reduced (~33 μg). It should be noted that load limitation makes it difficult to detect low-abundant proteins (Mansour and Coorssen, 2018, Timms and Cramer, 2008). Additional weaknesses of gel separation are gel-to-gel variation, expensive dyes, uncertain recovery from a gel piece, and low-throughput identification due to the lack of an automated procedure (Wu et al., 2006, Thelen and Peck, 2007, Gygi and Aebersold, 2000b, Mansour and Coorssen, 2018, Noaman et al., 2017, Vadivel and Arun, 2015).

Alternative simple methods for large-scale study have been developed and demonstrated that LC-ESI-MS systems can handle highly complex peptide mixtures without gel separation (Appella et al., 1995). Thus, the application of liquid chromatography to the analysis of peptide mixtures generated by the proteolysis of complex protein samples is a considerable step toward gel-independent proteomic technologies (Yates et al., 1997). Instead of omitting a protein gel electrophoresis step, separating peptides directly using one or more orthogonal chromatography steps following in-solution digestion results in better protein identification coverage (Figure B-5) (Schirle et al., 2003, Washburn et al., 2001). Additionally, SDS-PAGE-incompatible proteins, which include low-abundance proteins such as transcription factors, protein kinases, and other regulatory proteins, are detectable and quantifiable in LC-ESI-MS/MS; thus, the LC-based proteomics utilizes a routine procedure for global proteome analysis (Gygi and Aebersold, 2000a). The contemporary trend of disease research has been altered to high-throughput omics approaches that involve not only proteomics but also genomics, transcriptomics and metabolomics. In this context, it stands to reason that gel-free, LC-based mass spectrometry has became a dominant player in modern proteomics.

The high-throughput fractionation technique was established to develop various quantitative methods, which can provide better sensitivity, a broader dynamic range, and high accuracy. Chemical labeling using a stable isotope (e.g., ^{18}O and dimethyl) or isobaric tags (e.g., iTRAQ and TMT), which are simple and quick methods compatible from *in vitro* to clinical disease samples (Hsu et al., 2003, Yao et al., 2001, Thompson et al., 2003, Ross et al., 2004). Metabolic labeling (e.g., SILAC and AHA) can reach a higher labeling efficiency than others (Dieterich et al., 2006, Ong et al., 2002). Gel-based proteome quantification is analyzed on the gel via densitometric imaging. However, the quantification techniques described above can be utilized only when measured by MS and/or tandem MS.

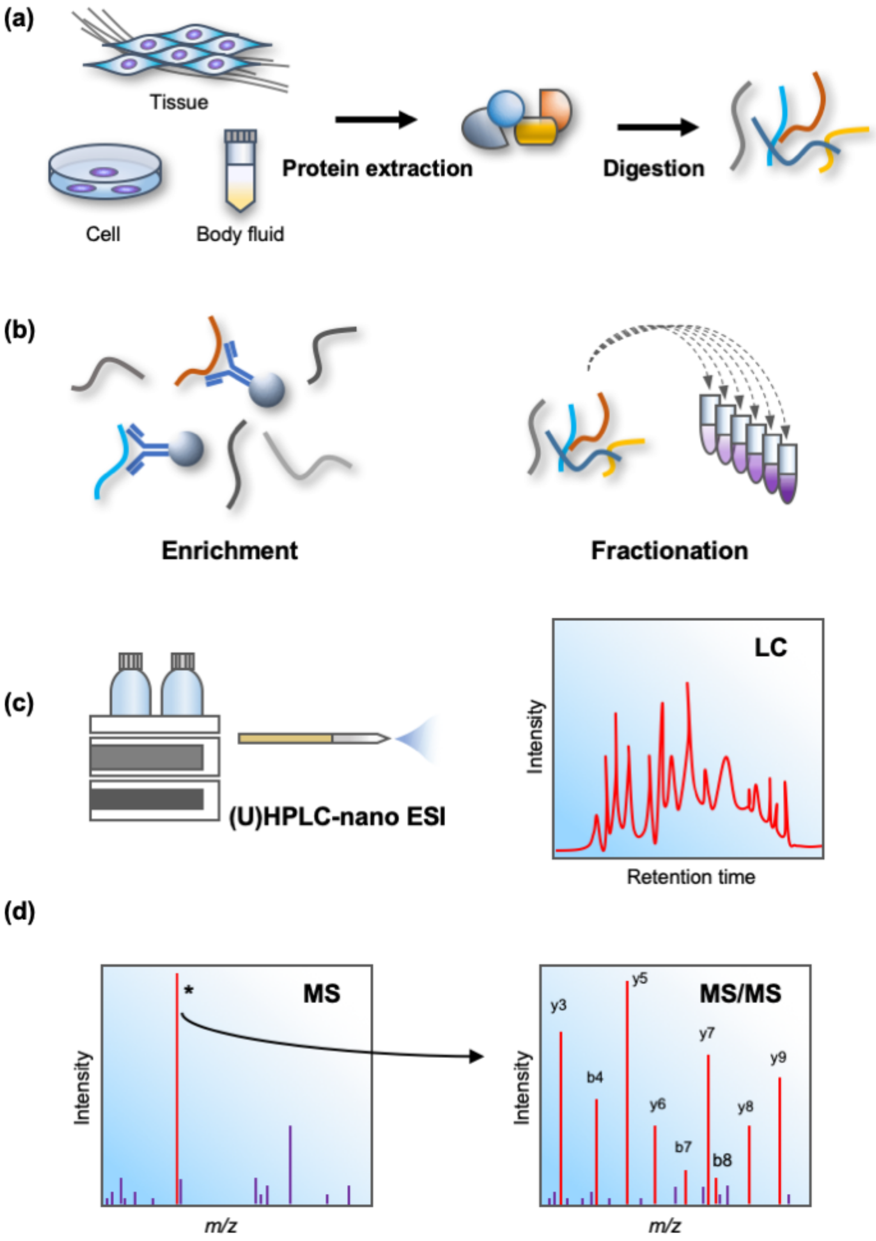


Figure B-5. Current gel-free proteomics workflow.

(a) Extracted proteins are digested without gel electrophoresis. **(b)** Peptide mixtures are enriched (e.g. tags, PTMs) and/or fractionated (e.g. SCX, High-pH RP) to reduce complexity, which contribute to enhanced sensitivity on MS. **(c)** Peptides are separated by reverse phase liquid chromatography, followed by ionization via nano ESI and are inducted to MS. **(d)** Each peptides is measured in MS spectrum, followed by selection and isolation, subsequently fragmented to get the sequence information from MS/MS spectrum.

4. PROTEOGENOMICS

The important step during proteogenomics is construction of a customized database of proteins sequences, which are obtained from genomic data and can be further used to investigate the model of interest (Figure B-6). Because the entire genome, exome and mRNA sequencing are available at a reasonable price, it is possible to explain the sequences of complete theoretical proteins that are present a specific structure to be evaluated (Sajjad et al., 2016).

Through proteogenomics different classes of peptide can be identified on the sample-sample specific genome. Novel peptides not found in any reference genome database include those that identify previously undiscovered protein-coding loci (intergenic peptides) and variant peptides (single amino acid variant, SAAV). These may also include peptide mapping to the untranslated regions of introns, peptides spanning the boundary between the coding region and neighboring intron region (exon extension), peptides spanning alternative splice junctions and out-of-frame peptides. Novel peptides may also provide us with evidence of protein expression for chimeric transcripts, transcripts thought to be noncoding RNAs, gene fusions and RNA-editing events, although such events are expected to be rare in proteomic data sets (Figure B-7) (Sajjad et al., 2016).

With the proteogenomics searches, peptides are identified based on customized protein sequence databases generated from genomic or transcriptomic information (Nesvizhskii, 2014). Search spaces for proteogenomics searches are typically larger than those in conventional searches because they often involve three- or six-frame translations of genomic sequences. Because of its large database size, it is necessary to consider the difference in the likelihood of identifying different classes of peptides (Figure B-8) (Ning and Nesvizhskii, 2010, Branca et al., 2014). Furthermore, these search spaces are heterogeneous, because the sequence content ranges from clearly existing, manually validated protein sequences to in silico-translated genomic regions without any prior evidence for their expression. Hence, extra measures need to be taken in the identification process to account for this heterogeneity.

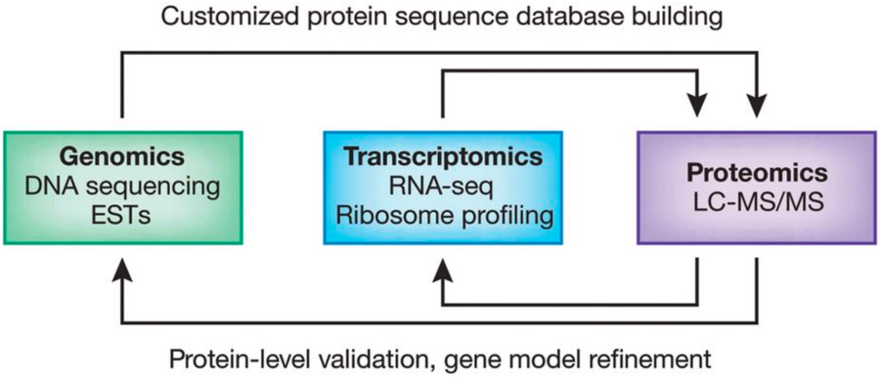

Customized protein sequence database building
Genomics
DNA sequencing
ESTs
Transcriptomics
RNA-seq
Ribosome profiling
Proteomics
LC-MS/MS
Protein-level validation, gene model refinement

Figure B-6. The concept of proteogenomics.

In a proteogenomic approach, genomic and transcriptomic data are used to generate customized protein sequence databases to help interpret proteomic data. In turn, the proteomic data provide protein-level validation of the gene expression data and help refine gene models. The enhanced gene models can help improve protein sequence databases for traditional proteomic analysis.

Adapted from Nesvizhskii, 2014

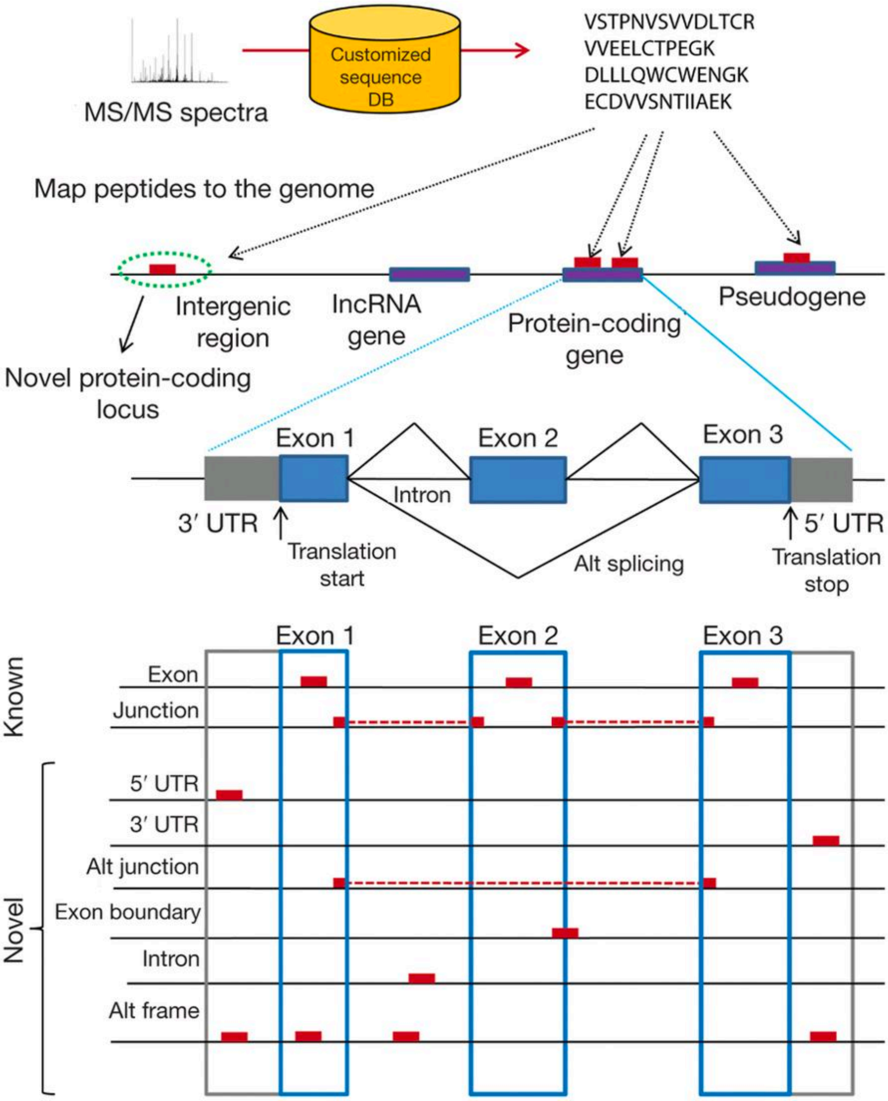


Figure B-7. Type of peptides identified in proteogenomics.

Peptides identified by searching customized protein sequence databases are mapped on the genome. Intergenic peptides map to regions located between annotated gene models, whereas intragenic peptides map to genomic regions contained within or in close proximity to an annotated gene model. Intragenic peptides can be further categorized according to the annotation of the corresponding gene model. The majority of peptide map to a protein-coding gene and can be divided into exon and exon-exon junction peptides. Novel peptides include peptides mapping to untranslated regions, intron peptides, peptides spanning the boundary between the coding sequence region and the neighboring UTR or intron region, peptides spanning alternative splice junctions, and out-of-frame peptides.

Adapted from Nesvizhskii, 2014

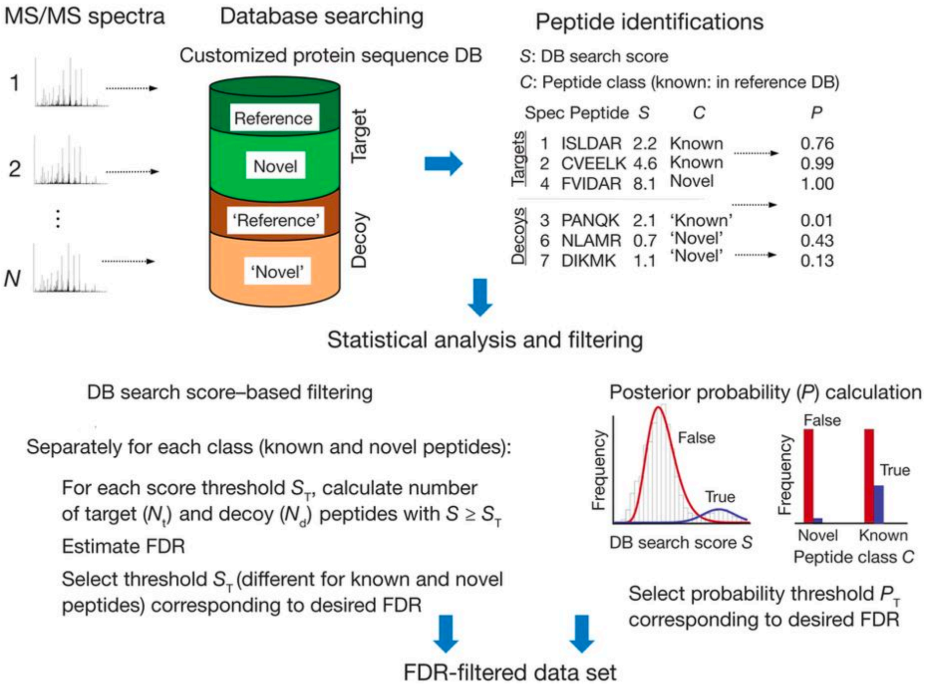

MS/MS spectra
1
2
⋮
N
Database searching
Customized protein sequence DB
Reference
Novel
'Reference'
'Novel'
Target
Decoy
Peptide identifications
S: DB search score
C: Peptide class (known: in reference DB)
Spec Peptide S C P
Targets
1 ISLDAR 2.2 Known 0.76
2 CVEELK 4.6 Known 0.99
4 FVIDAR 8.1 Novel 1.00
Decoys
3 PANQK 2.1 'Known' 0.01
6 NLAMR 0.7 'Novel' 0.43
7 DIKMK 1.1 'Novel' 0.13
Statistical analysis and filtering
DB search score–based filtering
Separately for each class (known and novel peptides):
For each score threshold S_T, calculate number of target (N_t) and decoy (N_d) peptides with $S \geq S_T$
Estimate FDR
Select threshold S_T (different for known and novel peptides) corresponding to desired FDR
Posterior probability (P) calculation
Frequency
False
True
DB search score S
Frequency
False
True
Novel
Known
Peptide class C
Select probability threshold P_T corresponding to desired FDR
FDR-filtered data set

Figure B-8. Statistical assessment of peptide identifications in proteogenomics.

MS/MS spectra are searched against a customized protein sequence database that includes target sequences for the organism of interest, i.e., a reference protein database and predicted protein sequences (containing novel peptides). In addition, two 'decoy' databases of the same sizes as the target reference and predicted databases are appended to the target databases. The best database peptide match for each spectrum is selected for further analysis. Peptide identifications are classified as known or novel. When simple database search score–based filtering is used, the numbers of target and decoy peptide identifications passing a certain score threshold are counted and used to estimate the FDR corresponding to that threshold. FDR analysis should be done separately for known and novel peptides (class-specific FDR) because of differences in the number of known and novel sequences in the searched customized sequence database and because of the lower likelihood of correctly identifying a novel peptide. For more advanced methods based on computing posterior peptide probabilities, both the database search scores and the peptide class (known or novel) should be taken into consideration.

Adapted from Nesvizhskii, 2014

The advancement of proteogenomics is best shown with research on cancer where tremendous advancement has been achieved. The progression of cancer is carried out by alteration in the genome and uncertainty that occurs because of the cascade of genomic variation that which includes mutation, copy number aberration or translocation and methylation (Hanahan and Weinberg, 2011). However, as massive cancer genome sequencing projects (Consortium, 2010, Weinstein et al., 2013) were gradually developed, it is now obvious that the association of cancer genotype and phenotype also needs a cancer proteotype description. The development of in high-throughput proteomics has enabled a consistent and significant methodology with the abilities to compare the genomics for blood samples and tumor analysis. The National Cancer Institute, in this context, has launched, in 2011, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Ellis et al., 2013).

A major development in CPTAC has been shown recently in an important study (Zhang et al., 2014) that described the proteogenomics of human colorectal cancer. This investigation evaluated five proteomics subtypes in The Cancer Genome Atlas cohort for colorectal cancer. Intriguingly, the 20q chromosome amplicon was linked to the prime universal alteration at the mRNA and proteins levels and data from proteomics showed 20q candidate targets and biomarkers for colorectal cancer therapy. The results also showed that the mRNA transcript quantity did not make a reliable prediction about protein quantity variance between cancers. It is noteworthy that the cancer taxonomy generated thus far totally depends on genomics and transcriptomics examination.

Although the proteogenomics approach shows novel progress in omics technology, with the expected results applied results in medicine and biology, it is in the initial phase, and further progress is required before it can be broadly adopted. More bioinformatics assimilation is needed to entirely use the complete information spectra acquired in genomics, transcriptomics and proteomics studies (Sajjad et al., 2016).

5. TARGETED PROTEOMICS

To verify potential protein biomarkers, it is necessary to build an adequate quantitative assay platform. There are two main ways to detect and quantify proteins: affinity reagent-based methods, exemplified by ELISA, Western blotting or immunohistochemistry staining, and MS-based peptide identification and quantification, which are mainly used for research and discovery proteomics. However, the dynamic range and number of proteins quantifiable using affinity reagent-based assays are limited (Leng et al., 2008, Ebhardt et al., 2015).

MRM is a mass spectrometry technique for the detection and quantification of specific, predetermined analytes with known fragmentation properties in complex backgrounds. MRM is used most effectively in an LC-MS system (Picotti and Aebersold, 2012), where a capillary chromatography column is connected in-line to the electrospray ionization source of the mass spectrometer. MRM exploits the unique capability of triple quadrupole (QqQ) mass spectrometers to act as mass filters and to selectively monitor a specific analyte molecular ion and one or several fragment ions generated from the analyte by collisional dissociation (Figure B-9) (Yost and Enke, 1978, Yost and Enke, 1979, Kondrat et al., 1978, Glish and Vachet, 2003).

Molecular ions within a mass range centered around the mass of the targeted peptide are selected in the first mass analyzer (Q1) and are fragmented at the peptide bonds by collision-activated dissociation (in Q2), and then one or several of the fragment ions uniquely derived from the targeted peptide are measured by the second analyzer (Q3) (Kuhn et al., 2004, Lange et al., 2008). Integration of the chromatographic peaks for each transition supports the relative or, if suitable heavy isotope-labeled reference standards are used, absolute quantification of the targeted peptide initially released from the protein and loaded on the LC-MS system. A suitably chosen set of MRM transitions therefore constitutes a specific assay to detect and quantify a target peptide and, by inference, a target protein in complex samples.

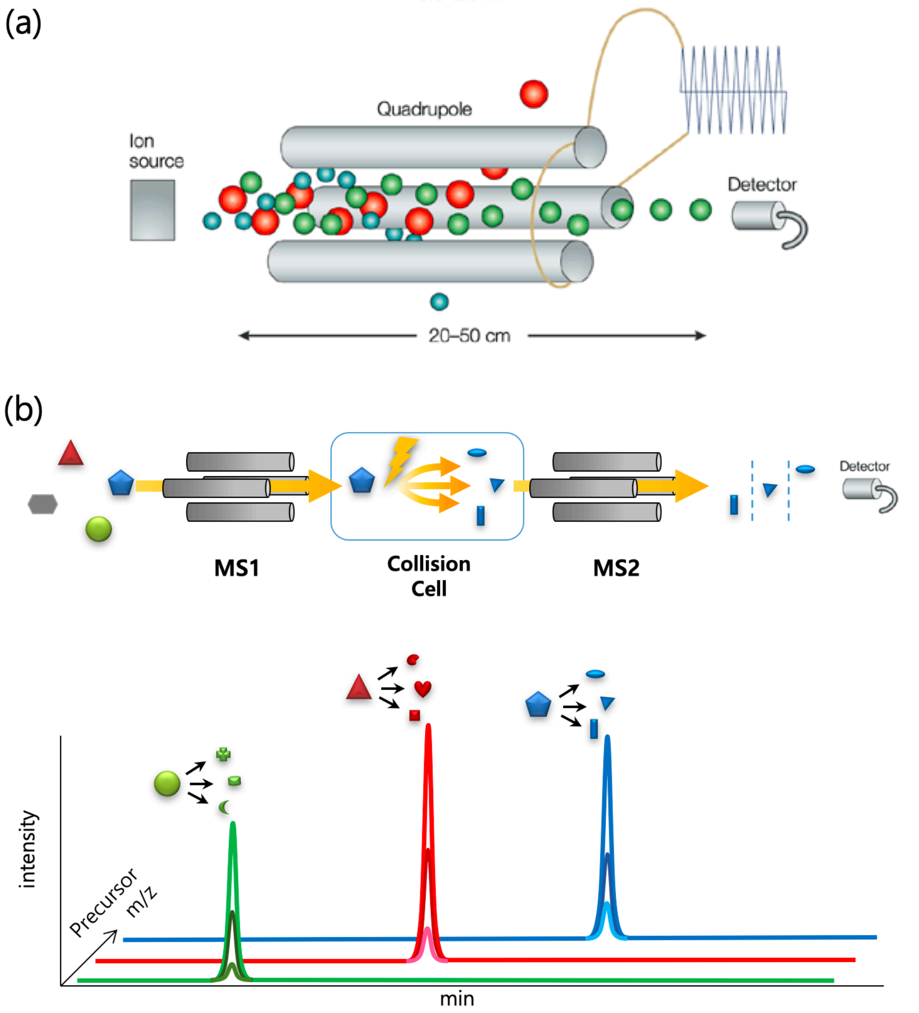


Figure B-9. Pictorial diagrams of the quadrupole during multiple reaction monitoring (MRM).

(a) In a quadrupole mass analyzer, the correct magnitude of the radio frequency and direct current voltages applied to the rods allows ions of a single m/z to maintain stable trajectories from the ion source to the detector, whereas ions with different m/z values are unable to maintain stable trajectories. (b) In the triple quadrupole (QqQ) instrument, which are suitable machine for MRM, the first and third quadrupoles are operated as mass spectrometers, whereas the second (middle) quadrupole acts as the collision region for collision induced dissociation (CID). Through HPLC eluting time, QqQ monitoring the fragmented ions and quantifying multiplexed targets.

Adopted from Glish et al., 2003

A targeted proteomics experiment consists of multiple steps: first, the generation of a hypothesis, a target list of proteins to test the hypothesis and a fit-for-purpose quantitation strategy; second, the study design and experimental planning; third, sample preparation; fourth, method refinement; fifth, data acquisition; and, finally, analysis and modeling (Gillette and Carr, 2013b, Picotti and Aebersold, 2012, Liebler and Zimmerman, 2013). Bioinformatics and computational proteomics are part of each step of the workflow (Figure B-10).

The findings of promising biomarkers or signatures from preclinical studies should be followed by clinical testing. The bulk of recent work has focused on candidate discovery and overcoming the associated challenges related to tissues and bodily fluids. Targeted proteomics can be used to validate biomarkers found in a project's discovery phase across many patient samples with high accuracy and reproducibility.

In 2006 the National Cancer Institute (USA) started the CPTAC to evaluate targeted and discovery technologies for quantitative analysis in tissues and biofluids. This program was renewed in 2011 as the CPTAC, which began focusing on applications (Ellis et al., 2013). CPTAC member laboratories applied the standardized methods of MRM and demonstrated reproducibility, precision, and sensitive quantitation in tissues and biofluids (Cox et al., 2014).

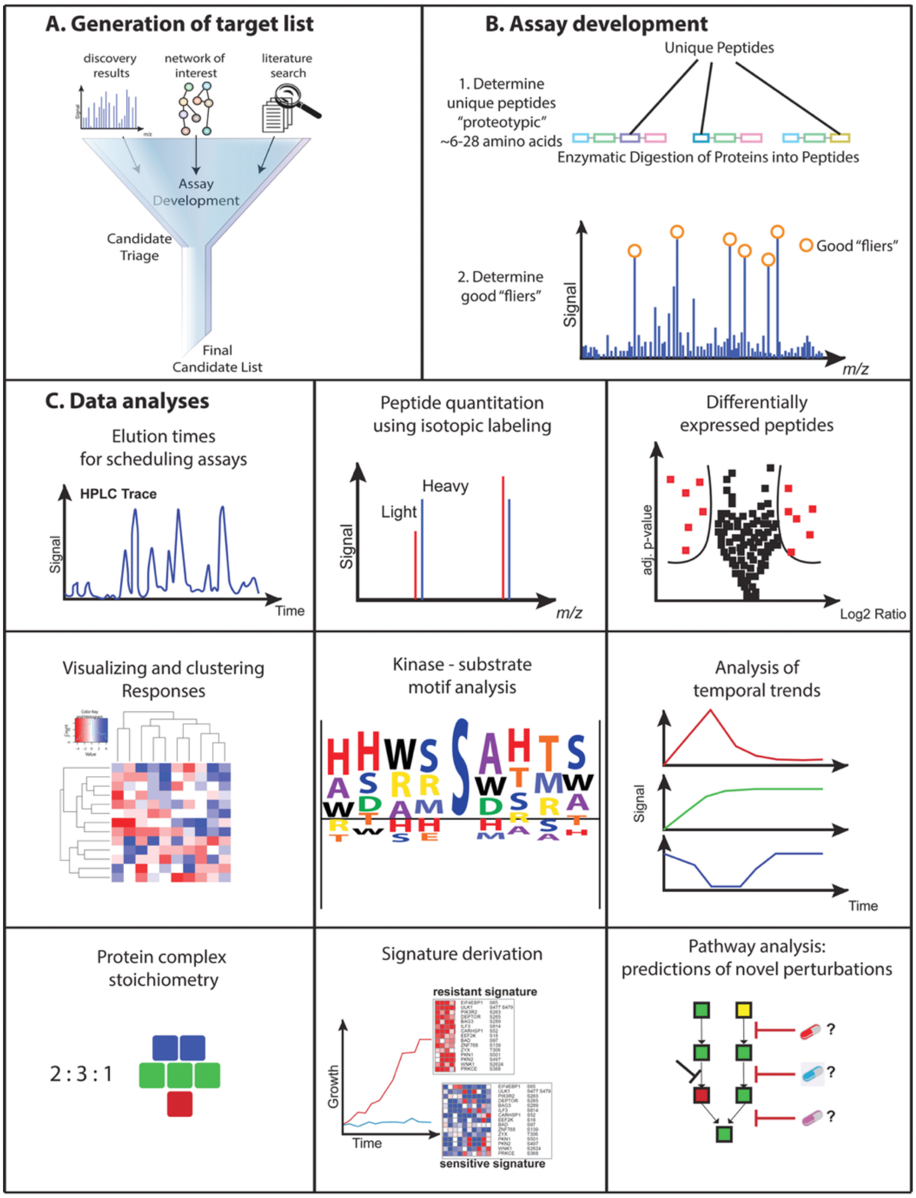

A. Generation of target list
discovery results
network of interest
literature search
Signal
m/z
Assay Development
Candidate Triage
Final Candidate List
B. Assay development
Unique Peptides
1. Determine unique peptides "proteotypic" ~6-28 amino acids
Enzymatic Digestion of Proteins into Peptides
2. Determine good "fliers"
Good "fliers"
Signal
m/z
C. Data analyses
Elution times for scheduling assays
HPLC Trace
Signal
Time
Peptide quantitation using isotopic labeling
Heavy
Light
Signal
m/z
Differentially expressed peptides
adj. p-value
Log2 Ratio
Visualizing and clustering Responses
Kinase - substrate motif analysis
Analysis of temporal trends
Signal
Time
Protein complex stoichiometry
2 : 3 : 1
Signature derivation
resistant signature
Growth
Time
sensitive signature
Pathway analysis: predictions of novel perturbations
?
?
?

Figure B-10. Typical targeted proteomics workflow.

(A) Discovery results from LC-MS/MS experiments, protein network modeling and literature search typically form the basis to generate the final candidate list to be quantified by MRM. (B) MRM assays for peptides are generated from extensive LC-MS/MS experiments under consideration of proteotypic peptides generated and best performing transitions per peptide. (C) Data anlysis starts with the primary LC-MS/MS performance examination. If spiked in, stable isotope labeled peptides serve as reference for consistent quantification. Statistical analysis of peptides quantified serve to identify peptides, and therefore proteins, changing in abundance. Further analysis include the clustering of data corresponding to proteins quantified and condition. If multiple kinase substrates were quantified, a consensus motif analysis could identify novel substrate motifs of a kinase. In case the conditions are time course data, the abundance of proteins can be plotted as a function of time. Using MRM-MS, protein stoichiometry of purified protein complexes can be determined (to be precise, this method requires newly synthesized externally calibrated reference peptides). The quantification of proteins and together with sample knowledge integration might lead to signatures which protein signature results in resistant or sensitive samples. The ultimate analysis is the protein network analysis leading to the prediction of novel perturbations.

Adapted from Ebhardt et al., 2015

Portions of this chapter were published as:

Kim YI, Cho JY. (2019) Gel-based proteomics in disease research: Is it still valuable? *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics.* **1867**(1):9-16.

CHAPTER I

Proteogenomic Study:

Variant Proteome and Transcriptome in Human Lung Adenocarcinoma Tissues

INTRODUCTION

The Chromosome-Centric Human Proteome Project (CHPP), in which teams focus on individual chromosomes, is a global initiative that characterizes proteins encoded by genes across the complete human genome. (Paik et al., 2012b) The proposed goals of C-HPP include the identification of all proteins encoded by each protein coding gene and the characterization of their localization, alternate splice variants, nonsynonymous variant-containing peptides, and major posttranslational modifications, including phosphorylation, glycosylation and acetylation, if any, by mass spectrometry (MS) and antibody-based methods. (Paik et al., 2012a)

Chromosome 9 (Chr 9) DNA is approximately 145 megabases in length. A total of 1,467 genes have been annotated from Chr 9 (GENCODE release 19), including genes implicated in male-to-female sex reversal, cancers and neurodegenerative diseases, along with 426 pseudogenes. (Humphray et al., 2004) Among these genes, 815 genes are currently classified as protein-coding genes. It has been reported that mutations of the genes on Chr 9 are associated with various types of cancer, including lung cancer. (Aravidis et al., 2012, Dagher et al., 2013, Narayanan et al., 2013)

When searching spectra with established annotated reference protein databases, it is not possible to evaluate many characteristics, including splicing variants, missense mutation, 5’ and 3’ UTR translated peptides, and pseudogene expression. Therefore, to identify proteins in wide, uncharacterized sections of the human genome, a customized database, such as a 6-frame open reading frame database or a sample-specific RNA-seq-origin protein database, is required. Several research groups have published disease specific novel protein identifications that were determined using proteogenomics approaches and custom databases. However, this is not a generalized analytic tool and it needs to be applied to various diseases. (Sheynkman et al., 2013, Zhang et al., 2014)

Previous has focused on the study of lung cancer biomarkers, as lung cancer still presents the highest mortality rate among all cancer-related deaths. (Siegel et al., 2015) However, there is still a lack of information regarding novel proteins or variant-based biomarkers and their pathogenesis related to lung cancer.

It has been shown that only approximately 16% of risk-associated loci harbor SNPs that affect coding sequences, and this absolves the majority of risk-associated loci from alterations to protein sequence. (Schaub et al., 2012) These synonymous, or silent, mutations are categorized as passenger events because they do not modify the protein sequence. They are therefore considered functionally irrelevant, although their potential involvement in tumorigenesis has been suspected. Recently, a compelling analysis suggested that such silent mutations can become oncogenic by altering transcript splicing and thereby affecting protein function. (Gartner et al., 2013, Supek et al., 2014, Sauna and Kimchi-Sarfaty, 2011) In this study, I performed RNA sequencing (RNA-seq) and MS-based proteomics analyses on lung adenocarcinoma (ADC) tissues and adjacent normal tissues. I identified 19 missing proteins across all chromosomes, tumor-specific 3 synonymous and 4 nonsynonymous SNPs at transcript level, 7 missense mutants at both protein and transcript levels, and an expressed pseudogene protein. For missense mutant identification, I searched RNA-seq and proteomics data after sample-specific protein database generation.

MATERIALS AND METHODS

Tissue samples

A total of 10 lung tissue specimens, including paired tumor and adjacent normal tissues from 5 patients with lung ADC, were obtained from the Samsung Medical Center Biobank (Seoul, Korea). Age and sex of the samples are 58 year male (ID11001563), 79 year male (ID13001628), 56 year male (ID11001794), 35 year female (ID10000306), and 74 year male (ID10004557). TNM stages were T4N2M0 in one patient (ID10004557) and T2N2M0 in remaining four patients. All samples were obtained and used in accordance with the study protocol approved by the Institutional Review Board at the Samsung Medical Center (Seoul, Korea) (IRB No. 2012-11-025-003). The tissue samples were obtained from dissected surgical specimens that were snap-frozen in liquid nitrogen immediately after surgery, with written informed consent from each patient. The specimens were kept at -80°C until RNA or protein extraction.

RNA-sequencing analysis

Total RNA was extracted from both adjacent normal and tumor tissues, and their quality was evaluated using an Agilent RNA 6000 Nano kit (Agilent Technologies, CA, USA). The cytoplasmic ribosomal RNA (rRNA) was depleted, and the cDNAs were subjected to end-repair and poly-(A) addition and then connected with sequencing adapters using the TruSeq RNA sample prep Kit (Illumina, CA, USA). The libraries from both tumor and adjacent normal tissues were sequenced by an Illumina HiSeq2500 sequencer (Illumina, CA, USA). RNA-seq data were aligned with the GENCODE human reference genome (release19, GRCh37.p13) and matched to Ensemble (release74) using STAR (version 2.4.0j). (Dobin et al., 2013) Abundance of both transcripts and genes was measured as Fragments Per Kilobase of exon per Million fragments mapped (FPKM). To detect genomic variants, aligned RNA-seq BAM files were processed by RVboost (version 0.1). (Wang et al., 2014)

Sample preparation for proteomics

For liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis, proteins were extracted from homogenized tissue samples using a NucleoSpin TriPrep kit (Macherey-Nagel GMH & Co KG, Düren, Germany) according to the manufacturer’s instructions. The protein concentration of diluted tissues was measured using Quick Start™ Bradford 1x Dye Reagent

(Bio-Rad Laboratories, Hercules, CA). Then, 80 μg of protein was prepared in 50 mM ammonium bicarbonate and reduced and alkylated by treatment with dithiothreitol (Bio-Rad Laboratories, Hercules, CA) and iodoacetamide (Sigma-Aldrich, St Louis, MO, USA). Trypsin (Promega, Madison, WI, USA) was added to digest samples at a protein-to-enzyme ratio of 50:1 (w/w), and the solution was incubated at 37 °C for 16 hours. Digested samples were separated into 10-12 fractions using high pH on a C18 column as first dimension.

LC-MS/MS analysis

Spectra raw data were acquired on an LTQ-Orbitrap (Thermo Fisher, San Jose, CA) with EASY-nLC II (Proxeon Biosystems, now Thermo Fisher Scientific). An autosampler was used to load 6-μL aliquots of the peptide solutions into an EASY-Column; C_{18} Trap-column of i.d. 100 μm, length 20 mm, and particle size of 5 μm (Thermo Scientific). The peptides were desalted and concentrated on the trap column for 15 min at a flow rate of 2 μL/min. Then, the trapped peptides were separated on an EASY-Column; C_{18} analytic-column of i.d. 75 μm and length 100 mm, and 3 μm particle size (120 Å from Thermo Scientific). The mobile phases were composed of 100% water (A) and 100% acetonitrile (ACN) (B), and each contained 0.1% formic acid. The voltage applied to produce the electrospray was 2.0 kV. During the chromatographic separation, the LTQ-Orbitrap was operated in a data-dependent acquisition mode. The MS data were acquired using the following parameters; full scans were acquired in Orbitrap at resolution 60,000 for each MS/MS measurement, six data-dependent collision induced dissociation (CID) MS/MS scans, CID scans were acquired in linear trap quadrupole (LTQ) with 10 ms activation time performed for each sample, 35% normalized collision energy (NCE) in CID, ±1.5 Da isolation window. Previously fragmented ions were excluded for 180 sec.

Then, the datasets generated by LTQ-Orbitrap were analyzed using the Proteome Discoverer (version 1.3.0.339, Thermo Fisher Scientific) and Scaffold (version 4.4.1, Proteome Software Inc., Portland, OR) Platform, and searched against the UniProt human protein database (release 2015_02) using SEQUEST and X!tandem. Peptide identifications were accepted if they could be established at a greater probability to achieve an FDR less than 1.0% by a Scaffold Local FDR algorithm. Protein identifications were accepted if they could be established at a greater than 95.0% probability and if they contained at least 2 identified unique peptides. Protein probabilities were assigned by the Protein Prophet algorithm. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (Vizcaíno et al., 2012) with the dataset identifier PXD002523 and DOI 10.6019/PXD002523.

Proteogenomic data analysis

To integrate the RNA-seq and LC-MS/MS datasets, I customized a sequence database, composed of SNP-containing or ORF-excluding transcripts, for each sample pair (Figure 1-1). SNP sequence data from individual patients was also processed by QUILTS (http://quilts.fenyolab.org/) to convert it to amino acid sequence for somatic non synonymous variant identification. Converted variant sequence and ensembl_human_37.70.fasta were combined to generate a working database, which was used for exclusive unique MS/MS spectrum searching for missense variation in peptides. The customized fasta format database is available from public data repository platform GitHub located at https://github.com/vetbio/2015jpr.

I also generated other databases from identified transcripts above FPKM 2 to find pseudogenes or novel proteins/peptides that contain long non-coding transcripts and splicing variants. These databases contain pseudogenes or long non-coding transcripts in the introns of coding genes, long non-coding transcripts that contain a possible coding gene in the intron on the same strand, short non-coding transcripts from the 3'UTR, and processed transcripts that do not contain an ORF. Nucleotide to amino acid sequence conversions were conducted by 3-frame translation and converted sequences were combined with human Fasta (UniProt 2015-02 rel.). MS/MS spectra searches were processed using the same platform mentioned above to identify proteins via proteogenomic data analysis with an estimated peptide FDR threshold of 1%, a protein probability threshold of 95%, and containing at least 1 unique peptide and with manual confirmation.

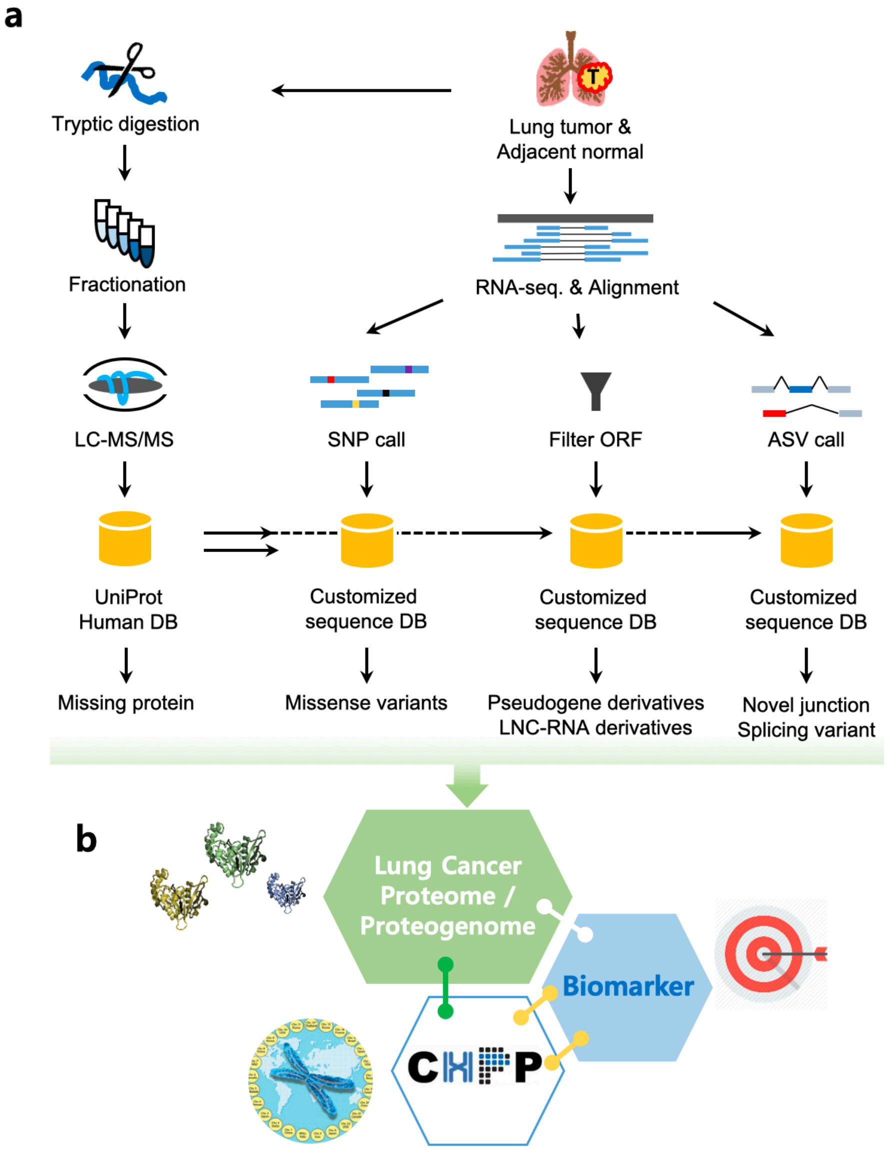


Figure 1-1. Schematic diagram of this study

(a) LC-MS/MS analysis was performed for proteome identification, and RNA-seq analysis was performed for customized database generation using a proteogenomic data acquisition process. To identify new proteins/peptides and transcript and protein mutants, customized databases composed of SNP- and ASV including or ORF-excluding transcripts were used. (b) Identified both proteome and proteogenome were used for C-HPP datamatrics to fill the gap towards the genome; and also utilized for the lung cancer biomarker discovery with multi-omics technologies.

RESULTS AND DISCUSSION

Recent updates to the chromosome 9 proteome

First, I revisited the neXtProt database and compared the protein lists on Chr 9 between neXtProt 2013.09.26 rel. (just prior to the submission of Chr 9 working group's last C-HPP special issue) and neXtProt 2015.04.28 rel. The total number of genes and proteins on Chr 9 was decreased by 11; from 826 to 815 for genes and from 821 to 810 for proteins (Figure 1-2a). The difference (5) in the number of genes and proteins is due to 4 overlapping genes: Interferon alpha-1/13 (Gene: Ifna13, Ifna1), Protein FAM74A1/A2 (Gene: Fam74a1, Fam74a2), Protein FAM74A4/A6 (Gene: Fam74a4, Fam74a6), and Protein FAM27A/B/C (Gene: Fam27a, Fam27b, Fam27c). The number of missing proteins was decreased from 170 to 133. Thus, in the last one and a half years, 39 missing proteins were identified, 2 missing proteins were downgraded to PE5, and 2 were eliminated from the neXtProt index. Meanwhile, 6 proteins that were demoted from PE1 to PE2 were newly added to the missing protein list (Figure 1-2b). The newly identified 39 missing proteins were contributed mostly from previous JPR issue (Ahn et al., 2013) and from two publications in Nature by the Pandey and Kuster groups in May 2014. (Kim et al., 2014, Wilhelm et al., 2014) These two Nature publications received a great attention because the authors showed a large amount of human proteome coverage in various tissues. However, they were criticized because the protein identification reliability used for protein identification was low. (Ezkurdia et al., 2014) This low reliability was shown by the high number of identified olfactory receptors despite the fact that there was no tissue included in the study where the expression of those proteins is to be expected. Thereafter, the C-HPP consortium set up strict protein identification criteria. Henceforth, to be identified as a missing protein, the peptide FDR should be less than 1%, two or more unique peptides should be identified, and the protein probability should be more than 95%. In the C-HPP previous special issue, I originally identified 45 missing proteins in Chr 9. However, after applying these strict criteria, only 16 out of 45 of these proteins were found to be valid, newly identified, missing proteins.

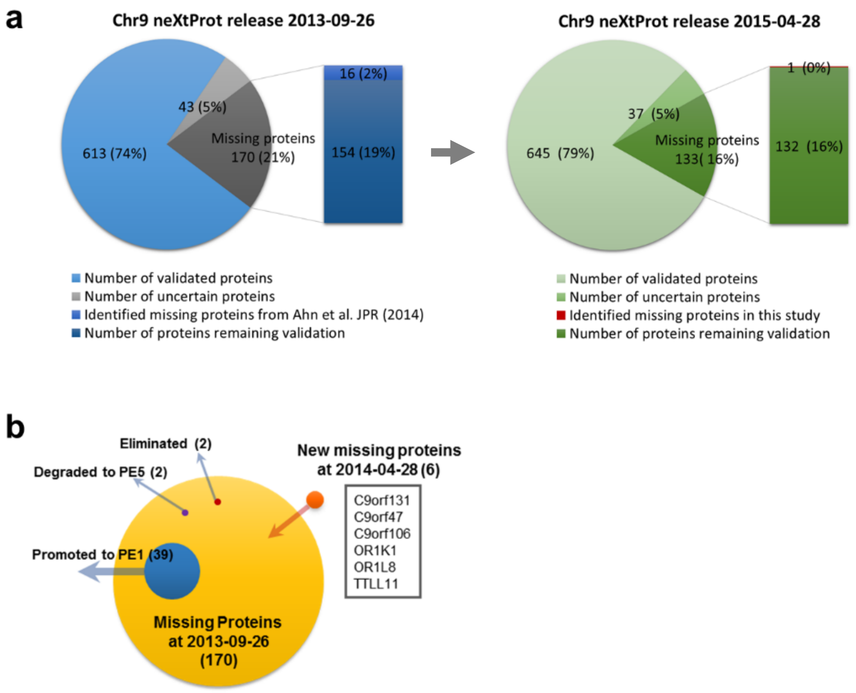


Figure 1-2. Recent updates to the chromosome 9 proteome

(a) Pie charts show Chr 9-encoded proteome statistics between neXtProt 2013.09.26 rel. and neXtProt 2015.04.28 rel. **(b)** Trends of Chr 9-encoded missing proteins. In the neXtProt 2014.04.28 rel., 43 proteins were expelled from the 2013.09.26 rel. and 6 proteins were newly added from PE1 to the missing protein list.

Newly detected missing proteins

To find new missing proteins, I performed LC-MS/MS proteomics analysis on 5 pairs of human lung ADC tissues and adjacent normal tissues. Proteome data were searched using the database from neXtProt release 2015-04-28. In the Protein Evidence (PE) 2-4 category, I identified 19 new missing proteins with the criteria of peptide FDR <1%, unique peptide ≥2 and protein probability >95% (Table 1-1). Out of 19 missing proteins, 13 were found only in lung ADC tissues (AKR7L, PHTF1, DNAH6, CTAGE9, TEX15, SPATA31A4, LRRC27, CCDC38, DNAH3, FBXW10, ZNF221, ZNF780A, BHLHB9), whereas only 4 were found only in normal tissues (ANKRD36, ZNF619, SLC35A4, CCDC178). One missing protein, SPATA31A4, was a Chr 9 gene-coded protein that was found in one of the ADC tissue samples. The peptide identified in the protein had a mis-cleaved arginine in the middle of its sequence. Thus, the CID may generate limited backbone cleavage and may usually produce poor tandem spectra. This may explain, in part, why this peptide was difficult to identify. Electron transfer dissociation (ETD) may help to increase the quality of the tandem spectra, if used. (Snijders et al., 2010)

From the missing protein identification data so far, and because different tissues express different protein sets, it is important to use in-depth analysis to analyze proteomes in diverse tissues to find the remaining missing proteins. In addition, using digestion enzymes other than trypsin, or using enzymes in combination with trypsin, such as Glu-C, chymotrypsin, or Lys-C proteases, may also be used to increase the sequence coverage and thus chance to detect peptides that holds splicing variants or specific isoforms for better proteome identification. Alternative fragmentation methods, such as ETD and ultraviolet photo dissociation (UVPD), will also improve the identification of missing proteins. (Tsiatsiani and Heck, 2015, Greer et al., 2015, Hendricks et al., 2014, Vasicek and Brodbelt, 2010, Yoon et al., 2009)

Table 1-1. List of newly detected missing proteins

neXtProt ID	Gene	Chr	PE	Description	Identified Sample	
					Normal	Tumor
NX_Q8NHP1	AKR7L	1	transcript level	Aflatoxin B1 aldehyde reductase member 4		10004557_T
NX_Q9UMS5	PHTF1	1	transcript level	Putative homeodomain transcription factor 1		11001563_T*
NX_A6QL64	ANKRD36	2	transcript level	Ankyrin repeat domain-containing protein 36A	11001563_N*	
NX_Q9C0G6	DNAH6	2	transcript level	Dynein heavy chain 6, axonemal		13001628_T
NX_Q8N2I2	ZNF619	3	transcript level	Zinc finger protein 619	10004557_N*	
NX_Q96G79	SLC35A4	5	transcript level	Probable UDP-sugar transporter protein SLC35A4	10004557_N*	
NX_Q5TEZ5	C6orf163	6	predicted	Uncharacterized protein C6orf163	11001794_N	11001563_T
NX_A4FU28	CTAGE9	6	transcript level	cTAGE family member 9		11001563_T
NX_Q9BXT5	TEX15	8	transcript level	Testis-expressed sequence 15 protein		11001563_T
NX_Q4VX67	SPATA31A4	9	homology	Spermatogenesis-associated protein 31A4		10000306_T

NX_Q9C0I9	LRRC27	10	transcript level	Leucine-rich repeat-containing protein 27		11001794_T*
NX_Q502W7	CCDC38	12	transcript level	Coiled-coil domain-containing protein 38		10004557_T
NX_Q8TD57	DNAH3	16	transcript level	Dynein heavy chain 3, axonemal		11001563_T
NX_Q5XX13	FBXW10	17	transcript level	F-box/WD repeat-containing protein 10		11001794_T
NX_Q5BJE1	CCDC178	18	transcript level	Coiled-coil domain-containing protein 178	11001563_N*	
NX_Q9UK13	ZNF221	19	transcript level	Zinc finger protein 221		11001563_T
NX_O75290	ZNF780A	19	transcript level	Zinc finger protein 780A		11001563_T*
NX_Q6PI77	BHLHB9	X	transcript level	Protein BHLHb9		11001563_T
NX_Q96LI9	CXorf58	X	transcript level	Putative uncharacterized protein CXorf58	11001563_N	

*RNA-seq analysis supported with FPKM>1

Abbreviation used: Chr, chromosome; PE, protein existence level

Proteome and transcriptome analysis of lung cancer tissues

In the LC-MS/MS proteomic analysis of the 5 pairs of ADC and adjacent normal tissues, 1257 proteins were identified (Figure 1-3e). To classify biological function, proteins and genes were clustered according to gene family, as annotated by the Human Genome Nomenclature Committee (HGNC; http://www.genenames.org/genefamily.html) (Figure 1-3f). The gene family with the most proteins and transcripts expressed was the Lipocalin family, which exhibits great functional diversity, with roles in the regulation of cell homeostasis and the modulation of the immune response, transporter, prostaglandin synthesis and as carrier proteins. (Flower, 1996) Five of the proteins and eight of the transcripts showed 100% coverage.

A total of 87946 transcripts from 18145 genes were identified in the RNA-seq analysis, using a cutoff of FPKM>1 for genes and >0.3 for transcripts in at least one specimen (Figure 1-3a & 3b). 676 genes (46% of 1467 genes) on chromosome 9 were detected, including 24 genes (18%) out of 133 missing protein genes (Figure 1-3a, 3c & 3d). Cancer tissues expressed more genes, but proportions of protein coding genes (~85%) and others were similar to normal tissues (Figure 1-3a). The number of unique genes found only in cancer tissues was 2206 (12%), whereas 14764 (81 %) of the expressed genes overlapped in cancer and normal tissues. The gene expression levels varied greatly across different tissues and specimens. A total of 11024 genes (60 %) were identified in all 10 specimens at the transcript level. However, 1784 genes (9 %) were detected only in one specimen, and 4547 genes (25%) were detected in less than half of the specimens.

As shown in Figure 1-3d, the RNA abundance of 24 chromosome 9-encoded missing proteins was shifted toward lower FPKM values. No corresponding protein was detected for these 24 genes in the proteome analysis. The missing protein, SPATA31A4, which was identified by proteomics, was not detected by RNA-seq analysis at an FPKM>1 level. Meanwhile, 7 (37%) out of 19 new missing proteins detected by MS had an FPKM>1. These findings indicate the extreme difficulty encountered when attempting to detect the remaining missing proteins using current proteomics techniques in one sample type of tissue.

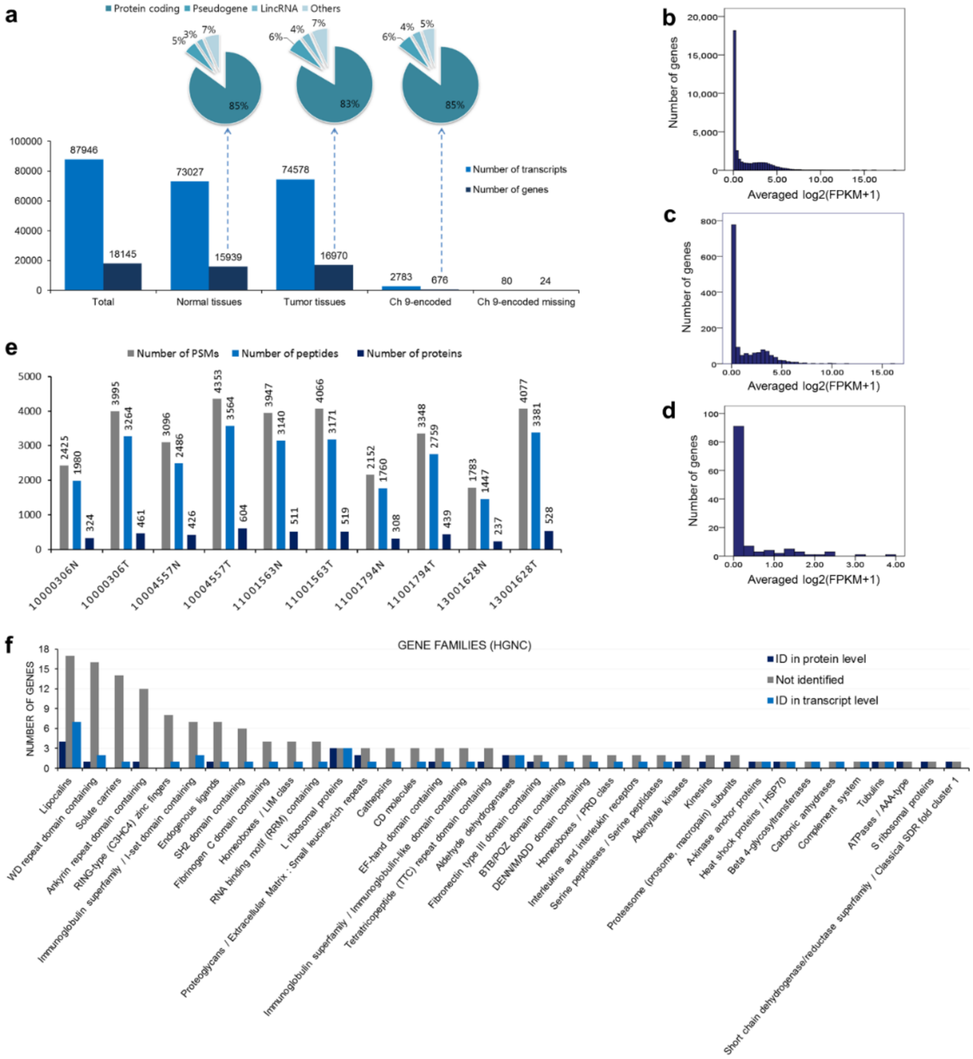


Figure 1-3. Categorization of the identified transcriptome and proteome in lung cancers

(a) Number of genes and transcripts identified in RNA-seq analysis. Transcript abundance distribution in total **(b)**, chromosome 9-encoded **(c)**, and missing chromosome 9-encoded genes **(d).** **(e)** A comparison of the number of peptide spectrum matches (PSMs), peptides, and proteins identified in LC-MS/MS analysis from 5 pairs of lung tissues. (**f)** Transcriptome and proteome expression enriched in major gene families encoded by Chr 9.

Synonymous and nonsynonymous SNPs in lung cancer tissues

Using the RVboost tool, 52910~63453 variants in five tumor and 56404~60680 variants were identified in five adjacent normal tissues, respectively. To extract tumor-specific somatic SNPs, variants derived from adjacent normal tissues were eliminated from the list of variants found in tumor tissues. As a result, 18263~23496 tumor-specific variants were acquired from 5 pairs of samples. The total number of enriched tumor-specific variants is 91985 in overall samples, containing 3172 chromosome 9-encoded variants (including 358 synonymous and 509 nonsynonymous mutations), which were distributed 625~821 per sample. Four nonsynonymous variants, in CDH17, HIST1H1T, SAPCD2, and ZNF695, and four synonymous variants, in CDH17, CST1, HNF1A, and CSMD2, were identified as common to all 5 tumor samples but none of normal adjacent tumors (Table 1-2).

Table 1-2. Identified tumor specific synonymous and nonsynonymous mutations in RNA-seq analysis

Chr	Gene name	Position	Ref	Alt	Codon change	Amino acid change	Effect	Exon ID
1	CSMD2	34285381	C	T	aaG/aaA	K379	Synonymous	NM_052896.ex.62
8	CDH17	95158382	C	T	ttG/ttA	L647	Synonymous	NM_001144663.ex.4
12	HNF1A	121435342	C	T	Ctg/Ttg	L459	Synonymous	NM_000545.ex.7
20	CST1	23729722	G	T	cgC/cgA	R91	Synonymous	NM_001898.ex.2
1	ZNF695	247162658	C	T	aGa/aAa	R84K	Nonsynonymous	NM_001204221.ex.4
6	HIST1H1T	26108168	G	A	Ctt/Ttt	L52F	Nonsynonymous	NM_005323.ex.1
8	CDH17	95143186	C	G	gaG/gaC	E734D	Nonsynonymous	NM_001144663.ex.3
9	SAPCD2	139964447	G	C	Cgc/Ggc	R156G	Nonsynonymous	NM_178448.ex.6

Abbreviation used: Chr, chromosome; Ref, reference nucleotide; Alt, altered nucleotide

The data was further analyzed to determine whether these synonymous mutations have any correlation with transcript expression levels. Synonymous mutations, which occur as a single nucleotide substitution in an exon region, do not modify coding amino acids and are considered functionally silent mutations. However, the possibility that a synonymous SNP might modulate transcriptional processes or induce tumorigenesis has been proposed. (Sheynkman et al., 2013, Narayanan et al., 2013, Supek et al., 2014) Four synonymous mutants that were common to all tumor samples were enriched, and three of these genes were identified as having transcript levels with an FPKM value greater than 0.3, with the exception of CSMD2 (Figure 4). Interestingly, the transcripts of CDH17, CST1, and HNF1A were detected only in tumor tissues and not in normal tissues. These data suggest that the synonymous SNPs at CDH17, CST1, and HNF1A in tumor tissues resulted in the selective enrichment of the mutant genes in these tissues and, therefore, to higher expression in tumor tissues than in adjacent normal tissues. Future study is needed to clarify the mechanisms involved in this selective expression of synonymous mutants in lung ADC.

Expression of CDH17, also known as an adhesion molecule, is restricted to the colon, intestine, and pancreas in adult humans. However, it has been reported to be reduced in colorectal and pancreatic cancers and induced in liver and gastric cancers. (Lee et al., 2010b) CDH17 is known as a liver-intestine cadherin and an onco-fetal gene and as a potential biomarker of hepatocellular carcinoma and gastric cancer. (Wong et al., 2003, Lee et al., 2010a) However, its synonymous SNP variant selectivity was not reported in lung cancers prior to this study. CST1 is expressed in the lacrimal gland, gall bladder and seminal vesicle in adult humans and is upregulated in the fetal submandibular gland. (Dickinson et al., 2002) CST1 regulates cysteine protease activity and is highly involved in gastric cancer and colorectal cancer. (Yoneda et al., 2009, Choi et al., 2009) HNF1A is a transcriptional activator that functions in the tissue specific expression of multiple genes, especially in pancreatic β-cells, liver and other tissues. (Ryffel, 2001) A frameshift mutation in exon 4 of HNF1A - Pro291fsinsC-HNF1A - causes HNF1A-MODY, which is a type of monogenic form of noninsulin-dependent diabetes mellitus. (Bell and Polonsky, 2001, Yamagata et al., 1996) Moreover, it has been reported that pancreatic cancer and hepatocellular carcinoma may be associated with HNF1A expression levels. (Pierce and Ahsan, 2011, Laumonier et al., 2007, Zucman-Rossi et al., 2006)

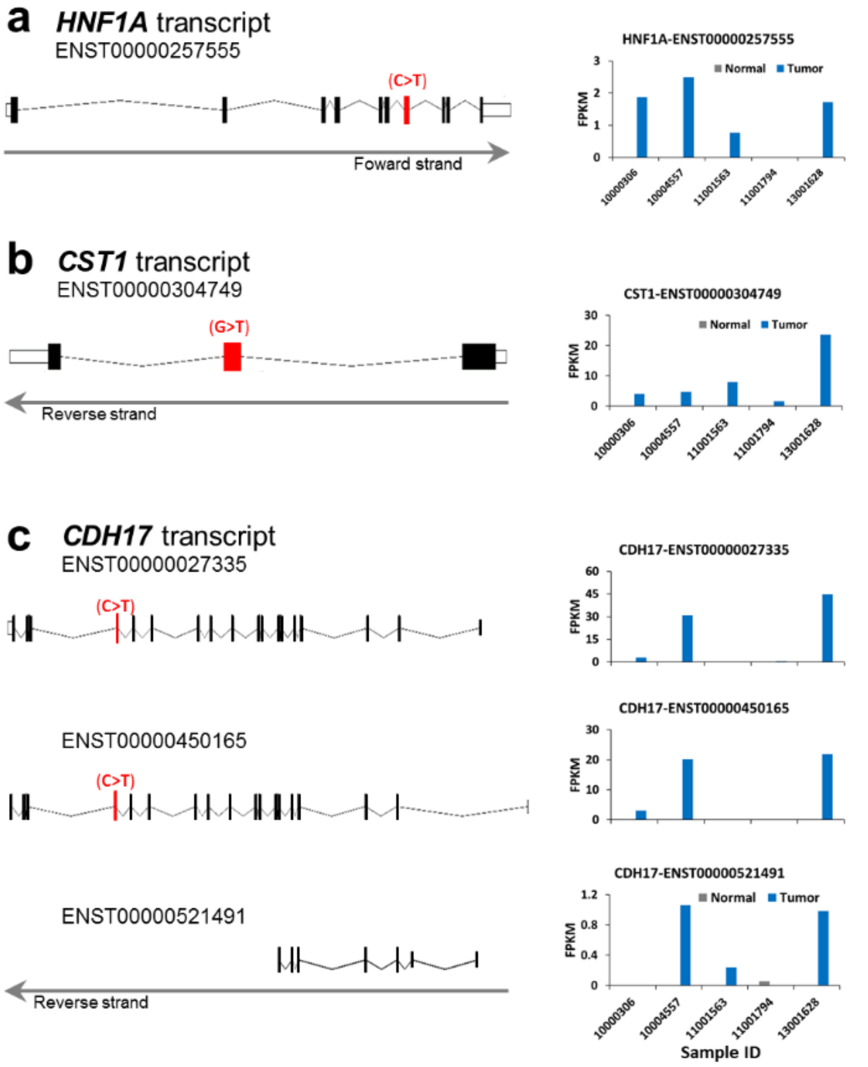


Figure 4. Lung tumor-specific synonymous variants

Tumor specific synonymous mutation exon map and transcript expression levels in lung tumor and adjacent normal tissues. **(a)** HNF1A **(b)** CST1 **(c)** CDH17.

Accordant nonsynonymous (missense) mutation identifications from proteogenomic analysis

Through the combined analysis of RNA-seq and proteomics in 5 pairs of lung ADC and adjacent normal tissues, this study first searched for any expressed SNP that caused amino acid changes in proteins. For this purpose, a sample-specific protein database was generated based on the RNA-seq sequences performed on the same samples. I used the web based 'QUILT' software (http://quilts.fenyolab.org/) that was developed by Dr. David Fenyo's lab at New York University. The RNA-seq in VCF Format was processed with QUILTS to generate each lung cancer tissue-specific reference protein database. Then, these proteome data were searched against this reference database to specifically select nonsynonymous mutants identified between ribonucleotide and amino acid sequences. The original search revealed a total of 16 nonsynonymous mutants. After manual validation for spectrum quality, 7 nonsynonymous mutants in 5 peptides of 5 different proteins remained (Table 1-3). Of these mutants, Lactotransferrin (LTF) was expressed in one tumor tissue sample and HDLBP was expressed in two tumor tissue samples. However, these mutations were not identified when the proteome data were searched using normal annotated protein databases. The two representative peptides from the LTF and HDLBP genes are presented on Figure 1-5. These mutations (LTF E535D and HDLBP S61A) were identified in lung cancer tissues but not in adjacent normal tissues.

Table 1-3. Identified nonsynonymous variant peptides in proteogenomic analysis

Chr	Gene name	Ref	Alt	Codon change	AA change	Exon ID	Identified sample	Identified peptide
2	HDLBP	A	C	Tct/Gct	S61A	NM_005336.ex.25	13001628 T	AACLESAQEPAGAWGNK
3	LTF	C	G	gaG/gaC	E535D	NM_001199149.ex.3	11001794 T	DVTVLQNTDGNNNDAWAK
3	TF	C	T	Cct/Tct	P589S	NM_001063	11001794 N/T	SVEEYANCHLAR
6	HLA-DRB5	A	C	Ttg/Gtg	L67V	NM_002125.ex.5	11001794 N	GIYNQEENVR

6	HLA-DRB5	C	T	Gac/Aac	D66N	NM_002125. ex.5	11001794 N	GIYNQEENVR
6	HLA-DRB5	T	C	gAc/gGc	D59G	NM_002125. ex.5	11001794 N	GIYNQEENVR
11	HBD	G	T	gCa/gAa	A23E	NM_000519. ex.3	11001563 N	VNVDEVGGEALGRL

Abbreviation used: Chr, chromosome; Ref, reference nucleotide; Alt, altered nucleotide

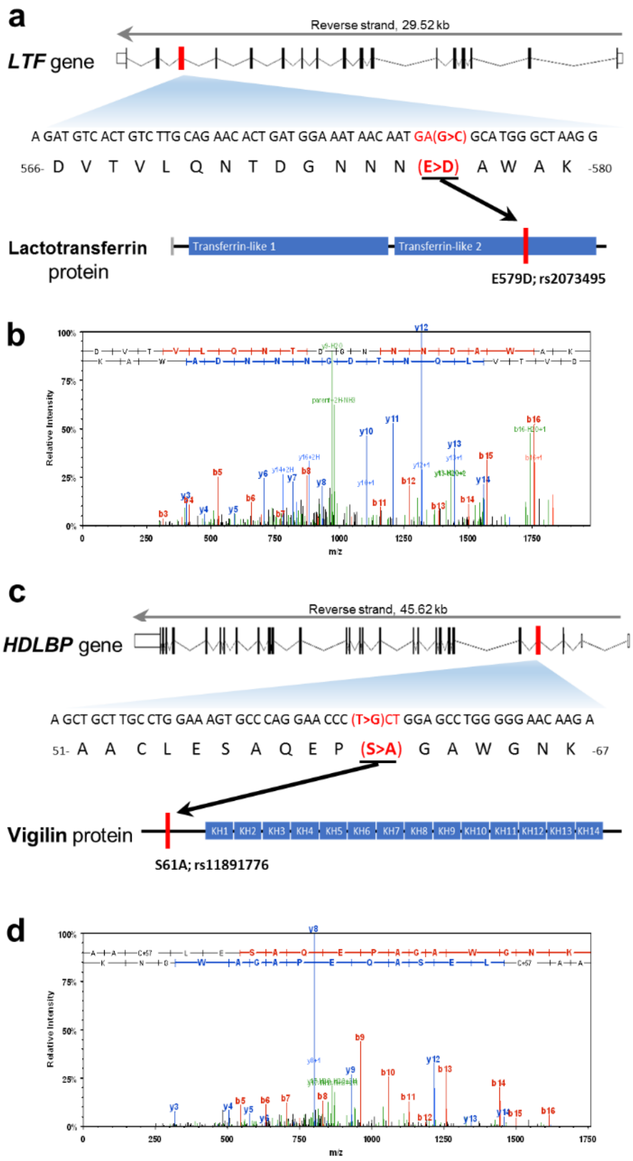


Figure 1-5. Identification of nonsynonymous variant peptides

(a) A representative example of the missense variant peptide from lactotransferrin (LTF). The red box indicates a region that contained a missense variant at the ribonucleotide and amino acid level. **(b)** Representative MS/MS spectrum of a missense variant peptide (DVTVLQNTDGNNN(E>D)AWAK) in lactotransferrin (LTF) **(c)** A representative example of the missense variant peptide HDLBP (Vigilin). The red box indicates a region that contained a missense variant at the ribonucleotide and amino acid level. **(d)** Representative MS/MS spectrum of a missense variant peptide (AACLESAQEP(S>A)GAWGNK) in Vigilin (HDLBP)

RNA-seq revealed a G→T mutation at position Chr3: 46480958 (release19, GRCh37.p13) in the LTF gene, and proteome data showed a matched glutamic acid to aspartic acid change at the 535 position (E535D) (Figure 1-5a). LTF is a strong iron binding member of the transferrin family that has antineoplastic, antibacterial, antimycotic, antiviral and anti-inflammatory activities. LTF is one of the genes most commonly inactivated in lung cancers by chromosomal elimination or epigenetic modulations. (Iijima et al., 2006) In this study, I found a novel NS mutation in LTF in lung cancers, displayed both as an expressed SNP and in peptide micro-sequencing by transcriptomics and proteomics approaches.

Another mutation was identified in High-density lipoprotein-binding protein (HDLBP), in which RNA-seq revealed a T→G mutation at position Chr2: 242203916 (release19, GRCh37.p13), and proteome data showed a matched serine to alanine change at peptide position 61 (S61A). HDLBP encodes the RNA-binding protein Vigilin, which contains 14 type I KH (hnRNP K homology) domains that participates in single strand nucleic acid binding and protein–protein interactions. (Grishin, 2001) HDLBP has been reported as a tumor suppressor that is frequently deleted or mutated in cancer cells. (van der Weyden et al., 2014, Molyneux et al., 2014) Again, my combined proteogenomic approach showed evidence of both an expressed SNP and an amino acid change.

Other nonsynonymous mutants identified are in Transferrin (TF P589S), hemoglobin delta (HBD A23E), and major histocompatibility complex, class II, DR beta 5 (HLA-DRB5 D59G, L67V, and D66N) (Table 1-3). I detected 3 nonsynonymous polymorphisms in an exon-5 region peptide of HLA-DRB5 protein, which belongs to the HLA class II beta chain paralogues that are expressed in antigen presenting cells such as B lymphocytes, dendritic cells, and macrophages. Since lung cancer recruits immune cells, this HLA-DRB5 might be presented by the immune cells in lung cancers. Among 6 exons of HLA-DRB5, exon 5 encodes the cytoplasmic tail. Within the DR molecule, the beta chain contains all the polymorphisms specifying the peptide binding specificities. Hundreds of DRB1 alleles have been described and typing for these polymorphisms is routinely done for tissue transplantation. (Reche and Reinherz, 2003) Thus, it should be noted that this HLA-DRB5 nonsynonymous variant is a form of naturally occurring polymorphisms and not a true cancer missense mutation.

I have not identified many nonsynonymous mutants that have been confirmed by proteomics. It is also notable that the four common nonsynonymous SNP variants in the genes CDH17, HIST1H1T, SAPCD2, and ZNF695, which were identified only in the 5 tumor samples and not in adjacent normal tissues, were not detected by my dual identification approach using both RNA-seq and proteomics. This is due in part to the low peptide coverage of protein identification in this discovery mode of proteomics analysis. To obtain wider coverage, and to therefore identify and confirm more missense mutants, it is required that I develop deep proteomics analysis that uses extensive fractionation, other digestion enzymes, and other ion dissociation methods, such as ETD.

So far, these results suggest that proteomics data analysis that uses a RNA-seq based, sample-specific protein reference database will help researchers to identify new variants and mutant proteins and their mutant sequences. Future studies of the missense mutations identified in LTF and HDLBP will reveal the function and role of these mutations in lung ADC.

Proteogenomic analysis identified a novel peptide derived from a pseudogene

I then performed proteogenomic data analysis to determine whether any pseudogene expression was detected. To select a confident list of expressed pseudogenes, a customized database was generated using measured pseudogene sequence at the transcript level that had an FPKM value above 2 (Figure 1-1). The selected transcript lists were converted to amino acid sequences by 3-frame translation and combined with the ENSEMBL_human_37.70.fasta database to determine an appropriate FDR cut-off within the score-spectra distribution by adding the effective number of the true positive population. However, by using this analysis with a customized database, I did not find any peptides from pseudogenes.

One pseudogene (EEF1A1P5) was identified with the UniProt human database (2015.02. rel.), which denotes the existence of this protein as 'uncertain'. EEF1A1P5 (Putative elongation factor 1-alpha-like 3) was detected all tissue samples. However, tumor tissues had 3.7-fold higher average quantitative values when normalized by total spectral counts (Figure 1-6). The sequence of EEF1A1P5 includes three GTP binding domains, which may result in the upregulation of protein synthesis by promoting translational elongation activities. Such a quantitative difference in pseudogene expression implies that EEF1A1P5 may be a potential lung cancer biomarker or causative gene. An increased rate of both global and/or specific protein synthesis is known to promote cell survival, angiogenesis, transformation, invasion and metastasis. Alterations to translation factors, including elongation factors of putative EEF1A1P5 and translational regulatory factors, have been reported in human cancers. (Silvera et al., 2010, Sonenberg, 1993, Anand et al., 2002, Ruggero and Pandolfi, 2003)

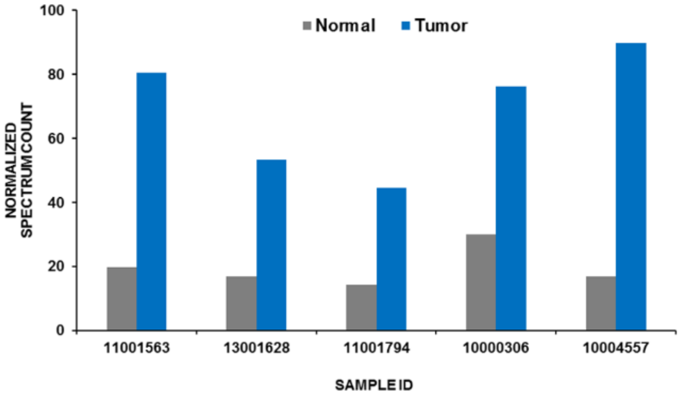


Figure 1-6. Detection of pseudogene expression at the protein level

The EEF1A1P5 pseudogene expression pattern was present at the most elevated levels in lung tissues. Quantitative values were used after normalizing by total spectra.

Similar proteogenomic analysis used for pseudogene identification has been performed to identify intron inclusion splicing variants, LNC RNA expression etc. However, in this analysis I did not identify any intron inclusion splicing variants and LNC RNA expressions by proteogenomic approaches. Alternatively, it is possible to find novel splice variants reads in reanalysis using other tools. If I use new tools such as MISO (Mixture of Isoforms) and Cufflinks as a de novo mode novel splicing events can be identified, because those tools have capacity to analyze all major types of alternative pre-mRNA processing at either the exon level or the isoform level. I am planning to analyze further for novel splicing variant aspect using additional tools such as MISO and Cufflinks.

CONCLUSION

Before C-HPP was launched in 2012, missing proteins on Chr 9 represented 20% of Chr 9-encoded proteins. At the time of this second publication, the remaining missing proteins represent 16% of Chr 9-encoded proteins (132 proteins). Proteome analysis using LTQ-Orbitrap in lung ADC tissues identified 19 missing proteins in whole human genome. RNA-seq analysis indicated that the remaining missing proteins are probably expressed at very low levels and the extreme difficulty encounters when attempting to detect the remaining missing proteins using current proteomics techniques in one sample type of tissue. Therefore, the use of various tissue types with more fractionation and high-speed mass spectrometry will be required to reveal the remaining missing proteins.

RNA-seq analysis detected RNA expression for 4 nonsynonymous and 4 synonymous mutations in all 5 tumor tissues but not in any of the adjacent normal tissues. I have also used a proteogenomic approach by combining proteomic data analysis with RNA-seq supported data from the same matched pairs of human lung tumor and adjacent normal tissues. The combined sample-specific proteogenomic analysis revealed 7 missense mutations from 5 peptides in lung tissues, demonstrating that variant gene and protein expression can be identified through this type of proteogenomic analysis. I also discovered peptides that were derived from the expression of a pseudogene. These results from the confirmation of both transcript and protein levels may present better disease specificity (Zhang et al., 2014, Alfaro et al., 2014) when analyzed in disease samples.

Analyzing mutations in genomic DNA level does not provide any information as to whether the mutated genes are expressed at the transcript level or, more importantly, whether they are expressed at the protein level and survive the protein degradation system and, thus, can cause cancers or be involved in tumorigenesis. However, my approach of using dual identifications of nonsynonymous variants expression by sample-specific RNA-seq based protein databases provides not only data confirmation but also direct information regarding diagnostic and/or drug target possibilities. Future studies will reveal the function of the missing proteins and confirm, quantitatively, the missense mutations between cancer tissues and normal tissues by MRM approaches, and thus diagnostic and therapeutic target possibilities in lung cancers.

This chapter was published as:

Kim YI, Lee J, Choi YJ, Seo J, Park J, Lee SY, Cho JY. (2015)

Proteogenomic Study beyond Chromosome 9: New Insight into Expressed Variant Proteome and Transcriptome in Human Lung Adenocarcinoma Tissues. *J Proteome Res.* **14**(12):5007-5016.

CHAPTER II

Multi-Panel Marker Development for the Differential Diagnosis of Lung Cancer and Lung Disease

INTRODUCTION

Lung cancer still presents high levels of mortality in cancer-related deaths (Siegel et al., 2014). Currently, lung cancer diagnosis largely depends on clinical imaging methods such as radiography, computed tomography (CT) and positron emission tomography (PET). However, these technologies are often incapable of distinguishing lung cancer from other lung abnormalities due to poor specificity (Henschke et al., 1999). Lesions from non-cancerous lung diseases can cause interference with solid tumor detection in imaging-based diagnosis, in which case only biopsy can definitively diagnose lung cancer (Ost et al., 2003). To overcome the current problems involved in diagnosing lung cancer apart from other respiratory diseases, the development of feasible, molecular marker-based differential diagnostic methods is highly desirable.

Several biochemical diagnostic molecules, called biomarkers, have been discovered from serum, allowing simple and non-invasive diagnosis. Several protein biomarkers are already used in the clinics for screening or monitoring therapy response, such as PSA for prostate cancer, CA125 for ovarian cancer, CA19-9 for pancreatic cancer, and CEA for colon cancer (Hanash et al., 2008).

Much research has been conducted in hopes of finding lung cancer diagnostic biomarkers in body fluids. At the present time, no biomarkers which discovered by proteomics are FDA-approved and used in clinical fields, (Kulasingam and Diamandis, 2008, Cho, 2007, Ludwig and Weinstein, 2005) and no biomarker has been developed for the differential diagnosis of lung cancers from other lung diseases.

Numerous biomarker candidates for lung cancer diagnosis have been discovered and reported. However, these biomarkers have never been tested as tools for the differential diagnosis of lung cancers and other lung diseases. For this purpose, discovered biomarker candidates should be validated on large scale using clinical samples. Conventional methods for validation, such as enzyme-linked immunosorbent assay (ELISA), immunohistochemistry (IHC), and western blotting, are based on immuno-affinity, which require costly antibodies for each biomarker. In contrast, nano-flow liquid chromatography triple quadrupole mass spectrometry (nLC-QqQ-MS) along with stable isotope dilution (SID) is the most widely used MS-based, antibody-free technology for quantifying multiple proteins with high-sensitivity, high-specificity, and high-reproducibility (Addona et al., 2009, Keshishian et al., 2009, Kuzyk et al., 2009, Gillette and Carr, 2013a). In this large-scale validation process, it is particularly useful to develop a method with improved sensitivity but without additional enrichment or pre-fractionation of samples. In addition, because no single marker is significantly differentially expressed in sera from lung cancer patients, focusing on the combination of two or more variables under consideration of statistical interaction would be particularly valuable for improving the differential diagnosis of lung cancer and other lung diseases (Lombardi et al., 1990).

In this study, 198 serum samples from patients with non-cancerous lung disease or lung cancer were subjected to nano-flow MRM to analyze the levels of multiple lung cancer biomarker candidates. Comparison of a panel of marker combinations using logistic regression generated meta-markers with improved capabilities for differential diagnosis of lung cancer from other lung diseases.

MATERIAL AND METHODS

Sample Collection

Serum samples from lung disease patients and lung cancer patients were obtained at Asan Medical Center (Table 2-1). Ninety nine serum samples of non-cancerous lung disease patients, who visited Asan Medical Center, department of pulmonology and critical care medicine, were collected from May, 2011 to April, 2013, and similarly ninety nine serum samples from lung cancer patients, who visited Asan Medical Center, department of pulmonology and critical care medicine, were collected from March, 2012 to February, 2013. All serum samples were collected using a serum-separating tube (SST). The SSTs were centrifuged to separate the serum from whole blood within 24 hours after venipuncture. Samples were stored at -70 °C until analysis. Informed consent was obtained from all donors (IRB 2011-0076).

Table 2-1. Clinical information of serum samples

	Training Set		**Validation Set**	
	Lung Disease	**Lung Cancer**	**Lung Disease**	**Lung Cancer**
Population	30	30	69	69
Age	57 (33-82)	61.8 (25-80)	56.8 (27-79)	60.4 (30-75)
Sex (Male/Female)	17/13	21/9	40/29	45/24
Disease diagnosis (TB/PN/ND/etc.)	3/12/9/6		14/20/17/18	
Cancer types				
Carcinoid tumor		0		1
NSCLC (1/2/3/4)		11/2/5/4		29/10/7/15
SCLC (LD, ED)		8/0		7/0
Smoking history (Curr, Ex, Non)	6/10/14	14/6/10	18/18/33	21/21/27

Abbreviations used: TB, tuberculosis; PN, pneumonia; ND, nodule; NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer; LD, limited disease; ED, extensive disease; Curr, current-smoker; Ex, ex-smoker; Non, non-smoker

Sample Preparation; In-solution Tryptic Digestion

Serum samples were diluted 20-fold with HPLC-grade water (Honeywell Burdick & Jackson, Muskegon, MI). The protein concentration of diluted serum was measured using Quick Start™ Bradford 1x Dye Reagent (Bio-Rad Laboratories, Hercules, CA). Then, 30 μg of protein was prepared in 50 mM ammonium bicarbonate and denatured by boiling at 100 °C for 20 min. Dithiothreitol (Bio-Rad Laboratories, Hercules, CA) and iodoacetamide (Sigma-Aldrich, St Louis, MO) were added for reduction and alkylation, both at a concentration of 10 mM. Trypsin (Roche, Mannheim, Germany) was added to samples at a protein-to-enzyme ratio of 50:1 (w/w) and incubated at 37 °C for 16 hours. The digested peptide mixtures were cleaned using Pierce C18 spin column (Thermo Scientific, Rockford, IL, USA) according to the manufacturer's instructions, dried, and stored at -20 °C until analysis. Dried tryptic digests were resuspended in 30 μL of 0.1% formic acid for mass spectrometric analysis.

Target Peptide Selection

Tryptic target peptides were selected using the Skyline program (64-bit, Version 1.4.0.4421) (MacCoss Laboratory, University of Washington, Seattle, WA). Peptides containing NXT/NXS and RP/KP motifs were excluded, and peptides with lengths of 7 to 24 amino acids were selected. Carbamidomethyl cysteine structural modifications were included. To select proteotypic peptides, unique peptides were sorted using Uniprot human database (2012.11 released version). To select easily detectable peptides by nLC-QqQ-MS, the NIST Q-Tof database was also used as a reference. Stable isotope synthetic peptides used in this study are as follows: β–Galactosidase; LNVENPK, AHGS; EHAVEGDCDFQLLK, ITIH1; LDAQASFLPK, CLUS; ASSIIDELFQDR (JPT Peptide Technologies GmbH, Berlin, Germany), SERPINA4; GDATVFFILPNQGK, PON1; YVYIAELLAHK (21st Century Biochemicals, MA, USA).

nLC-MRM-MS Analysis

Liquid chromatography was conducted using a 1260 Infinity LC system with Chip Cube (Agilent technologies, Santa Clara, CA). The tryptic digest was separated in HPLC-Chip, which consisted of a 40 nL enrichment column and a 75 μm x 150 mm analytical column packed with 300 Å C18. Separation was performed using binary gradients with buffers A (HPLC-grade water in 0.1% formic acid solution) and B (acetonitrile in 0.1% formic acid solution). The column was initially equilibrated and eluted at a flow rate of 0.4 μL/min for the nano-flow pump and 4 μL/min for the capillary-flow pump. The 30-min LC schedule with 14-min gradient was programmed as follows: 1–15 min, 3–40% B for reverse phase separation; 15–18 min, 40–80% B to wash the chip column; finally, 12 min in 3% B to equilibrate the chip column prior to the injection of the next sample.

A triple quadruple mass spectrometer (6490 Agilent technologies, Santa Clara, CA) was used with the following parameters: positive ion mode, a drying gas flow rate of 11 L/min at 150 °C, and MS1 and MS2 set to unit resolution. For relative quantification of 60 sera, dynamic MRM was conducted with a cycle time set to 500 ms, and the minimum and maximum dwell times were 19.36 and 198.55 ms, respectively. For SID-MRM validation of sera, dynamic MRM was conducted with cycle time of 500 ms, and the minimum and maximum dwell times were 18.16 and 123.12 ms, respectively. Delta retention time was 4 min. The data were visualized by Skyline (64-bit, Version 1.4.0.4421) (MacCoss Laboratory, University of Washington, Seattle, WA).

Statistical Analysis

Normalized MRM data were statistically analyzed using the T-test, and $P < 0.05$ was considered statistically significant. Combinations of markers were analyzed using the binary logistic regression. All of the analyses were performed using SPSS Statistics 23 (IBM Corp., NY, USA).

RESULTS

Target Screening for Differential Diagnosis of Lung Cancer and Other Lung Diseases

To study biomarker combinations using targeted proteomics, the targets were selected from previous studies, and the analytic conditions were optimized (Figure 2-1). Fifteen potential single biomarkers were selected from Ahn's previous lung cancer biomarker study (Ahn et al., 2014a) and a publication from another research group, which analyzed secretome of 23 cancer cell lines (Table 2-1) (Wu et al., 2010).

Relative quantification using MRM was conducted, and the data were normalized to the LNVENPK SIS-peptide derived from β–galactosidase of *Escherichia coli*, which had been used as an internal standard (Figure 2-2). A training set of human serum samples consisting of 30 cases of non-cancerous lung diseases (LD) and 30 cases of lung cancer (LC), which are subsets of total 198 samples, were subjected to MRM analysis (Figure 2-3). Eight target proteins showed significant differences between serum samples of LD and LC in the MRM analysis ($P < 0.05$). All but one candidate marker (CLUS) were observed to be down-regulated in LC compared to LD.

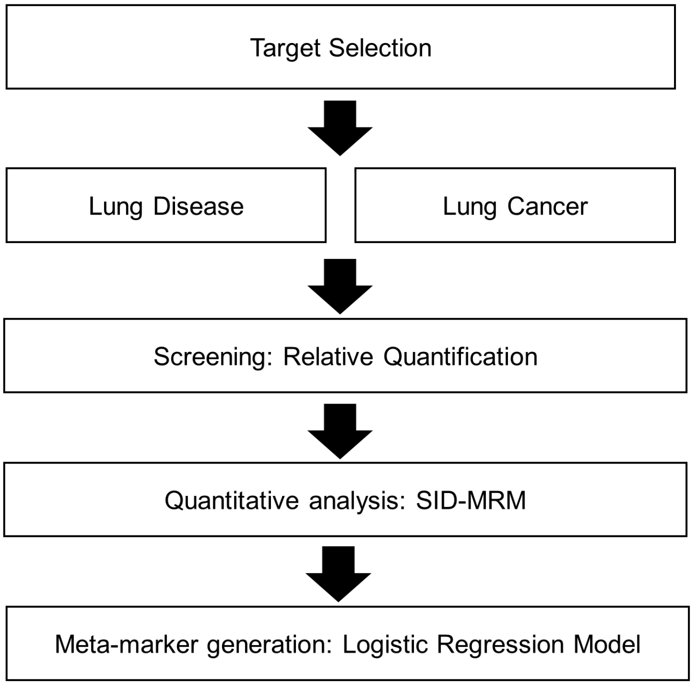

Target Selection
Lung Disease
Lung Cancer
Screening: Relative Quantification
Quantitative analysis: SID-MRM
Meta-marker generation: Logistic Regression Model

Figure 2-1. Schematic diagram of overall workflow

Table 2-2. Specifications of total target biomarker candidates

UniProtKB	Name	Peptide sequence	Description
P04003	C4BPA	QSSSYSFFK	C4b-binding protein alpha chain
P02647	APOA1	THLAPYSDELR	Apolipoprotein A-I
P00736	C1R	MDVFSQNMFCAGHPSLK	Complement C1r subcomponent
P02765	AHSG	EHAVEGDCDFQLLK	Alpha-2-HS-glycoprotein
P51884	LUM	NNQIDHIDEK	Lumican
P33151	CADH5	ELDSTGTPTGK	Cadherin-5
P19827	ITIH1	LDAQASFLPK	Inter-alpha-trypsin inhibitor heavy chain H1
P10909	CLUS	ASSIIDELFQDR	Clusterin
P05543	THBG	FSISATYDLGATLLK	Thyroxine-binding globulin

Q86UE4	LYRIC	TLPPATSTEPSVILSK	Protein LYRIC
O00501	CLD5	EFYDPSVPVSQK	Claudin-5
P02748	C9	RPWNVASLIYETK,	Complement component 9
P80108	GPLD1	TLLLVGSPTWK,	Phosphatidylinositol-glycan-specific phospholipase D
P27169	PON1	YVVYIAELLAHK	Serum paraoxonase 1
P29622	SERPINA4	GDATVFFILPNQGK	Kallistatin

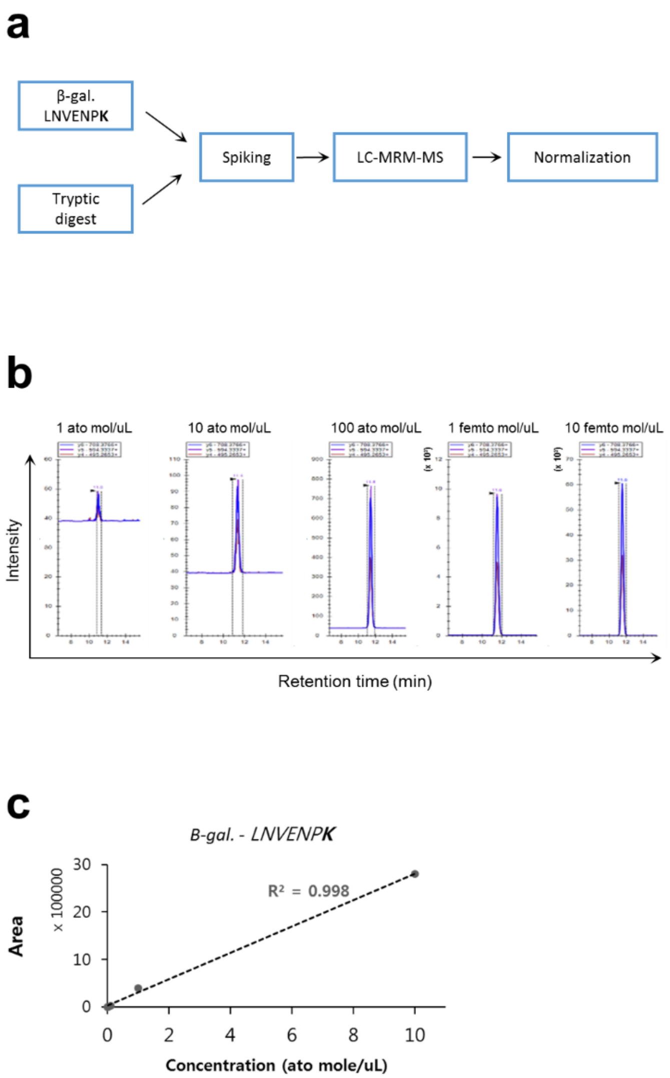


Figure 2-2. Specifications of LNVENPK derived from β–galactosidase as internal standard

(a) Schematic flow of relative quantification. **(b)** Peaks obtained by spiking increased amount of LNVENPK. **(c)** Standard curve analysis to demonstrate linearity of LNVENPK.

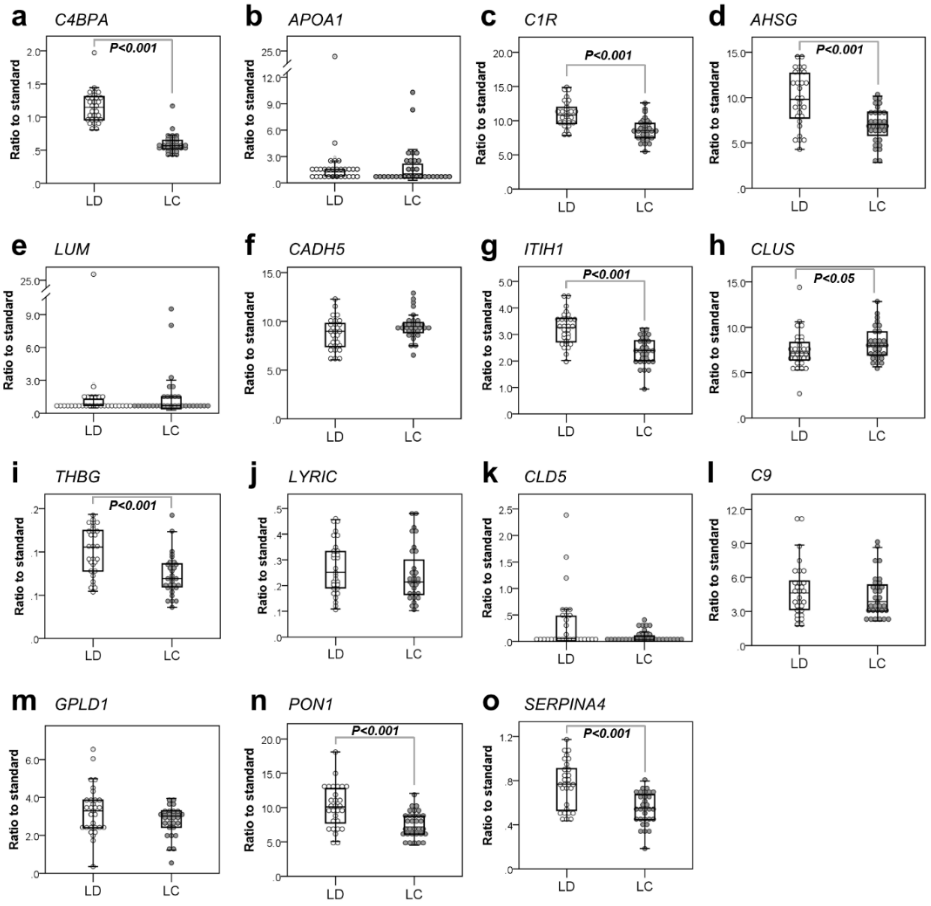


Figure 2-3. Relative quantitation on crude sera of a training set

(a) - (o) Quantification results of each biomarker candidates in sera of non-cancer lung disease patients (N=30) and lung cancer patients (N=30). All of relative quantitation value was normalized by peak area of spiked LNVENPK peptide which derived from β-galactosidase.

SID-MRM development for selected biomarker candidate

To quantify biomarker candidate proteins using MRM methods in clinical samples, the accuracy of the peptide quantification must be measured first. To confirm the precise quantitative accuracy of the MRM performance, the stable isotope dilution (SID) technique was used. Stable isotope-labeled standard (SIS)-peptides were tested at several peptide concentrations to determine the optimal endogenous detection ranges for each target (Gillette and Carr, 2013a). SIS peptides were synthesized with the stable isotope-labeled ^{13}C and ^{15}N incorporated at the C-terminal lysine or arginine residue for the selected 8 protein candidates. Then, the SIS-peptides were measured by MRM at a 50 femtomole per microliter concentration to determine whether the peptides are detectable in nLC-QqQ-MS. The peptides corresponding to the selected 8 proteins C4BPA, C1R, AHGS, ITIH1, CLUS, SERPINA4, THBG and PON1 are QSSSYSFFK, MDVFSQNMFCAGHPSLK, EHAVEGDCDFQLLK, LDAQASFLPK, ASSIIDELFQDR, GDATVFFILPNQGK, FSISATYDLGATLLK and YVYIAELLAHK, respectively. To establish quantification methods for these peptides, I performed collision energy optimization for selection of well-detected charge states and transitions to be measured (Figure 2-4 and Table 2-3). QSSSYSFFK (C4BPA), MDVFSQNMFCAGHPSLK (C1R), and FSISATYDLGATLLK (THBG) showed poor reproducibility in liquid chromatography, and thus, these proteins were excluded from further analysis.

Standard curves for the trypsin-digested crude sera were constructed for the five SIS-peptides (Figure 2-5). The sum of the total ion area corresponded to the known quantities for each peptide. All analyses were performed in triplicate. All standard curves showed good linearity ($0.93 < R^2 < 0.99$) when the SIS-peptides were spiked in crude sera.

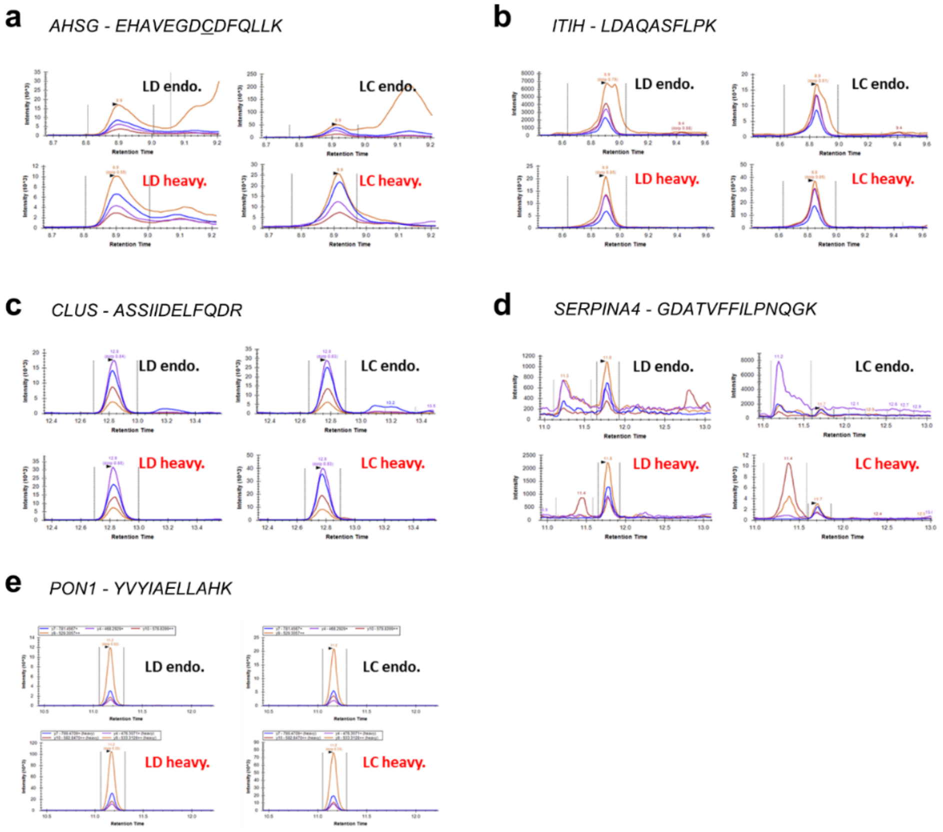


Figure 2-4. Representative extracted ion chromatogram (EIC) in the tryptic digested crude sera were constructed for five SIS-peptides

(a)-(e) EIC show sets of transitions and retention time of each target peptides in LD and LC crude sera.

Table 2-3. Selected transitions of five biomarker candidates for SID-MRM

Protein	Type	Peptide sequence		Product Ion (m/z)	Ion Name	CE (eV)
	Retention Time (min)	Peptide molecular mass (Da)	Precursor Ion (m/z)			
AHSG	Light	EHAVEGDCDFQLLK		147.1	y1+	15.2
	8.8	1659.8	554.3	373.3	y3+	15.2
				648.4	y5+	15.2
				763.4	y6+	15.2
	Heavy	EHAVEGDCDFQLLK *		155.1	y1+	15.2
	8.8	1667.8	556.9	381.3	y3+	15.2
				656.4	y5+	15.2
				771.4	y6+	15.2
ITIH1	Light	LDAQASFLPK		244.2	y2+	17.9
	8.8	1088.6	545.3	662.4	y6+	17.9
				861.5	y8+	17.9
				976.5	y9+	17.9

	Heavy	LDAQASFLPK *		252.2	y2+	17.9
	8.8	1096.6	549.3	670.4	y6+	17.9
				869.5	y8+	17.9
				984.5	y9+	17.9
CLUS	Light	ASSIIDELFQDR		418.2	y3+	22.6
	12.8	1392.7	697.4	565.3	y4+	22.6
				922.4	y7+	22.6
				1035.5	y8+	22.6
	Heavy	ASSIIDELFQDR*		428.2	y3+	22.6
	12.8	1402.7	702.4	575.3	y4+	22.6
				932.4	y7+	22.6
				1045.5	y8+	22.6
SERPINA4	Light	GDATVFFILPNQGK		543.3	y5+	24.4
	11.7	1505.8	753.9	769.5	y7+	24.4
				916.5	y8+	24.4
				1063.6	y9+	24.4
	Heavy	GDATVFFILPNQGK*		551.3	y5+	24.4
	11.7	1513.8	757.9	777.5	y7+	24.4

				924.5	y8+	24.4
				1071.6	y9+	24.4
PON1	Light	YVYIAELLAHK		468.3	y4+	11.1
	11.1	1319.7	440.6	781.5	y7+	11.1
				529.3	y9++	11.1
				578.8	y10++	11.1
	Heavy	YVYIAELLAHK*		476.3	y4+	11.1
	11.1	1327.7	443.3	789.5	y7+	11.1
				533.3	y9++	11.1
				582.8	y10++	11.1

R*; 13C(6)15N(2) labeled argine, K*; 13C(6) 15N(4) labeled lysin

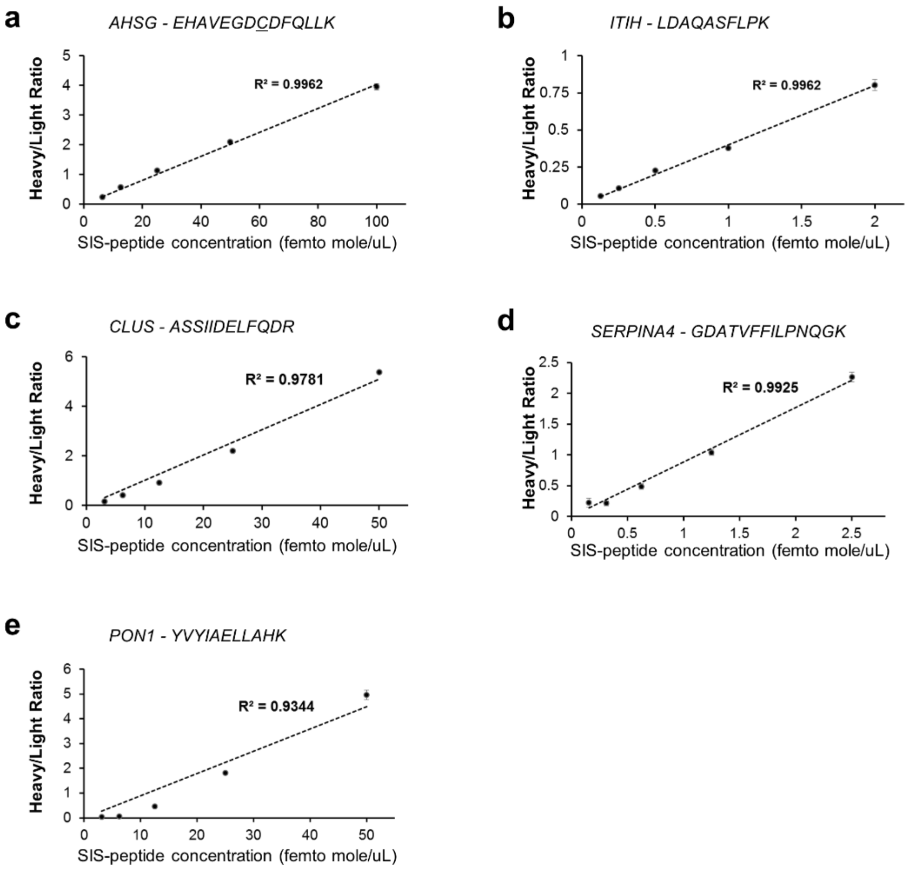


Figure 2-5. Standard curves in the tryptic digested crude sera were constructed for five SIS-peptides

(a)-(e) Standard curve generated by spiking serially increasing amount of stable isotope-labeled synthetic (SIS)-peptide. Each analysis conducted in triplicate.

SID-MRM Analysis of Selected Targets in Crude Serum

Using developed SID-MRM methods, seven targets were validated in sera from 99 non-cancerous lung disease patients (LD) and 99 lung cancer patients (LC). The results showed that only one target protein (SERPINA4) showed statistically significant changes between LD and LC (Figure 2-6). SERPINA4 was significantly lower ($P < 0.001$) in the sera of lung cancer patients than in lung disease patients. However, other target proteins did not show any significant differences between the two groups. To calculate the differential diagnostic power, ROC curves were generated for each protein . SERPINA4 had the highest area-under-curve (AUC) values of 0.836, respectively (Figure 2-6 d).

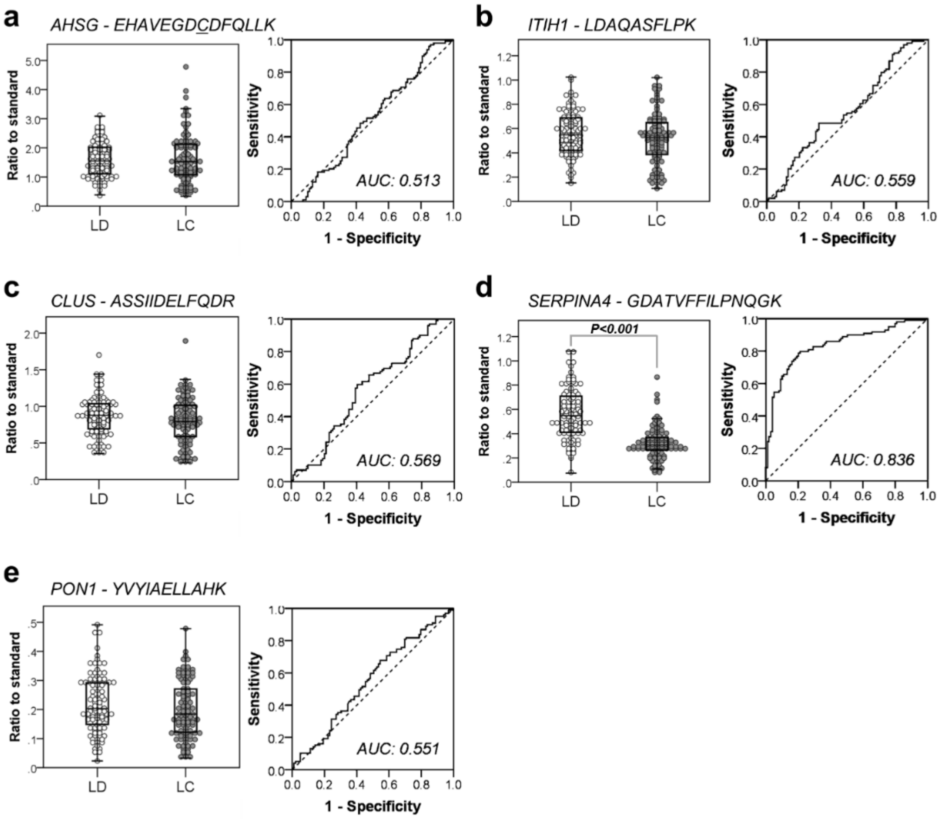


Figure 2-6. SID-MRM validation of selected targets in crude sera

SID-MRM analysis showed that (d) SERPINA4 was significantly lower in lung cancer patients (LC; n=99) compared to lung disease patients group (LD; n=99) (P<0.05). The other proteins, **(a) – (c)** and **(e)** had no statistical difference between two groups.

Modeling for Meta-marker Generation

To improve the diagnostic power, I made combinations of the candidate results using a logistic regression model. SID-MRM results from 30 samples from each group were used as a training set, and the other 69 data samples for each group were used as the test set. The quantitative data of individual biomarkers had AUC-value differences between the training set and the test set below 0.1 for all targets (Figure 2-7).

The SERPINA4 result was significant and had high AUC values (Figure 3). Therefore, I combined the quantitative protein data and patient clinical information with the SERPINA4-results to assess any potential increase in diagnostic ability. I included statistical interactions for variables with no significance as single markers to generate the best fitting model. The estimated logistic regression statistic values for each variable revealed age, PON1, and ITIH1 have significant interactions (Table 2-4).

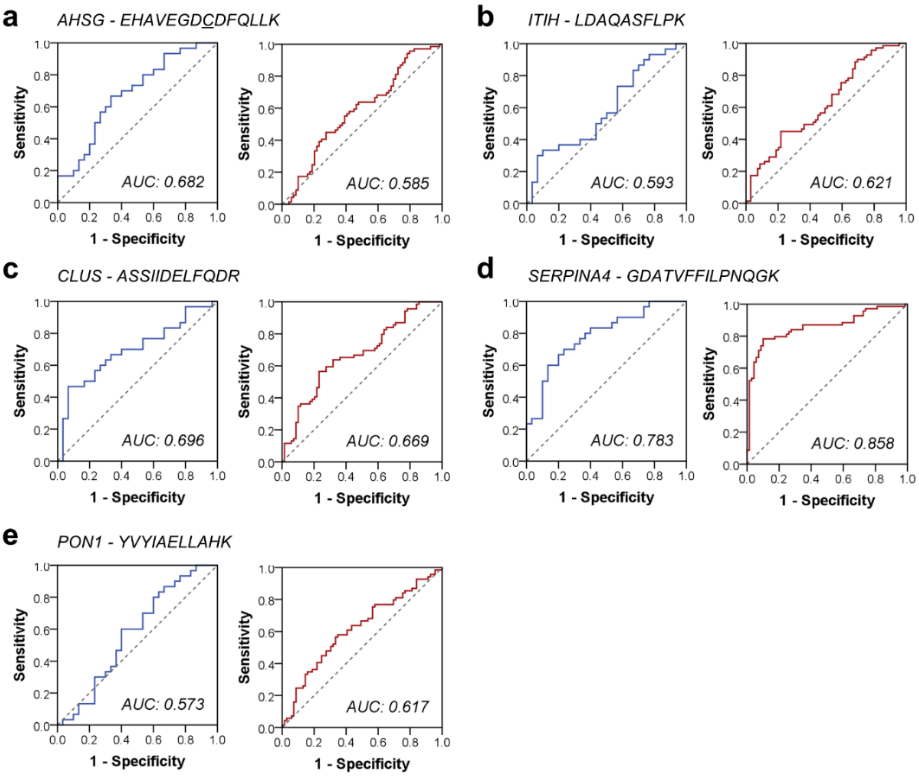


Figure 2-7. ROC curve analysis of selected targets in crude sera of training and validation set

(a)-(e) ROC curve is made of using SID-MRM analysis data. Training set (left panel; blue lined graph) is consisted of lung cancer patients' sera (LC; n=30) compared to lung disease patients' (LD; n=30). In validation set (right panel; red lined graph) represent sera of lung cancer patients (LC; n=69) compared to lung disease patients (LD; n=69).

Table 2-4. Estimated Logistic Regression Statistic Values in Training set

	Score	Sig.
SERPINA4	12.597	0.0004
Age	2.468	0.116
PON1	0.567	0.452
CLUS	6.048	0.014
AHSG	6.971	0.008
ITIH1	1.736	0.188
C1R	0.179	0.672
Smoking	3.293	0.070
Sex	1.148	0.284
SERPINA4 by Age	5.786	0.016
SERPINA4 by PON1	3.988	0.046
SERPINA4 by CLUS	2.723	0.099
SERPINA4 by AHSG	0.460	0.498
SERPINA4 by ITIH1	3.858	0.050
SERPINA4 by C1R	2.836	0.092
SERPINA4 by Smoking	0.349	0.555
SERPINA4 by SEX	0.107	0.744

Abbreviations used: Sig, significance

This strategy produced two types of models. The first model included PON1 (PON1, SERPINA4, age, SERPINA4*PON1, and SERPINA4*age) (Table 2-5). The second model included ITIH1 (ITIH1, SERPINA4, age, SERPINA4*ITIH1, and SERPINA4*ITIH1) (Table 2-6). However, the ITIH1 model showed a poor fit in the logistic regression (Table 2-7, 8, 9). Thus, I subsequently excluded the ITIH1 model.

Then, the equation was tested by using a validation data set to determine if the PON1 model is reliable. The results showed the model correctly fit the data for both the training set and the validation set (Figure 2-8). In addition, the resulting logistic regression model distinguish tuberculosis and stage1 NSCLC patients with AUC value of 0.881 (Figure 2-10).

Table 2-5. Logistic Regression Analysis for PON1, SERPINA4, and age included Modeling in Training set

	Beta	Std. error	Wald	Sig.	Odd ratio	95% CI for odd ratio	
						Lower	Upper
SERPINA4	-4.760	19.275	0.061	0.805	8.569.E-03	3.359.E-19	2.186.E+14
Age	0.155	0.143	1.178	0.278	1.168.E+00	8.826.E-01	1.544.E+00
PON1	47.400	18.004	6.932	0.008	3.849.E+20	1.823.E+05	8.129.E+35
SERPINA4 by Age	-0.220	0.270	0.659	0.417	8.029.E-01	4.726.E-01	1.364.E+00
SERPINA4 by PON1	-19.302	28.175	0.469	0.493	4.141.E-09	4.307.E-33	3.981.E+15
Constant	-11.202	10.235	1.198	0.274	1.365.E-05		

Abbreviations used: SE, standard error; Sig., significance; CI, confidence interval

Table 2-6. Logistic Regression Analysis for ITIH1 included Modeling in Training set

	Beta	Std. error	Wald	Sig.	Odd ratio	95% CI for odd ratio	
						Lower	Upper
SERPINA4	10.889	16.071	0.459	0.498	1.866.E-05	3.901.E-19	8.926.E+08
AGE	0.045	0.109	0.170	0.680	1.046.E+00	8.450.E-01	1.294.E+00
ITIH1	10.858	7.098	2.340	0.126	5.195.E+04	4.717.E-02	5.721.E+10
AGE by SERPINA4	0.037	0.222	0.028	0.868	1.038.E+00	6.713.E-01	1.604.E+00
ITIH1 by SERPINA4	-5.063	13.063	0.150	0.698	6.326.E-03	4.811.E-14	8.318.E+08
Constant	-3.807	7.938	0.230	0.632	2.222.E-02		

Abbreviations used: SE, standard error; Sig., significance; CI, confidence interval

Table 2-7. Model Summary

Model	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
PON1_model	40.145	0.512	0.683
ITIH1_model	54.302a	0.382	0.509

Table 2-8. Omnibus Tests of Model Coefficients

Model	Chi-square	df	Sig.
PON1_model	43.033	5	0.000000036
ITIH1_model	28.875	5	0.000024531

Table 2-9. Hosmer and Lemeshow Test

Model	Chi-square	df	Sig.
PON1_model	2.234	8	0.972984584
ITIH1_model	11.230	8	0.188990469

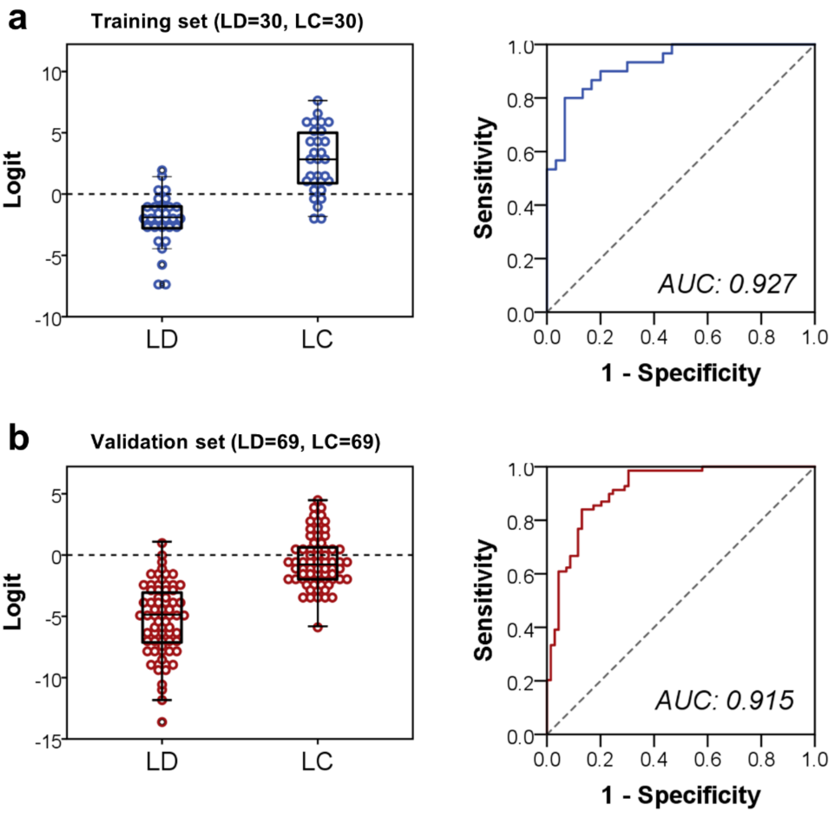


Figure 2-8. Differential diagnostic capability of biomarker combinations in both the training set and validation set

The model built based on SERPINA4, PON1 and Age in the training set **(a)** and the combination was applied to validation set **(b)**.

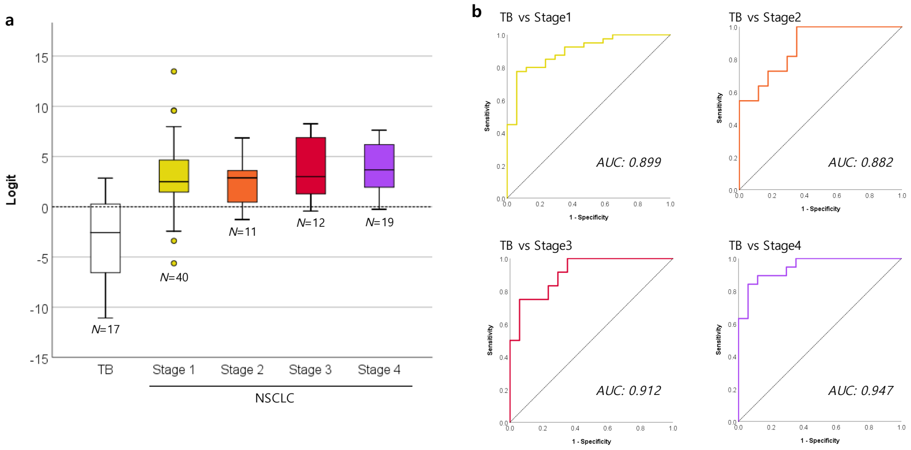


Figure 2-9. Differential diagnostic capability of biomarker combination between tuberculosis and NSCLC by stage

(a) The box plot indicates logit of TB and all stages of NSCLC. **(b)** The marker combination that generated from logistic regression distinguished TB and stage 1-4 of NSCLC with AUC value of 0.899, 0.882, 0.912, and 0.947, respectively.

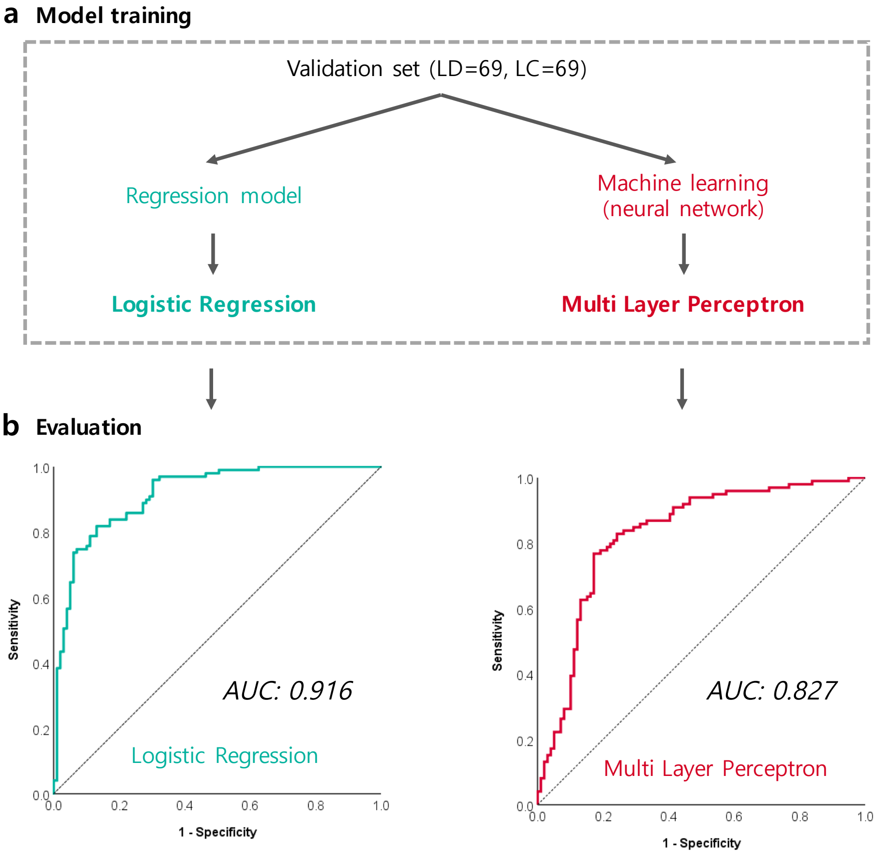


Figure 2-10. The ROC curves and AUCs for logistic regression and multi layer perceptron

(a) To maximize robustness of models, validation set, which has larger size of samples, was used to train both of models. **(b)** Comparison between logistic regression and multi layer perceptron showed different AUCs that of 0.916 and 0.827, respectively. Total 198 samples was used for ROC analysis.

DISCUSSION

Diagnosis of lung cancer mainly relies on imaging technologies such as radiography, CT, and PET scans. Radiography is convenient but has low sensitivity. CT is commonly used as a lung cancer diagnostic method; however, it is not an ideal method due to radiation exposure and cost. PET is the most sensitive of the three methods, but it is also the most expensive. Biopsies, such as fine needle aspiration (FNA), are the most traditional and reliable method; however, biopsies are a highly invasive procedure. With the primary use of imaging diagnostics, many lung cancer patients are conflated with patients with lung-associated symptoms. Many of the lung-associated symptoms, especially tuberculosis, pneumonia and lung nodules which are selected for this study, interfere with the diagnosis of lung cancer in imaging-based diagnosis. In this study, seven biomarker candidates were selected and validated as differential diagnostic biomarkers by MRM analysis, and the biomarker combination of SERPINA4, PON1 and age was shown to form the optimal differential diagnostic meta-marker.

This study was designed for the validation of each protein marker candidates and development of meta-marker model for the differential diagnosis of lung cancer (LC) from other lung diseases (LD). Therefore, sample set was made up non-cancerous lung disease and lung cancer patients without healthy population. To verify the meta-marker generated is statistically reliable, LD group and LC group are designed to be composed of equal numbers.

Along with the development of high-throughput technologies, many studies have reported biomarker candidates. However, validating the candidates in large scale samples is still considered to be a hurdle for biomarker development. Large-scale clinical validation is highly difficult due to the number of targets and samples.

Traditional immunoassay-based quantitative methods are suitable for single biomarker validation, making it inappropriate to multiplex marker validation. MS-based assay is sufficiently cost-effective than immunoassay when the analysis is performed to multiplex biomarkers. Even with current price of stable isotope labeled internal standard and MS machine, it makes MRM assay costs highly competitive in comparison to current cost in the clinical immunoassays operated by clinical hospital laboratory (Diamandis, 2009, Percy et al., 2014, Kato et al., 2011, Carr and Anderson, 2008).

Nano-flow MRM, an MS-based quantitative proteomics technology, provides simultaneous validation with high sensitivity with a little amount of sample, however, compared to standard-flow MRM unstable reproducibility should be optimized (Addona et al., 2009, Keshishian et al., 2009, Kuzyk et al., 2009, Gillette and Carr, 2013a). Therefore, SIS-peptide as an internal standard should be spiked for accurate quantitative analysis. Considering cost effectiveness, relative quantification using global standards (in this study, the LNVENPK synthetic peptide derived from 🚲-galactosidase was used) is possible, which provides high productivity in multi-target screens.

When protein quantitative analysis using mass spectrometer were performed, selection of unique peptide derived from a protein is important. Especially, hydrophobic and/or high reactivity amino acid composition of peptide is influential factor to liquid chromatography. For example, three peptides derived from C4BPA, C1R and THBP are excluded because they are composed at least 30% hydrophobic amino acid and/or methionine, a highly oxidative residue.

To closely reflect clinical environment, minimal sample preparation, without abundant serum protein depletion or pre-fractionation, is a valid approach, if detectable range of nano-flow MRM is considered. The normal concentrations of AHSG (72 μg/mL), ITIH1 (10.42 μg/mL), CLUS (152.36 μg/mL), SERPINA4 (22.1 μg/mL), and PON1 (59.3 μg/mL) are reported in human blood (Chambers et al., 2013, Chao et al., 1996, Kujiraoka et al., 2000). Also, when Agilent 6490 mass spectrometer coupled with standard-flow ESI was used to quantify the proteins in human crude plasma samples, lower limit of quantitation (LLOQ) is reported to be 751 ng/mL. Considering Agilent 6490 coupled with nano-flow ESI was used in this study, direct detection of target peptide/protein by nLC-QqQ-MS is appropriate (Wilm and Mann, 1996).

Despite vast numbers of biomarker studies in past two decades, there remains no approved valid differential diagnostic biomarker for lung cancers (Diamandis, 2010, Konforte and Diamandis, 2013). Currently, there is no reliable blood biomarker for differential diagnosis of lung cancer from other lung diseases. Previous studies have focused on single biomarkers. However, there are known technological limitations associated with identifying biomarkers within a complex biological system such as cancer because there are multiple differentially expressed proteins. The limitation of a single protein biomarker could be overcome by combining multiple marker panels to form a meta-marker with improved diagnostic value via weighting on significant marker, considering interaction, and compensate quantitation errors. Statistical models including logistic regression and machine learning enable the selection of meta-markers that show improved diagnostic power (Leichtle et al., 2013). In this study, logistic regression show better distinguish ability than multi layer perceptron that is representative neural network algorithm (Figure 2-10).

Logistic regression has been used in clinical statistics to estimate causes of disease and in analyses of combinations of multi-marker panels (Xiao et al., 2013, Zhang et al., 2013). Logit, which derives from logistic regression as a representative value, is defined as the log-value of (probability/1-probability). In this study, positive logit values imply a diagnosis of cancer and negative logit values imply non-cancerous lung disease. The meta-marker presented in this study is a combination result based on this logistic model, which functions well in distinguishing lung cancer patients from non-cancerous lung disease patients. Currently, low-dose CT is used and reported to be effective for lung cancer screening. Considering high-false positive rates of that method (23.3%) , however, meta-marker developed in this study show better false positive rate (2/69) (Team, 2011).

The results demonstrated that only the significant variables are needed for the marker panel. However, a superficial relationship combination could miss complex biological interactions. These interactions may explain biological phenomena more accurately. In this study, the quantitative data from five proteins and the patient clinical information were used to produce a high quality model. Through these considerations, the biomarker candidates lacking significant results (PON1, age) as single markers were found to have significant positive effects on the meta-marker function. The logistic model maximizes the cooperative effect using optimum weighting to elucidate the value of each variable. In summary a meta-marker by the PON1, SERPINA4, and age combination is the most potent differential meta-marker with a suitable number of proteins (Aviram and Rosenblat, 2005).

Serpin peptidase inhibitor, clade A, member 4 (SERPINA4) is known as a serine protease inhibitor and heparin-binding protein. SERPINA4 regulates angiogenesis, inflammatory reactions, and blood pressure (Zhu et al., 2007, Wang et al., 2005, Chen et al., 1997, Chao et al., 1997). SERPINA4 inhibits vascular endothelial growth factor (VEGF) or basic fibroblast growth factor (bFGF)-induced angiogenesis and tumor growth (Miao et al., 2002, Miao et al., 2003). Because of its anti-tumor effect via anti-inflammatory and anti-angiogenic activity, SERPINA4 has been studied as a potential therapeutic in laboratory trials (Shiau et al., 2010, Zhu et al., 2007, Wang et al., 2005, Chen et al., 1997). At this point, the results indicating a down-regulation of SERPINA4 in lung cancer patient serum compared to other lung diseases are in accordance with previously reported studies.

Serum paraoxonase 1 (PON1) is mainly expressed in the liver and is secreted into the blood. PON1 is hydrolytic enzyme that processes organophosphate substrates and is associated with high density lipid (HDL (Aviram et al., 1998). It has been suggested that PON1 protects cells against lipid oxidation, but the antioxidant mechanism remains unknown (Aviram and Rosenblat, 2005). In the Ahn's previous study, PON1 was found in decreased levels in the sera of small cell lung cancer (SCLC); reversely, the degree of PON1 fucosylation was increased (Ahn et al., 2014b). The down-regulation of PON1 in cancer is also reported with endometrial, ovarian, pancreatic, and other lung cancers (Arioz et al., 2009, Camuzcuoglu et al., 2009, Elkiran et al., 2007, Akcay et al., 2003). This study showed that the non-cancerous lung disease patients had further decreased levels of PON1 compared to lung cancer patients. It is also reported that PON1 activity was significantly lower in pulmonary tuberculosis than normal individuals (Naderi et al., 2011). Larger-scale validation with a normal group included or antioxidant-related target validation may help explain why PON1 appears at even lower levels in the lung disease group in comparison with the lung cancer group.

In conclusion, this study presents meta-markers for the differential diagnosis of lung cancer and non-cancerous lung diseases. These meta-markers were determined by calculated values obtained from logistic regression models under consideration of statistical interaction. The meta-markers are combinations of data about not only protein levels as measured by MRM but also clinical information, and have the potential to enhance the differential diagnostic power between lung cancers and other lung diseases.

CONCLUSION

In this study, I attempted to identify the best meta-marker by the combination of the MRM quantitative values from a panel of proteins. Theese results showed that a meta-marker combination of SERPINA4, PON1 and age improved sensitivity and specificity when used together as a biomarker for the differential diagnosis between lung cancers and non-cancerous lung diseases, even PON1 did not show significance as a single bio signature. The results thus indicate that the combination of several potential biomarkers, determined via modeling under consideration of statistical interaction, would likely provide better diagnostic specificity and sensitivity than a single biomarker for the differential diagnosis between lung cancer and lung disease patients.

This chapter was published as:

Kim YI, Ahn JM, Sung HJ, Na SS, Hwang J, Kim Y, Cho JY. (2016)

Meta-markers for the differential diagnosis of lung cancer and lung disease.

J Proteomics. **148**:36-43.

GENERAL CONCLUSION

Personalized medicine in oncology has taken great strides, with predictive biomarkers guiding both therapy and monitoring of disease progression or remission (Ong et al., 2012). Recent advances in -omics technologies have led to the emergence of personalized medicine for complex diseases. Taken together, MS-based proteomics has been evolved to more sensitive, accurate, efficient, and diverse applications. Furthermore, the field of proteomics continues to develop currently. In this context, to break through the complete human proteome, the C-HPP consortium has been encouraged to apply various new technologies of each participant, and this approach has resulted in 40% of missing proteins (3,325 proteins) being claimed from 2012 to 2018.

In these studies, I also tried to apply various state-of-the-art technologies to tissues or blood from lung cancer patients. In the proteome discovery phase including known and novel proteins, total RNA-seq was allowed to build a personal transcriptome database, followed by proteogenomic analysis. This advanced approach suggests the example of personalized medicine based on cross-omics profiles.

In addition, using MRM technology, which is a suitable platform to verify protein marker candidates, I developed a lung cancer differential diagnostic marker panel. The combined marker shows better diagnostic specificity and sensitivity than a single biomarker. Statistical modeling needs as many reliable data sets as possible; therefore, MRM, which easily enables the multiplexing assay, is well suited for combination rather than immunoassay. I anticipate that this combined blood protein signature might provide a complementary strategy to the established image-based lung cancer diagnosis to differentiate other lung diseases.

Considering that this series of proteome/proteogenome studies is utilizes state-of-the-art omics technology that spans the discovery phase to verification phase, this dissertation might provide insight into establishing an effective pipeline for protein-based lung cancer biomarker development.

REFERENCES

ADDONA, T. A., ABBATIELLO, S. E., SCHILLING, B., SKATES, S. J., MANI, D. R., BUNK, D. M., SPIEGELMAN, C. H., ZIMMERMAN, L. J., HAM, A. J., KESHISHIAN, H., HALL, S. C., ALLEN, S., BLACKMAN, R. K., BORCHERS, C. H., BUCK, C., CARDASIS, H. L., CUSACK, M. P., DODDER, N. G., GIBSON, B. W., HELD, J. M., HILTKE, T., JACKSON, A., JOHANSEN, E. B., KINSINGER, C. R., LI, J., MESRI, M., NEUBERT, T. A., NILES, R. K., PULSIPHER, T. C., RANSOHOFF, D., RODRIGUEZ, H., RUDNICK, P. A., SMITH, D., TABB, D. L., TEGELER, T. J., VARIYATH, A. M., VEGA-MONTOTO, L. J., WAHLANDER, A., WALDEMARSON, S., WANG, M., WHITEAKER, J. R., ZHAO, L., ANDERSON, N. L., FISHER, S. J., LIEBLER, D. C., PAULOVICH, A. G., REGNIER, F. E., TEMPST, P. & CARR, S. A. 2009. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol,* 27**,** 633-41.

AEBERSOLD, R. & MANN, M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature,* 537**,** 347.

AHN, J.-M., KIM, M.-S., KIM, Y.-I., JEONG, S.-K., LEE, H.-J., LEE, S. H., PAIK, Y.-K., PANDEY, A. & CHO, J.-Y. 2013. Proteogenomic analysis of human chromosome 9-encoded genes from human samples and lung cancer tissues. *Journal of proteome research,* 13**,** 137-146.

AHN, J. M., SUNG, H. J., YOON, Y. H., KIM, B. G., YANG, W. S., LEE, C., PARK, H. M., KIM, B. J., KIM, B. G., LEE, S. Y., AN, H. J. & CHO, J. Y. 2014a. Integrated glycoproteomics demonstrates fucosylated serum paraoxonase 1 alterations in small cell lung cancer. *Mol Cell Proteomics,* 13**,** 30-48.

AHN, J. M., SUNG, H. J., YOON, Y. H., KIM, B. G., YANG, W. S., LEE, C., PARK, H. M., KIM, B. J., KIM, B. G., LEE, S. Y., AN, H. J. & CHO, J. Y. 2014b. Integrated Glycoproteomics Demonstrates Fucosylated Serum Paraoxonase 1 Alterations in Small Cell Lung Cancer. *Molecular & Cellular Proteomics,* 13**,** 30-48.

AKCAY, M. N., YILMAZ, I., POLAT, M. F. & AKCAY, G. 2003. Serum paraoxonase levels in gastric cancer. *Hepatogastroenterology,* 50 Suppl 2**,** cclxxiii-cclxxv.

ALFARO, J. A., SINHA, A., KISLINGER, T. & BOUTROS, P. C. 2014. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature methods,* 11**,** 1107.

ALTELAAR, A. M. & HECK, A. J. 2012. Trends in ultrasensitive proteomics. *Current opinion in chemical biology,* 16**,** 206-213.

ANAND, N., MURTHY, S., AMANN, G., WERNICK, M., PORTER, L. A., CUKIER, I. H., COLLINS, C., GRAY, J. W., DIEBOLD, J. & DEMETRICK, D. J. 2002. Protein elongation factor EEF1A2 is a putative oncogene in ovarian cancer. *Nature genetics,* 31**,** 301.

APPELLA, E., PADLAN, E. & HUNT, D. 1995. Analysis of the structure of naturally processed peptides bound by class I and class II major histocompatibility complex molecules. *Exs,* 73**,** 105-119.

ARAVIDIS, C., PANANI, A. D., KOSMAIDOU, Z., THOMAKOS, N., RODOLAKIS, A. & ANTSAKLIS, A. 2012. Detection of numerical abnormalities of chromosome 9 and p16/CDKN2A gene alterations in ovarian cancer with fish analysis. *Anticancer research,* 32**,** 5309-5313.

ARIOZ, D. T., CAMUZCUOGLU, H., TOY, H., KURT, S., CELIK, H. & EREL, O. 2009. Assessment of serum paraoxonase and arylesterase activity in patients with endometrial cancer. *Eur J Gynaecol Oncol,* 30**,** 679-82.

AVIRAM, M. & ROSENBLAT, M. 2005. Paraoxonases and cardiovascular diseases: pharmacological and nutritional influences. *Curr Opin Lipidol,* 16**,** 393-9.

AVIRAM, M., ROSENBLAT, M., BISGAIER, C. L., NEWTON, R. S., PRIMO-PARMO, S. L. & LA DU, B. N. 1998. Paraoxonase inhibits high-density lipoprotein oxidation and preserves its functions. A possible peroxidative role for paraoxonase. *J Clin Invest,* 101**,** 1581-90.

BELL, G. I. & POLONSKY, K. S. 2001. Diabetes mellitus and genetically programmed defects in β-cell function. *Nature,* 414**,** 788.

BRANCA, R. M., ORRE, L. M., JOHANSSON, H. J., GRANHOLM, V., HUSS, M., PÉREZ-BERCOFF, Å., FORSHED, J., KÄLL, L. & LEHTIÖ, J. 2014. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods,* 11**,** 59.

BURGESS, M. W., KESHISHIAN, H., MANI, D., GILLETTE, M. A. & CARR, S. A. 2014. Simplified and efficient quantification of low abundance proteins at very high multiplex by targeted mass spectrometry. *Molecular & Cellular Proteomics***,** mcp. M113. 034660.

CAMUZCUOGLU, H., ARIOZ, D. T., TOY, H., KURT, S., CELIK, H. & EREL, O. 2009. Serum paraoxonase and arylesterase activities in patients with epithelial ovarian cancer. *Gynecol Oncol,* 112**,** 481-5.

CARR, S. A. & ANDERSON, L. 2008. Protein quantitation through targeted mass spectrometry: the way out of biomarker purgatory? *Clinical chemistry,* 54**,** 1749-1752.

CHAMBERS, A. G., PERCY, A. J., YANG, J., CAMENZIND, A. G. & BORCHERS, C. H. 2013. Multiplexed quantitation of endogenous proteins in dried blood spots by multiple reaction monitoring-mass spectrometry. *Molecular & Cellular Proteomics,* 12**,** 781-791.

CHAO, J., SCHMAIER, A., CHEN, L.-M., YANG, Z. & CHAO, L. 1996. Kallistatin, a novel human tissue kallikrein inhibitor: levels in body fluids, blood cells, and tissues in health and disease. *Journal of Laboratory and Clinical Medicine,* 127**,** 612-620.

CHAO, J. L., STALLONE, J. N., LIANG, Y. M., CHEN, L. M., WANG, D. Z. & CHAO, L. 1997. Kallistatin is a potent new vasodilator. *Journal of Clinical Investigation,* 100**,** 11-17.

CHEN, L. M., CHAO, L. & CHAO, J. 1997. Adenovirus-mediated delivery of human kallistatin gene reduces blood pressure of spontaneously hypertensive rats. *Hum Gene Ther,* 8**,** 341-7.

CHEVALIER, F. 2010. Highlights on the capacities of "Gel-based" proteomics. *Proteome Sci,* 8**,** 23.

CHO, W. C. 2007. Potentially useful biomarkers for the diagnosis, treatment and prognosis of lung cancer. *Biomed Pharmacother,* 61**,** 515-9.

CHOI, E. H., KIM, J.-T., KIM, J. H., KIM, S.-Y., SONG, E. Y., KIM, J. W., KIM, S.-Y., YEOM, Y. I., KIM, I.-H. & LEE, H. G. 2009.

Upregulation of the cysteine protease inhibitor, cystatin SN, contributes to cell proliferation and cathepsin inhibition in gastric cancer. *Clinica Chimica Acta,* 406**,** 45-51.

CONSORTIUM, I. C. G. 2010. International network of cancer genome projects. *Nature,* 464**,** 993.

COX, H. D., LOPES, F., WOLDEMARIAM, G. A., BECKER, J. O., PARKIN, M. C., THOMAS, A., BUTCH, A. W., COWAN, D. A., THEVIS, M. & BOWERS, L. D. 2014. Interlaboratory agreement of insulin-like growth factor 1 concentrations measured by mass spectrometry. *Clinical chemistry,* 60**,** 541-548.

DAGHER, J., DUGAY, F., VERHOEST, G., CABILLIC, F., JAILLARD, S., HENRY, C., ARLOT-BONNEMAINS, Y., BENSALAH, K., OGER, E. & VIGNEAU, C. 2013. Histologic prognostic factors associated with chromosomal imbalances in a contemporary series of 89 clear cell renal cell carcinomas. *Human pathology,* 44**,** 2106-2115.

DALTON, W. S. & FRIEND, S. H. 2006. Cancer biomarkers—an invitation to the table. *Science,* 312**,** 1165-1168.

DIAMANDIS, E. P. 2009. Protein quantification by mass spectrometry: Is it ready for prime time? *Clinical chemistry,* 55**,** 1427-1430.

DIAMANDIS, E. P. 2010. Cancer biomarkers: can we turn recent failures into success? *J Natl Cancer Inst,* 102**,** 1462-7.

DICKINSON, D., THIESSE, M. & HICKS, M. 2002. Expression of type 2 cystatin genes CST1-CST5 in adult human tissues and the developing submandibular gland. *DNA and cell biology,* 21**,** 47-65.

DIETERICH, D. C., LINK, A. J., GRAUMANN, J., TIRRELL, D. A. & SCHUMAN, E. M. 2006. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). *Proceedings of the National Academy of Sciences,* 103, 9482-9487.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29, 15-21.

DRABOVICH, A. P., MARTINEZ-MORILLO, E. & DIAMANDIS, E. P. 2015. Toward an integrated pipeline for protein biomarker development. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics,* 1854, 677-686.

EBHARDT, H. A., ROOT, A., SANDER, C. & AEBERSOLD, R. 2015. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics,* 15, 3193-3208.

ELIAS, J. E. & GYGI, S. P. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods,* 4, 207.

ELKIRAN, E. T., MAR, N., AYGEN, B., GURSU, F., KARAOGLU, A. & KOCA, S. 2007. Serum paraoxonase and arylesterase activities in patients with lung cancer in a Turkish population. *BMC Cancer,* 7, 48.

ELLIS, M. J., GILLETTE, M., CARR, S. A., PAULOVICH, A. G., SMITH, R. D., RODLAND, K. K., TOWNSEND, R. R., KINSINGER, C., MESRI, M. & RODRIGUEZ, H. 2013. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery,* 3, 1108-1112.

EZKURDIA, I., VÁZQUEZ, J. S., VALENCIA, A. & TRESS, M. 2014. Analyzing the first drafts of the human proteome. *Journal of proteome research,* 13, 3854-3855.

FLOWER, D. R. 1996. The lipocalin protein family: structure and function. *Biochemical Journal,* 318**,** 1-14.

GARTNER, J. J., PARKER, S. C., PRICKETT, T. D., DUTTON-REGESTER, K., STITZEL, M. L., LIN, J. C., DAVIS, S., SIMHADRI, V. L., JHA, S. & KATAGIRI, N. 2013. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proceedings of the National Academy of Sciences,* 110**,** 13481-13486.

GEYER, P. E., HOLDT, L. M., TEUPSER, D. & MANN, M. 2017. Revisiting biomarker discovery by plasma proteomics. *Molecular systems biology,* 13**,** 942.

GILLETTE, M. A. & CARR, S. A. 2013a. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nat Methods,* 10**,** 28-34.

GILLETTE, M. A. & CARR, S. A. 2013b. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nature methods,* 10**,** 28.

GLISH, G. L. & VACHET, R. W. 2003. The basics of mass spectrometry in the twenty-first century. *Nature reviews Drug discovery,* 2**,** 140.

GREER, S. M., PARKER, W. R. & BRODBELT, J. S. 2015. Impact of protease on ultraviolet photodissociation mass spectrometry for bottom-up proteomics. *Journal of proteome research,* 14**,** 2626-2632.

GRISHIN, N. V. 2001. KH domain: one motif, two folds. *Nucleic acids research,* 29**,** 638-643.

GYGI, S. P. & AEBERSOLD, R. 2000a. Mass spectrometry and proteomics. *Curr Opin Chem Biol,* 4**,** 489-94.

GYGI, S. P. & AEBERSOLD, R. 2000b. Using mass spectrometry for quantitative proteomics. *Trends in Biotechnology,* 18, 31-36.

GYGI, S. P., CORTHALS, G. L., ZHANG, Y., ROCHON, Y. & AEBERSOLD, R. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A,* 97, 9390-5.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *cell,* 144, 646-674.

HANASH, S. M., PITTERI, S. J. & FACA, V. M. 2008. Mining the plasma proteome for cancer biomarkers. *Nature,* 452, 571-9.

HENDRICKS, N. G., LAREAU, N. M., STOW, S. M., MCLEAN, J. A. & JULIAN, R. R. 2014. Bond-specific dissociation following excitation energy transfer for distance constraint determination in the gas phase. *Journal of the American Chemical Society,* 136, 13363-13370.

HENSCHKE, C. I., MCCAULEY, D. I., YANKELEVITZ, D. F., NAIDICH, D. P., MCGUINNESS, G., MIETTINEN, O. S., LIBBY, D. M., PASMANTIER, M. W., KOIZUMI, J., ALTORKI, N. K. & SMITH, J. P. 1999. Early Lung Cancer Action Project: overall design and findings from baseline screening. *Lancet,* 354, 99-105.

HSU, J.-L., HUANG, S.-Y., CHOW, N.-H. & CHEN, S.-H. 2003. Stable-isotope dimethyl labeling for quantitative proteomics. *Analytical chemistry,* 75, 6843-6852.

HUMPHRAY, S., OLIVER, K., HUNT, A., PLUMB, R., LOVELAND, J., HOWE, K., ANDREWS, T., SEARLE, S., HUNT, S. & SCOTT, C. 2004. DNA sequence and analysis of human chromosome 9. *Nature,* 429, 369.

HUSTOFT, H. K., MALEROD, H., WILSON, S. R., REUBSAET, L., LUNDANES, E. & GREIBROKK, T. 2012. A critical review of trypsin digestion for LC-MS based proteomics. *Integrative Proteomics*. InTech.

IIJIMA, H., TOMIZAWA, Y., IWASAKI, Y., SATO, K., SUNAGA, N., DOBASHI, K., SAITO, R., NAKAJIMA, T., MINNA, J. D. & MORI, M. 2006. Genetic and epigenetic inactivation of LTF gene at 3p21. 3 in lung cancers. *International journal of cancer,* 118**,** 797-801.

JEMAL, A., BRAY, F., CENTER, M. M., FERLAY, J., WARD, E. & FORMAN, D. 2011. Global cancer statistics. *CA: a cancer journal for clinicians,* 61**,** 69-90.

JENSEN, O. N. 2004. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current opinion in chemical biology,* 8**,** 33-41.

JUNG, K.-W., WON, Y.-J., KONG, H.-J. & LEE, E. S. 2018. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2015. *Cancer Research and Treatment: Official Journal of Korean Cancer Association,* 50**,** 303.

JUNG, K.-W., WON, Y.-J., OH, C.-M., KONG, H.-J., CHO, H., LEE, D. H. & LEE, K. H. 2015. Prediction of cancer incidence and mortality in Korea, 2015. *Cancer research and treatment: official journal of Korean Cancer Association,* 47**,** 142.

KATO, H., NISHIMURA, T., IKEDA, N., YAMADA, T., KONDO, T., SAIJO, N., NISHIO, K., FUJIMOTO, J., NOMURA, M. & ODA, Y. 2011. Developments for a growing Japanese patient population: facilitating new technologies for future health care. *Journal of proteomics,* 74**,** 759-764.

KESHISHIAN, H., ADDONA, T., BURGESS, M., MANI, D. R., SHI, X., KUHN, E., SABATINE, M. S., GERSZTEN, R. E. & CARR, S. A. 2009. Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Mol Cell Proteomics,* 8**,** 2339-49.

KIM, M.-S., PINTO, S. M., GETNET, D., NIRUJOGI, R. S., MANDA, S. S., CHAERKADY, R., MADUGUNDU, A. K., KELKAR, D. S., ISSERLIN, R. & JAIN, S. 2014. A draft map of the human proteome. *Nature,* 509**,** 575.

KIM, Y.-I., AHN, J.-M., SUNG, H.-J., NA, S.-S., HWANG, J., KIM, Y. & CHO, J.-Y. 2016. Meta-markers for the differential diagnosis of lung cancer and lung disease. *Journal of proteomics*, 148, 36-43.

KIM, Y.-I. & CHO, J.-Y. 2018. Gel-based proteomics in disease research: Is it still valuable? *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1867, 9-16.

KIM, Y.-I., LEE, J., CHOI, Y.-J., SEO, J., PARK, J., LEE, S.-Y. & CHO, J.-Y. 2015. Proteogenomic study beyond chromosome 9: new insight into expressed variant proteome and transcriptome in human lung adenocarcinoma tissues. *Journal of proteome research*, 14, 5007-5016.

KIM, Y. J., SERTAMO, K., PIERRARD, M.-A., MESMIN, C. D., KIM, S. Y., SCHLESSER, M., BERCHEM, G. & DOMON, B. 2015. Verification of the biomarker candidates for non-small-cell lung cancer using a targeted proteomics approach. *Journal of proteome research,* 14**,** 1412-1419.

KONDRAT, R. W., MCCLUSKY, G. A. & COOKS, R. G. 1978. Multiple reaction monitoring in mass spectrometry/mass spectrometry for direct analysis of complex mixtures. *Analytical chemistry,* 50**,** 2017-2021.

KONFORTE, D. & DIAMANDIS, E. P. 2013. Is early detection of cancer with circulating biomarkers feasible? *Clin Chem,* 59**,** 35-7.

KUHN, E., WU, J., KARL, J., LIAO, H., ZOLG, W. & GUILD, B. 2004. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. *Proteomics,* 4**,** 1175-1186.

KUJIRAOKA, T., OKA, T., ISHIHARA, M., EGASHIRA, T., FUJIOKA, T., SAITO, E., SAITO, S., MILLER, N. E. & HATTORI, H. 2000. A sandwich enzyme-linked immunosorbent assay for human serum paraoxonase concentration. *Journal of lipid research,* 41**,** 1358-1363.

KULASINGAM, V. & DIAMANDIS, E. P. 2008. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat Clin Pract Oncol,* 5**,** 588-99.

KUZYK, M. A., SMITH, D., YANG, J., CROSS, T. J., JACKSON, A. M., HARDIE, D. B., ANDERSON, N. L. & BORCHERS, C. H. 2009. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol Cell Proteomics,* 8**,** 1860-77.

LANGE, V., PICOTTI, P., DOMON, B. & AEBERSOLD, R. 2008. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology,* 4**,** 222.

LAUMONIER, H., BLANC, J. F., LAURENT, C., POUSSIN, K., RULLIER, A., LE BAIL, B., CUNHA, A. S., SARIC, J., TRILLAUD, H. & BALABOUD, C. P. High HFN1 alpha mutation frequency in liver adenomatosis. HEPATOLOGY, 2007. JOHN WILEY & SONS INC 111 RIVER ST, HOBOKEN, NJ 07030 USA, 430A-430A.

LEE, H. J., NAM, K. T., PARK, H. S., KIM, M. A., LAFLEUR, B. J., ABURATANI, H., YANG, H. K., KIM, W. H. & GOLDENRING, J. R. 2010a. Gene expression profiling of metaplastic lineages identifies CDH17 as a prognostic marker in early stage gastric cancer. *Gastroenterology,* 139**,** 213-225. e3.

LEE, K.-S., CHANG, H.-S., LEE, S.-M. & PARK, E.-C. 2015. Economic burden of cancer in Korea during 2000-2010. *Cancer research and treatment: official journal of Korean Cancer Association,* 47**,** 387.

LEE, N. P., POON, R. T., SHEK, F. H., NG, I. O. & LUK, J. M. 2010b. Role of cadherin-17 in oncogenesis and potential therapeutic implications in hepatocellular carcinoma. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer,* 1806**,** 138-145.

LEGRAIN, P., AEBERSOLD, R., ARCHAKOV, A., BAIROCH, A., BALA, K., BERETTA, L., BERGERON, J., BORCHERS, C. H., CORTHALS, G. L. & COSTELLO, C. E. 2011. The human proteome project: current state and future direction. *Molecular & cellular proteomics,* 10**,** M111. 009993.

LEICHTLE, A. B., DUFOUR, J. F. & FIEDLER, G. M. 2013. Potentials and pitfalls of clinical peptidomics and metabolomics. *Swiss Med Wkly,* 143**,** w13801.

LENG, S. X., MCELHANEY, J. E., WALSTON, J. D., XIE, D., FEDARKO, N. S. & KUCHEL, G. A. 2008. ELISA and multiplex technologies for cytokine measurement in inflammation and aging research. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences,* 63**,** 879-884.

LIEBLER, D. C. & ZIMMERMAN, L. J. 2013. Targeted quantitation of proteins by mass spectrometry. *Biochemistry,* 52**,** 3797-3806.

LOMBARDI, C., TASSI, G. F., PIZZOCOLO, G. & DONATO, F. 1990. Clinical significance of a multiple biomarker assay in patients with lung cancer. A study with logistic regression analysis. *Chest,* 97, 639-44.

LUDWIG, J. A. & WEINSTEIN, J. N. 2005. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer,* 5, 845-56.

MANSOUR, V. J. & COORSSEN, J. R. 2018. Quantitative Gel Electrophoresis. *Proteomics in Domestic Animals: from Farm to Systems Biology*. Springer.

MIAO, R. Q., AGATA, J., CHAO, L. & CHAO, J. 2002. Kallistatin is a new inhibitor of angiogenesis and tumor growth. *Blood,* 100, 3245-3252.

MIAO, R. Q., CHEN, V., CHAO, L. & CHAO, J. 2003. Structural elements of kallistatin required for inhibition of angiogenesis. *American Journal of Physiology-Cell Physiology,* 284, C1604-C1613.

MIDTHUN, D. E. & JETT, J. R. Update on screening for lung cancer. Seminars in respiratory and critical care medicine, 2008. © Thieme Medical Publishers, 233-240.

MOLYNEUX, S. D., WATERHOUSE, P. D., SHELTON, D., SHAO, Y. W., WATLING, C. M., TANG, Q.-L., HARRIS, I. S., DICKSON, B. C., THARMAPALAN, P. & SANDVE, G. K. 2014. Human somatic cell mutagenesis creates genetically tractable sarcomas. *Nature genetics,* 46, 964.

NADERI, M., HASHEMI, M., KOMIJANI-BOZCHALOEI, F., MOAZENI-ROODI, A. & MOMENIMOGHADDAM, M. 2011. Serum paraoxonase and arylesterase activities in patients with pulmonary tuberculosis. *Pathophysiology,* 18, 117-20.

NARAYANAN, V., GUTMAN, J., POLLYEA, D. & JIMENO, A. 2013. Omacetaxine mepesuccinate for the treatment of chronic myeloid leukemia. *Drugs Today,* 49**,** 447.

NESVIZHSKII, A. I. 2014. Proteogenomics: concepts, applications and computational strategies. *Nature methods,* 11**,** 1114.

NIE, S., SHI, T., FILLMORE, T. L., SCHEPMOES, A. A., BREWER, H., GAO, Y., SONG, E., WANG, H., RODLAND, K. D. & QIAN, W.-J. 2017. Deep-dive targeted quantification for ultrasensitive analysis of proteins in nondepleted human blood plasma/serum and tissues. *Analytical chemistry,* 89**,** 9139-9146.

NING, K. & NESVIZHSKII, A. I. 2010. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC bioinformatics,* 11**,** S14.

NOAMAN, N., ABBINENI, P. S., WITHERS, M. & COORSSEN, J. R. 2017. Coomassie staining provides routine (sub) femtomole in-gel detection of intact proteoforms: Expanding opportunities for genuine Top-down Proteomics. *Electrophoresis,* 38**,** 3086-3099.

O'FARRELL, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem,* 250**,** 4007-21.

ONG, F. S., DAS, K., WANG, J., VAKIL, H., KUO, J. Z., BLACKWELL, W.-L. B., LIM, S. W., GOODARZI, M. O., BERNSTEIN, K. E. & ROTTER, J. I. 2012. Personalized medicine and pharmacogenetic biomarkers: progress in molecular oncology testing. *Expert review of molecular diagnostics,* 12**,** 593-602.

ONG, S.-E., BLAGOEV, B., KRATCHMAROVA, I., KRISTENSEN, D. B., STEEN, H., PANDEY, A. & MANN, M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics,* 1**,** 376-386.

OST, D., FEIN, A. M. & FEINSILVER, S. H. 2003. Clinical practice. The solitary pulmonary nodule. *N Engl J Med,* 348**,** 2535-42.

PAIK, Y.-K., JEONG, S.-K., OMENN, G. S., UHLEN, M., HANASH, S., CHO, S. Y., LEE, H.-J., NA, K., CHOI, E.-Y. & YAN, F. 2012a. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature biotechnology,* 30**,** 221.

PAIK, Y.-K., OMENN, G. S., UHLEN, M., HANASH, S., MARKO-VARGA, G. R., AEBERSOLD, R., BAIROCH, A., YAMAMOTO, T., LEGRAIN, P. & LEE, H.-J. 2012b. Standard guidelines for the chromosome-centric human proteome project. *Journal of proteome research,* 11**,** 2005-2013.

PARKIN, D. M., BRAY, F., FERLAY, J. & PISANI, P. 2005. Global cancer statistics, 2002. *CA: a cancer journal for clinicians,* 55**,** 74-108.

PERCY, A. J., CHAMBERS, A. G., YANG, J., HARDIE, D. B. & BORCHERS, C. H. 2014. Advances in multiplexed MRM-based protein biomarker quantitation toward clinical utility. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics,* 1844**,** 917-926.

PICOTTI, P. & AEBERSOLD, R. 2012. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nature methods,* 9**,** 555.

PIERCE, B. L. & AHSAN, H. 2011. Genome-wide" pleiotropy scan" identifies HNF1A region as a novel pancreatic cancer susceptibility locus. *Cancer research***,** canres. 0124.2011.

RECHE, P. A. & REINHERZ, E. L. 2003. Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *Journal of molecular biology,* 331**,** 623-641.

RIFAI, N., GILLETTE, M. A. & CARR, S. A. 2006. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature biotechnology,* 24**,** 971.

ROSS, P. L., HUANG, Y. N., MARCHESE, J. N., WILLIAMSON, B., PARKER, K., HATTAN, S., KHAINOVSKI, N., PILLAI, S., DEY, S. & DANIELS, S. 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics,* 3**,** 1154-1169.

RUGGERO, D. & PANDOLFI, P. P. 2003. Does the ribosome translate cancer? *Nature Reviews Cancer,* 3**,** 179.

RYFFEL, G. 2001. Mutations in the human genes encoding the transcription factors of the hepatocyte nuclear factor (HNF) 1 and HNF4 families: functional and pathological consequences. *Journal of molecular endocrinology,* 27**,** 11-29.

SAJJAD, W., RAFIQ, M., ALI, B., HAYAT, M., ZADA, S. & KUMAR, T. 2016. Proteogenomics: New Emerging Technology. *HAYATI Journal of Biosciences,* 23**,** 97-100.

SAUNA, Z. E. & KIMCHI-SARFATY, C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics,* 12**,** 683.

SCHAUB, M. A., BOYLE, A. P., KUNDAJE, A., BATZOGLOU, S. & SNYDER, M. 2012. Linking disease associations with regulatory information in the human genome. *Genome research,* 22**,** 1748-1759.

SCHIRLE, M., HEURTIER, M.-A. & KUSTER, B. 2003. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics,* 2**,** 1297-1305.

SHEVCHENKO, A., TOMAS, H., HAVLI, J., OLSEN, J. V. & MANN, M. 2006. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature protocols,* 1**,** 2856.

SHEYNKMAN, G. M., SHORTREED, M. R., FREY, B. L. & SMITH, L. M. 2013. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Molecular & Cellular Proteomics***,** mcp. O113. 028142.

SHIAU, A. L., TEO, M. L., CHEN, S. Y., WANG, C. R., HSIEH, J. L., CHANG, M. Y., CHANG, C. J., CHAO, J., CHAO, L., WU, C. L. & LEE, C. H. 2010. Inhibition of experimental lung metastasis by systemic lentiviral delivery of kallistatin. *BMC Cancer,* 10**,** 245.

SIEGEL, R., MA, J. M., ZOU, Z. H. & JEMAL, A. 2014. Cancer Statistics, 2014. *Ca-a Cancer Journal for Clinicians,* 64**,** 9-29.

SIEGEL, R. L., MILLER, K. D. & JEMAL, A. 2015. Cancer statistics, 2015. *CA: a cancer journal for clinicians,* 65**,** 5-29.

SIEGEL, R. L., MILLER, K. D. & JEMAL, A. 2017. Cancer statistics, 2017. *CA: a cancer journal for clinicians,* 67**,** 7-30.

SILVERA, D., FORMENTI, S. C. & SCHNEIDER, R. J. 2010. Translational control in cancer. *Nature Reviews Cancer,* 10**,** 254-266.

SINITCYN, P., RUDOLPH, J. D. & COX, J. 2018. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science,* 1**,** 207-234.

SNIJDERS, A. P., HUNG, M.-L., WILSON, S. A. & DICKMAN, M. J. 2010. Analysis of arginine and lysine methylation utilizing peptide separations at neutral pH and electron transfer dissociation mass spectrometry. *Journal of the American Society for Mass Spectrometry,* 21**,** 88-96.

SONENBERG, N. 1993. Translation factors as effectors of cell growth and tumorigenesis. *Current opinion in cell biology,* 5**,** 955-960.

STEEN, H. & MANN, M. 2004. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology,* 5**,** 699.

SUPEK, F., MIÑANA, B., VALCÁRCEL, J., GABALDÓN, T. & LEHNER, B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell,* 156**,** 1324-1335.

TEAM, N. L. S. T. R. 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine,* 365**,** 395.

THELEN, J. J. & PECK, S. C. 2007. Quantitative proteomics in plants: choices in abundance. *The Plant Cell,* 19**,** 3339-3346.

THOMPSON, A., SCHÄFER, J., KUHN, K., KIENLE, S., SCHWARZ, J., SCHMIDT, G., NEUMANN, T. & HAMON, C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry,* 75**,** 1895-1904.

TIMMS, J. F. & CRAMER, R. 2008. Difference gel electrophoresis. *Proteomics,* 8**,** 4886-4897.

TSIATSIANI, L. & HECK, A. J. 2015. Proteomics beyond trypsin. *The FEBS journal,* 282**,** 2612-2626.

VADIVEL, A. & ARUN, K. 2015. Gel-based proteomics in plants: time to move on from the tradition. *Frontiers in plant science,* 6**,** 369.

VAN DER WEYDEN, L., RANZANI, M. & ADAMS, D. J. 2014. Cancer gene discovery goes mobile. *Nature genetics,* 46**,** 928.

VASICEK, L. & BRODBELT, J. S. 2010. Enhancement of ultraviolet photodissociation efficiencies through attachment of aromatic chromophores. *Analytical chemistry,* 82**,** 9441-9446.

VIZCAÍNO, J. A., CÔTÉ, R. G., CSORDAS, A., DIANES, J. A., FABREGAT, A., FOSTER, J. M., GRISS, J., ALPI, E., BIRIM, M. & CONTELL, J. 2012. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research,* 41**,** D1063-D1069.

WANG, C., DAVILA, J. I., BAHETI, S., BHAGWATE, A. V., WANG, X., KOCHER, J.-P. A., SLAGER, S. L., FELDMAN, A. L., NOVAK, A. J. & CERHAN, J. R. 2014. RVboost: RNA-seq variants prioritization using a boosting method. *Bioinformatics,* 30**,** 3414-3416.

WANG, C. R., CHEN, S. Y., WU, C. L., LIU, M. F., JIN, Y. T., CHAO, L. & CHAO, J. 2005. Prophylactic adenovirus-mediated human kallistatin gene therapy suppresses rat arthritis by inhibiting angiogenesis and inflammation. *Arthritis Rheum,* 52**,** 1319-24.

WASHBURN, M. P., WOLTERS, D. & YATES III, J. R. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology,* 19**,** 242.

WASINGER, V. C., CORDWELL, S. J., CERPA-POLJAK, A., YAN, J. X., GOOLEY, A. A., WILKINS, M. R., DUNCAN, M. W., HARRIS, R., WILLIAMS, K. L. & HUMPHERY-SMITH, I. 1995. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis,* 16**,** 1090-4.

WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R. M., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C., STUART, J. M. & NETWORK, C. G. A. R. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics,* 45**,** 1113.

WESTERMEIER, R. 2016. 2D gel-based Proteomics: there's life in the old dog yet. *Arch Physiol Biochem,* 122**,** 236-237.

WILHELM, M., SCHLEGL, J., HAHNE, H., GHOLAMI, A. M., LIEBERENZ, M., SAVITSKI, M. M., ZIEGLER, E., BUTZMANN, L., GESSULAT, S. & MARX, H. 2014. Mass-spectrometry-based draft of the human proteome. *Nature,* 509**,** 582.

WILM, M. & MANN, M. 1996. Analytical properties of the nanoelectrospray ion source. *Analytical chemistry,* 68**,** 1-8.

WILM, M., SHEVCHENKO, A., HOUTHAEVE, T., BREIT, S., SCHWEIGERER, L., FOTSIS, T. & MANN, M. 1996. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature,* 379**,** 466.

WISHART, D. S., JEWISON, T., GUO, A. C., WILSON, M., KNOX, C., LIU, Y., DJOUMBOU, Y., MANDAL, R., AZIAT, F. & DONG, E. 2012. HMDB 3.0—the human metabolome database in 2013. *Nucleic acids research,* 41**,** D801-D807.

WONG, B. W., LUK, J. M., NG, I. O., HU, M. Y., LIU, K. D. & FAN, S. 2003. Identification of liver–intestine cadherin in hepatocellular carcinoma—a potential disease marker. *Biochemical and biophysical research communications,* 311**,** 618-624.

WU, C. C., HSU, C. W., CHEN, C. D., YU, C. J., CHANG, K. P., TAI, D. I., LIU, H. P., SU, W. H., CHANG, Y. S. & YU, J. S. 2010. Candidate serological biomarkers for cancer identified from the secretomes of 23 cancer cell lines and the human protein atlas. *Mol Cell Proteomics,* 9**,** 1100-17.

WU, W. W., WANG, G., BAEK, S. J. & SHEN, R.-F. 2006. Comparative study of three proteomic quantitative methods, DIGE, cICAT, and iTRAQ, using 2D gel-or LC– MALDI TOF/TOF. *Journal of proteome research,* 5**,** 651-658.

XIAO, B., WANG, Y., LI, W., BAKER, M., GUO, J., CORBET, K., TSALIK, E. L., LI, Q. J., PALMER, S. M., WOODS, C. W., LI, Z. G., CHAO, N. J. & HE, Y. W. 2013. Plasma microRNA signature as a noninvasive biomarker for acute graft-versus-host disease. *Blood,* 122**,** 3365-3375.

YAMAGATA, K., ODA, N., KAISAKI, P. J., MENZEL, S., FURUTA, H., VAXILLAIRE, M., SOUTHAM, L., COX, R. D., LATHROP, G. M. & BORIRAJ, V. V. 1996. Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3). *Nature,* 384**,** 455.

YAO, X., FREAS, A., RAMIREZ, J., DEMIREV, P. A. & FENSELAU, C. 2001. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Analytical chemistry,* 73**,** 2836-2842.

YATES, J. R., MCCORMACK, A. L., SCHIELTZ, D., CARMACK, E. & LINK, A. 1997. Direct analysis of protein mixtures by tandem mass spectrometry. *Journal of protein chemistry,* 16**,** 495-497.

YONEDA, K., IIDA, H., ENDO, H., HOSONO, K., AKIYAMA, T., TAKAHASHI, H., INAMORI, M., ABE, Y., YONEDA, M. & FUJITA, K. 2009. Identification of Cystatin SN as a novel tumor marker for colorectal cancer. *International journal of oncology,* 35**,** 33-40.

YOON, S. H., MOON, J. H., CHUNG, Y. J. & KIM, M. S. 2009. Influence of basic residues on dissociation kinetics and dynamics of singly protonated peptides: time-resolved photodissociation study. *Journal of mass spectrometry,* 44**,** 1532-1537.

YOST, R. & ENKE, C. 1978. Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society,* 100**,** 2274-2275.

YOST, R. & ENKE, C. 1979. Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Analytical chemistry,* 51**,** 1251-1264.

ZHANG, B., WANG, J., WANG, X., ZHU, J., LIU, Q., SHI, Z., CHAMBERS, M. C., ZIMMERMAN, L. J., SHADDOX, K. F. & KIM, S. 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature,* 513**,** 382.

ZHANG, X., GUO, J., FAN, S. F., LI, Y. Y., WEI, L. L., YANG, X. Y., JIANG, T. T., CHEN, Z. L., WANG, C., LIU, J. Y., PING, Z. P., XU, D. D., WANG, J. X., LI, Z. J., QIU, Y. Q. & LI, J. C. 2013. Screening and Identification of Six Serum microRNAs as Novel Potential Combination Biomarkers for Pulmonary Tuberculosis Diagnosis. *Plos One,* 8.

ZHU, B., LU, L., CAI, W., YANG, X., LI, C., YANG, Z., ZHAN, W., MA, J. X. & GAO, G. 2007. Kallikrein-binding protein inhibits growth of gastric carcinoma by reducing vascular endothelial growth factor production and angiogenesis. *Mol Cancer Ther,* 6**,** 3297-306.

ZUCMAN-ROSSI, J., JEANNOT, E., VAN NHIEU, J. T., SCOAZEC, J. Y., GUETTIER, C., REBOUISSOU, S., BACQ, Y., LETEURTRE, E., PARADIS, V. & MICHALAK, S. 2006. Genotype–phenotype correlation in hepatocellular adenoma: new classification and relationship with HCC. *Hepatology,* 43, 515-524.

국문초록

프로테오지노믹스 기법을 이용한 폐암 바이오마커 연구

김 용 인

서울대학교 대학원
수의학과 수의생명과학 전공

지도교수 조 제 열

정밀의료 패러다임의 등장 이후, 질환의 진단 및 치료를 위해서 바이오마커에 대한 수요는 높아지고있다. 가설기반연구는 전통적으로 당연하게 사용되오던 연구수행체계이지만, 바이오마커 발굴에서 필연적으로 마주치게되는 광범위한 스크리닝 작업에서는 효율성의 한계를 드러낸다. 오믹스기술의 등장과 함께 질환연구의 패러다임은 증거기반 대규모 타겟발굴방식으로 변화하고 있다.

단백질은 생체 기능조절에 직접적으로 관여하는 물질이기 때문에 바이오마커로 활용할 수 있는 가장 이상적인 물질로 여겨진다. 질량분석기를 이용한 단백체분석은 단백질을 직접 정성 및 정량할 수 있을 뿐만 아니라 매우 생산성이 높아 질환 바이오마커 발굴에 유용하다. 이 논문에서는 질량분석기를 이용하여 폐암 바이오마커의 발굴을 위한 고도화된 분석기법인 프로테오지노믹스 기법의 적용과, 스크리닝된 바이오마커후보 단백질의 정량검증 및 폐암 감별진단 조합마커의 생성연구에 대하여 알아본다.

CHAPTER I 에서는 인간염색체기반 단백체프로젝트 (C-HPP)의 일환으로 수행된 염색체 9 번에 대한 단백체연구가 포함되어있다. 미확인 단백질과 유전단백체에서 발견되지 않았던 시그니처를 밝혀내기 위해 LC-MS/MS 분석과 RNA-seq 차세대염기분석기법을 적용하여 샘암종 폐암환자 5 명의 정상-종양조직을 분석하였다. 염색체중심-인간단백체프로젝트의 2013 년 리포트에서는 neXtProt 인간단백체 데이터베이스를 기준으로 염색체 9 번에서 170 개의 미확인 단백질이 있는 것으로 알려졌으며, 본 논문의 연구가 진행된 2015 년에는 133 개가 계속 미확인상태로 남아있었다. 본 논문의 단백체분석에서는 19 개의 미확인 단백질을 동정할 수 있었으며, 그 중에서 염색체 9 번에 해당하는 단백질은 SPATA31A4 한 개 였다. RNA-seq 분석으로는 샘종폐암조직 5 개에서 공통적으로 검출되면서 정상조직에서는 검출되지 않는 nonsynonymous SNP

5 개 (CDH17, HIST1H1T, SAPCD2, ZNF695) 그리고 synonymous SNP 3 개를 발굴할 수 있었다.

프로테오지노믹 분석을 위해서 각 시료별 RNA-seq 데이터를 가공하여 맞춤형 데이터베이스를 구축하였다. 이렇게 생성된 시료맞춤형 데이터베이스를 단백체 질량분석데이터 검색에 활용하여 5 개 유전자(LTF, HDLBP, TF, HBD, HLA-DRB5)에 해당하는 7 개의 돌연변이를 검출하였다. 두 개의 돌연변이는 정상조직에서는 검출되지 않고 암조직에서만 검출되었다. 또한, 이 결과에서는 정상-암조직 모두에서 위유전자 (EEF1A1P5) 펩티드를 검출할 수 있었다.

CHAPTER II 에서는 다중반응검지법 (MRM) 을 이용한 단백질 바이오마커 검증과 조합마커 구성에 대한 연구를 서술하였다. 폐암과 다른 폐질환은 감별이 어렵기 때문에 폐암은 오진단 위험이 큰 질병이다. 따라서 혈청기반의 폐암감별진단 바이오마커개발의 필요성은 널리 인정되고있다. 이 단원에서는 폐암환자와 대조군폐질환 환자 198 명의 혈청시료를 활용하여 일곱개의 폐암바이오마커 후보단백질을 나노유속 액체크로마토그래피-다중반응검지법으로 정량하였다.

후보단백질을 개별로 분석하였을 때에는 SERPINA4 만이 통계적으로 유의성있게 혈중농도가 감소하는 것으로 나타났다.

다중반응검지법 전체데이터를 임상정보와 함께 로지스틱회귀모델에 적용하여 하나의 조합마커로 만들 수 있었다. 이 과정에서 개별마커로는 통계적인 유의성이 두드러지지 않지만 간섭효과를 만들어낼 수 있는 변수를 고려하여 모델링을 진행하였다. 최종적으로 SERPINA4, PON1, 나이를 조합하였을 때 가장 최적의 조합마커가 생성되었다. 이 조합마커는 AUC 0.915 의 감별진단 성능을 보여주었으며, 모델을 만드는데 사용되었던 시료와는 별개의 검증군에서도 성능은 유지되었다. 이와 같이 통계모델을 이용하여 생성한 조합마커는 개별 분자마커를 이용했을 경우보다 개선된 폐암 감별진단능력을 보여줄 수 있음을 제시한다.

주요어: 폐암, 액체크로마토그래피-탠덤질량분석기;
프로테오지노믹스; 단백체학; 다중반응검지법

학번: 2013-21536