



의학박사 학위논문

C18 columns 으로 단백질 정제와 농축 후 MALDI-TOF MS 시스템을 이용한 혈청 단백질 분석을 통한 유방암 진단

Serum Protein Breast Cancer Diagnosis using MALDI-TOF MS System after protein purification and concentration by C18 columns

> 울산대학교 대학원 의 학 과 최 수 정

C18 columns 으로 단백질 정제와 농축 후 MALDI-TOF MS 시스템을 이용한 혈청 단백질 분석을 통한 유방암 진단

지도교수 이 새 별

이 논문을 의학박사 학위 논문으로 제출함

2022년 08월

울산대학교대학원

- 의학과
- 최 수 정

최수정의 의학박사학위 논문을 인준함

- 심사위원 장수환 (인)
- 심사위원 이새별 (인)
- 심사위원 김희정 (인)
- 심사위원 정일용 (인)
- 심사위원 곽성찬 (인)

울산대학교 대학원

2022년 08월

국문요약

목적

유방암은 여성에서 가장 많이 발생하는 암으로 발병률은 전세계적으로 증가하고 있다. 따라서, 조기 유방암 진단에 도움이 될 수 있는 바이오 마커를 발견하는 것은 매우 중요하다. 실제로 혈액을 통한 CA15-3, CEA 등의 종양표지자가 있으나 특이도와 민감도가 낮아 그 효용성이 떨어진다. 진단에 유용한 단백질/펩타이드의 바이오 마커를 확립하기 위한 단백질 연구가 폭넓게 진행되어 왔다. 임상 단백질체학은 질병 발생 중 변화를 연구하기 위해 다양한 펩타이드와 단백질의 분석에 초점을 맞추고 있다. 임상 단백질체학은 여러 질병의 임상 진단 뿐 만 아니라 다양한 암 치료 방법에 있어서 반응 정도를 평가하는 데까지 활용 가능 하다. 이것은 건강한 대조군과 암 환자의 검체 사이의 단백질 발현 또는 프로파일 차이를 비교하여 확인할 수 있으며 1,000 ~ 10,000Da 의 저 분자량 혈청 펩타이드의 경우 중요한 바이오 마커 정보를 포함하고 있어, 많은 질병의 조기 진단을 위한 마커로 사용될 수 있다고 보고되었다. 단백질을 분석함에 있어, 인간의 혈청 샘플은 쉽게 얻을 수 있으며 다양한 기법과 기구에 의해 분리 및 분석할 수 있는 장점을 가지고 있어 진단 도구로 유용하다.

대상 및 방법

본 연구는 서울 아산병원에서 건강 검진 시 획득한 30 명의 건강한 사람의 혈청 샘플과 30 명의 유방암 환자의 혈청 샘플로 총 60 개의 혈청 샘플을 얻었다. 혈청 단백질 또는 펩타이드 분자는 Peptide Cleanup C18 pipette tips (Cat # 5188-5239, Agilent Technologies)를 사용하여 프로토콜에 따라 정제 및 농축하여 MALDI-TOF MS 를 통해 질량 분석하였다. 질량 스펙트럼은 m/z 및 절대 강도 값을 추출하여 TIC 로 정규화 하였으며 heatmap, PCA plots 및 ROC 곡선을 작성했다. m/z 값과 그에 상응하는 TIC 정규화 된 강도 값은 NosID[™] (NosQuest, 대한민국)이라는 소프트웨어를 통해 특이도, 민감도 및 정확도 값과 함께 기록된다.

결과

Heatmap을 통해 건강한 대조군 30명과 유방암 환자 30명의 통계적으로 유의한 m/z 피크의 패턴 차이를 보여줄 수 있었고 m/z 마커에 대한 건강한 대조군과 유방암 환 자 사이의 정규화 된 강도 값은 통계적으로 유의한 차이를 보여주었다. 특히 4,150 m/z 에서 분명한 차이를 볼 수 있었다. 본 연구의 진단 능력을 분석하기 위한 ROC 곡선을 그렸으며 0.932의 AUC 값으로 검증되었다. NosID[™] 프로그램을 사용하여 높은 특이도,

1

민감도 및 정확도 값을 얻어 확인하였고 건강한 사람과 유방암 환자 사이에서 TIC 정규 화 강도 값에서 유의하게 차이를 보이는 6개의 질량 값을 확인하였다.

결론

본 연구에 사용된 샘플과 향후 연구에 사용된 샘플 간에 겹치는 m/z 값이 많이 있는 경우 이러한 m/z 값은 더 뛰어난 유방암 바이오 마커가 될 것이다. 또한 샘플 준비 단계의 변동을 최소화하기 위해 새롭고 개선된 방법이 요구되는데 실제로 변화에 취약할 수 있는 단계 중 하나인 Peptide Cleanup C18 pipet tip을 사용할 때의 혈청 단백질 정제 단계를 균일 화할 수 있는 더 많은 연구가 향후에 도움이 될 수 있으며, 이는 데이터를 보다 균일하고 정확하게 만들 수 있다. 또한 자동화된 샘플 준비 시스템을 사용하면 시 스템이 인적 오류 및 변형을 제거하므로 보다 정확한 데이터 수집 과 바이오 마커 확인 에 유리할 것이다.

중심단어

-breast neoplasms, biomarker, heatmap, MALDI-TOF MS, Peptide

국문요약	1-2
표 및 그림 차례	4
약어목록	5
서론	6
연구대상 및 연구방법	8
1.연구 대상	
2. 연구 방법	
1) 혈청 검체 준비	
2) 단백질/펩타이드 정화 및 농축	
3) MLDI-TOF MS에 의한 질량분석 해석 데이터 처리 및 정규화	
4) 통계분석	
5) 암 검진 및 진단 분석	
결과	
고찰	
결론	22
참고문헌	23
영문요약	27

표 및 그림 차례

 Table 1. Patient demographic and clinical information

Table 2. NosID results for the healthy controls and breast cancer patients using the p-value weighted algorithm.

Table 3. NosID results for the healthy controls and breast cancer patients using the random forest-based algorithm.

Table 4. m/z values and TIC normalized intensity values between the healthy controls (n = 30) and the breast cancer patients (n = 30)

Table 5. Additional m/z values and TIC normalized intensity values between the healthy controls (n = 30) and the breast cancer patients (n = 30)

Table 6. Human serum protein biomarkers (Table 4) and their possible protein identities based on the comparison to the list from the UniProt database.

Table 7. Human serum protein biomarkers (Table 5) and their possible protein identities based on the cancer biomarker.

Fig. 1. TIC normalized intensity comparisons between the healthy controls (n = 30) and the breast cancer patients (n = 30) for the different masses (m/z)

Fig. 2. Heatmap clusters showing the pattern differences of the statistically significant m/z peaks between 30 healthy controls and 30 breast cancer patients.

Fig. 3. Waterfall plot showing the spectral pattern differences between the statistically significant m/z peaks between 30 healthy controls and 30 breast cancer patients.

Fig. 4. Principal Component Analysis (PCA) plots of the TIC normalized intensity values between the two groups. **Left**: 2D PCA plot; **Right**: 3D PCA plot

Fig. 5. The ROC curves for the healthy controls and breast cancer patients

Fig. 6. Scatter plot of m/z values and TIC normalized intensity values between the healthy controls (n = 30) and the breast cancer patients (n = 30) A: 2161m/z B:4153 m/z C: 4600 m/z D: 7618 m/z E: 8503 m/z F: 9575 m/z

약어목록

- CA 15-3: Cancer antigen 15-3
- CEA: Carcinoembryonic Antigen
- ELISA: Enzyme-linked immunosorbent assay
- LC-MS/MS: Liquid chromatography tandem-mass spectrometry
- MALDI-TOF MS: Mass-assisted laser desorption/ionization time-of-flight mass spectrometry
- m/z: mass-to-charge ratio
- TFA: Trifluoroacetic acid
- ACN: Acetonitrile
- TIC: Total Ion Count
- PCA plot: Principal Component Analysis plot
- ROC: Receiver operating characteristic
- AUC: Area under the curve
- SELDI-TOF MS: Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry

유방암은 여성에서 가장 많이 발생하는 암으로 WHO 세계보건기구에 따르면 2020년 한 해 동안 230만 명의 여성들에게 발생하였고 약 68만 5천 명의 여성이 유방암으로 사망하였다. 또한 발병률은 전세계적으로 증가하고 있다[1]. 따라서, 조기 유방암 진단 에 도움이 될 수 있는 바이오 마커를 발견하는 것은 중요하다. 실제로 혈액을 통해 Cancer antigen 15-3 (CA15-3), Carcinoembryonic antigen (CEA) 등의 종양표지자가 있으 나 특이도와 민감도가 낮아 그 효용성이 떨어진다. 진단에 유용한 단백질/펩타이드 바 이오 마커를 확립하기 위한 단백질 연구가 폭넓게 진행되어 왔다. 다양한 단백질체학 기술은 관절낭액, 혈액 및 소변 샘플의 단백질을 분석함으로써 류마티스 관절염과 같 은 다양한 질병의 임상 진단의 발전을 가져왔다[2]. 또한, 예측 바이오 마커의 발견은 어떤 환자가 다양한 암 치료 방법에 있어서 더 잘 반응할지를 식별하는 데 매우 중요 하다[3-7]. 임상 단백질체학은 질병 발생 중의 변화를 연구하기 위해 다양한 펩타이드 와 단백질의 분석에 초점을 맞추고 있다. 이것은 건강한 대조군과 암 환자의 검체 사 이의 단백질 발현 또는 프로파일 차이를 비교함으로써 이루어진다[3, 8-11]. 또한 1,000 ~ 10,000Da의 저 분자량 혈청 펩타이드는 중요한 바이오 마커 정보를 포함하고 있으 며 많은 질병의 조기 진단을 위한 마커로 사용될 수 있다고 보고되었다[12-13]. 단백질 을 분석함에 있어, 인간의 혈청 샘플은 쉽게 얻을 수 있으며 암 관련 바이오 마커 발 견을 위한 임상적으로 가장 중요한 체액 중 하나이다[14]. 일반적으로 건강검진시 환자 의 혈액 샘플을 이용하여 혈청을 얻을 수 있기 때문에 이를 활용하여 다양한 기법과 기구에 의해 분리 및 분석할 수 있다.

Enzyme-linked immunosorbent assay (ELISA)[12-13]와 같은 면역측정 기법이 마커 분석 에 자주 사용된다. ELISA는 특정 암 바이오 마커의 정량화 분석은 가능하지만 다중 바이 오 마커 분석에는 몇 가지 제한이 있고 개발 초기 비용이 높다[12-13]. 반면에, 질량 분 석법은 높은 민감도 때문에 펩타이드와 단백질을 포함한 다양한 작은 분자와 큰 분자를 분석하는 데 자주 사용되어 왔다[11,15-16]. Liquid chromatography tandem-mass spectrometry (LC-MS/MS)은 높은 특이도와 민감도 및 정량적 재현성 때문에 혈청 단백 질 분석에 일반적으로 사용되지만, 높은 처리량 분석에는 적합하지 않다[12-13]. 이와 관 련하여 Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS)은 신속한 정성 및 정량 분석을 위한 새로운 기술이다. MALDI-TOF MS의 경우, 레이저로 이온화를 통해 다양한 분자를 활성화시키고, 활성화된 분자는 검출기를 향해 가속된다. 여기서 각기 다른 이온에 대한 비행 시간 값이 측정되고 이는 질량 대

6

전하 비율 값 mass-to-charge ratio (m/z)으로 변환된다. 본 연구에서는 정제된 혈청단백 질을 MALDI-TOF MS 시스템으로 직접 분석하였으며, 단백질 신호를 이용하여 건강한 대 조군을 유방암 환자 샘플과 비교하였다.

이 연구에서 사용된 총 60개의 샘플 중 30개는 건강한 대조군이었고 나머지 30개는 유방암 환자에서 채취한 것으로 2020.08.01부터 2020.12.31 동안 서울아산병원에서 수집 되었다. 혈청 검체를 먼저 희석한 후 C18 column purification technique 기법을 사용하여 정제하였다. 정제된 단백질은 유방암 환자와 건강한 대조군을 구별하는 데 사용될 수 있 는 의미 있는 단백질 바이오 마커를 식별하기 위해 분석되었다. 본 연구의 주요 목적은 MALDI-TOF MS에 의한 정제 혈청 단백질 분석이 건강한 유방암 환자를 진단하고 의미 있는 m/z 마커를 식별해 내는데 충분한지 여부를 결정하는 것이다.

연구대상 및 방법

1. 연구대상

본 연구는 서울 아산병원에서 2020.08.01부터 2020.12.31 동안 건강 검진 시 획득한 30명의 건강한 사람의 혈청 샘플과 30명의 유방암 환자의 혈청 샘플로 총 60개의 혈청 샘플을 얻었다. 모든 참여자가 사전 동의서에 서명했으며, 본 연구에 사용된 모든 혈청 샘플은 기관심사위원회(IRB 승인 번호: 2018-1234)의 승인을 받았다. 최소 환자 연령은 34세, 최대 환자 연령은 64세였다. 유방암 검체의 경우 20개는 2기, 나머지 10개는 3기 였다.

본 연구는 서울 아산병원에서 건강 검진 시 획득한 30명의 건강한 사람의 혈청 샘플 과 30명의 유방암 환자의 혈청 샘플로 총 60개의 인간 혈청 샘플을 얻었다.

2. 연구방법

1) 혈청 검체 준비

단백질 정제 전에 혈청 샘플은 0.5% Trifluoroacetic acid (TFA) (v/v)를 함유한 물로 먼저 1:10 희석되었다. C18 silica pipette tips를 사용하기 전에 혈청 시료 5 μL를 TFA 0.5%의 탈 이온수 (Deionized Water: DW) 45μL에 혼합하였다.

2) 단백질/펩타이드 정화 및 농축

혈청 단백질 또는 펩타이드 분자는 Peptide Cleanup C18 pipette tip를 사용하여 제공된 프로토콜에 따라 정제 및 농축하였다. C18 Pipette tip을 10-uL pipet에 부착하고, 10 μL의 Acetonitrile (ACN) 용액을 tip에 흡인하여 분주한 다음, 10 μL의 TFA를 물에 분주하여 제 조하였다. 그런 다음 희석 혈청 검체의 10 μL를 20회 흡입 및 분주하여 팁에 시료를 결 합하였다. 팁은 물에 0.1% TFA 용액으로 두 번 씻은 후 결합 단백질과 펩타이드들은 팁 에 0.1% TFA가 함유된 60% ACN 용액의 7.5uL를 흡입하여 10번 원심분리관으로 용출하 였다.

3) MALDI-TOF MS에 의한 질량분석 해석

질량분석 분석은 IDSys LT(ASTA, 한국)를 사용하여 수행되었다. 용출된 단백질과 펩타 이드 샘플 4µL를 0.1% TFA 용액과 70% ACN의 20mg/mL 4µL와 혼합하고 혼합물의 1.5µL 를 96 target MALDI plates (ASTA)에 샘플 당 네 곳에 분주한다. 각각의 MALDI plate spot 위치마다 40회의 레이저를 각기 다른 30개 위치에 쏘았다. 그 결과 분석의 질량 범위는 2000~50,000이었고 MALDI-TOF MS 분석은 선형 양성 모드에서 수행되었다. 레이저로 이온화를 통해 다양한 분자를 활성화시키고, 활성화된 분자는 검출기를 향해 가속된다. 여기서 각기 다른 이온에 대한 비행 시간 값이 측정되고 이는 질량 대 전하 비율 값 mass-to-charge ratio (m/z)으로 변환된다.

4) 데이터 처리 및 정규화

원시 질량 스펙트럼은 IDSys LT 프로그램을 사용하여 모든 질량 스펙트럼에 나타난 8 개의 주요 m/z 피크를 사용하여 질량 보정되었다. 보정된 질량 스펙트럼은 m/z 및 절대 강도 값을 추출하여 피크 목록으로 변환되었으며, 결과 피크 목록을 XML 파일로 내보내 고 내보낸 XML 파일은 CSV 파일로 변환되었으며, 그 후 통계적으로 분석될 임상 샘플 에 대한 질량 및 강도 값을 수집하였다. 통계 분석을 수행하기 전에, 모든 절대 강도 데 이터는 각 강도 값을 각 표본의 강도 값의 합으로 나누어 총 이온 계수 Total Ion Count(TIC)로 정규화 하였다.

5) 통계 분석

TIC 정규화 데이터는 본 연구의 모든 통계 분석의 기초로 사용되었다. 우선, heatmap, Principal Component Analysis (PCA) plots 및 Receiver Operating Characteristic (ROC) 곡 선이 표시되었다. Heatmap과 PCA plots은 Perseus 프로그램을 사용하여 만들어졌다. 프 로그램에 대한 입력은 분류에 따라 정렬된 대조군과 유방암환자의 순으로 임상 샘플에 대한 질량 대 TIC 정규화 된 강도를 나열한 텍스트 파일로 다중 표본 분산 분석은 p-값 0.05를 기준으로 수행되었으며 유의한 값의 마커만 heatmap과 PCA plots을 만드는데 사 용하였다. 동일한 추가 정규화 없이 다변량 ROC 곡선 설정에 대한 random forest algorithm 하에 Metaboanalyst 라는 온라인 툴을 사용하여 ROC 곡선을 작성했다.

9

6) 암 검진 및 진단 분석

m/z 값과 그에 상응하는 TIC 정규화 된 강도 값은 NosID[™] (NosQuest, 대한민국)이라 는 소프트웨어를 통해 나타냈다. 이 프로그램은 진단 알고리즘을 적용하는 데 사용하였 는데 TIC 정규화 된 강도 값과 두 임상 그룹 간의 차이를 기반으로 컷오프 값을 자동으 로 계산했으며, 이 값은 검체가 건강한 대조군 또는 유방암 샘플로 선별되었는지 여부를 결정하기 위한 진단 기준 값으로 사용되었다. 두 가지 다른 진단 알고리즘이 사용되었는 데, 하나는 p-값 가중치 방법에 기초한 것이고 다른 하나는 random forest modeling에 기초한 것이다. 결과는 NosID[™] 분석 후 특이도, 민감도 및 정확도 값으로 기록되었다. 정제된 단백질과 펩타이드는 아산병원의 건강한 대조군 (N, n=30) 및 유방암군(BR, n=30) 각각에서 MALDI-TOF MS에 의해 분석되었다. 환자 특성은 표 1과 같다.

Characteristics	Breast Cancer Patients	Controls
No. of patients	30	30
Gender		
Male	0	0
Female	30	30
Age (Year)		
Mean (SD)	45 (5.8)	52 (7.7)
Range	36-56	34-64
Tumor Stage (No. of patients)		
Stage 1	0	
Stage 2	20	
Stage 3	10	
Stage 4	0	
CA15-3		
Mean (SD)	11.05 (4.75)	
Range	3.9-22.7	
Receptor types (No. of patients)		
HR+/HER2-	24	
HR+/HER2+	2	
HR-/HER2+	0	
HR-/HER2-	3	
Unknown	1	

건강한 대조군과 유방암 환자 간의 단백질 TIC 정규화 강도 차이

관측된 m/z 피크에 대한 모든 절대 강도 값은 TIC에 의해 정규화 되었다. 건강한 대조 군(n = 30)과 유방암 환자(n = 30) 사이의 TIC 표준화 강도 차이를 그림 1에서 나란히 비 교하였다. 그림 1에 표시된 m/z 값 중에는 유의하게 차이가 없는 분자량 (m/z)을 포함하 고 있다.

통계적으로 유의한 질량 후보들을 결정하기 위해, heatmap을 생성함으로써 TIC 정규화 된 강도 값을 시각화 하였다. ANOVA 분산 분석은 p-value 분계점이 0.05인 두 그룹에 대해 수행되었고 Z-score는 통계적으로 유의한 질량(p-값 <0.05)에 대해 수행되었다. 통 계적으로 유의한 총 24개의 질량이 검출되었으며, 해당 heatmap clustering 결과는 그림 2와 같다.



Fig. 1. TIC normalized intensity comparisons between the healthy controls (n = 30) and the breast cancer patients (n = 30) for the different masses (m/z)

그림 2는 건강한 대조군과 유방암 환자군 간의 패턴 차이를 명확히 보여주고 있다. 여 기서 빨간색은 TIC 정규화 된 강도 값이 평균보다 높고 녹색은 그 반대임을 나타낸다. 즉, 적색 영역에 해당하는 질량은 상향 조절 마커가 되는 반면 녹색 영역에 해당하는 질 량은 특정 표본 그룹에 대한 하향 조절 마커가 된다(그림 2).



Fig. 2. Heatmap clusters showing the pattern differences of the statistically significant m/z peaks between 30 healthy controls and 30 breast cancer patients.

그림 3은 그림 2에 표시된 m/z 마커에 대한 건강한 대조군과 유방암 환자 사이의 정 규화 된 강도 값 사이의 가시적인 차이를 보여주고 있으며, 정규화 된 강도 값은 통계적 으로 유의하다(p <0.05). 특히 4,150 m/z에서 분명한 차이를 볼 수 있다.



Fig. 3. Waterfall plot showing the spectral pattern differences between the statistically significant m/z peaks between 30 healthy controls and 30 breast cancer patients

건강한 대조군과 유방암 환자 사이의 표본 분포(PCA plot)

그림 4는 각 샘플이 어떻게 분포되어 있는지 또는 샘플이 유사성을 기반으로 군집화 되는지를 보여준다. 파란색 사각형은 정상 샘플에 해당하고 빨간색 사각형은 유방암 샘 플에 해당한다(그림 4). 각 PCA 플롯은 두 샘플 그룹이 군집화 된 패턴을 표시했다. 그림 4의 군집화 된 모습은 두 그룹이 일반적으로 구별 가능하지만 일부 겹치는 영역이 있음 을 보여준다.



Fig. 4. Principal Component Analysis (PCA) plots of the TIC normalized intensity values between the two groups. **Left**: 2D PCA plot; **Right**: 3D PCA plot. Blue: controls Red: breast cancer patients

혈청 단백질 및 펩티드의 TIC 정규화 데이터의 진단 성능/능력

연구의 진단 능력을 분석하기 위해 heatmap 분석과 동일한 데이터 m/z 값 및 TIC 정 규화 강도)를 사용하여 ROC곡선을 그렸다. 유방암 환자의 ROC 곡선을 구성하는데 사 용된 알고리즘은 Random Forest model을 기반으로 했으며 MetaboAnalyst 라는 온라인 도구를 사용하였다. 결과 ROC 곡선 은 그림 5에서 보여주고 있다.

그림 5에 기반하여, 이연구의 진단 능력은 0.932의 AUC 값으로 검증되었다. 이 곡선은 식별 성능이 현재 연구에 대해 탁월했으며 이전에 표시된 heatmap, PCA plot 및 ROC 곡 선에 의해 더욱 강화되었음을 나타낸다 (그림 2, 4, 5).



Fig. 5. The ROC (receiver operating characteristic) curves for the healthy controls and breast cancer patients. **AUC :** Area under the Curve

그림 5. 건강한 대조군과 유방암 환자에 대한 ROC 곡선은 진단 컷 오프를 나타낸다. 가 장 많은 마커 수의 Area under the Curve(AUC) 값은 0.932였다.

NosID에 의한 진단 결과

실제 암 검진이나 진단은 NosID™라는 자체 프로그램을 사용하여 수행되었다. 프로그 램은 m/z 값과 강도 값으로 구성된 입력 데이터를 가져오며, 본 연구에서는 강도 값을 추가 데이터 정규화 없이 TIC 정규화 된 강도 값으로 전환했다. 앞서 언급한 바와 같이 두 가지 알고리즘 방법이 연구에 사용되었으며 그 결과 특이도, 민감도 및 정확도 값은 표 2 및 표 3에 나와 있다.

Table 2. NosID results for the healthy controls and breast cancer patients using the p-value weighted algorithm.

Samples	Specificity	Sensitivity	Accuracy
30 Healthy vs 30 BR	71.7%	87.5%	79.6%

Table 3. NosID results for the healthy controls and breast cancer patients using the random forest-based algorithm.

Samples	Specificity	Sensitivity	Accuracy
30 Healthy vs 30 BR	93.3%	94.9%	94.2%

표 2 와 표 3 두개의 표는 두 진단 알고리즘에 대한 높은 특이도, 민감도 및 정확도 결과 값을 나타낸다. 특이도는 얼마나 많은 건강한 대조군이 진단 분석에 의해 올바르 게 식별되었는지를 나타내고, 민감도는 분석에 의해 얼마나 많은 환자 샘플이 올바르게 식별되었는지를 나타낸다. 정확도 값은 진단의 전반적인 정확성을 나타낸다. 두 테이블은 두 임상 그룹 사이에 TIC 강도 차이에 상당한 차이가 있음을 보여 주었으며 이러한 차이 의 컷오프 값을 사용하여 유방암을 진단하거나 선별할 수 있다.

TIC 정규화 강도의 차이를 기반으로 m/z 값 분석

p-값 가중치 방법 또는 Random Forest model을 기반으로 각 m/z에 대한 TIC 정규화 강도 값 간의 차이를 분석하는 것 외에도 단순히 평균을 취하여 각 m/z에 대한 TIC 정 규화 강도 값을 비교하는 추가 분석이 수행되었다. 건강한 대조군의 TIC 정규화 강도 값 과 각 m/z에 대한 유방암 샘플의 강도 값 사이의 p-값을 계산하여 값에 눈에 띄는 차이 가 있는지 확인하기위한 것으로 가장 차별화되는 바이오 마커 만을 얻기 위해 0.001보다 작은 p-값을 갖는 m/z만 선택하고 결과 m/z 값과 해당하는 TIC 정규화 강도 값을 정리 하였다(표 4). 6개의 질량 값은 건강한 사람과 유방암 환자 사이에서 TIC 정규화 강도 값 에서 유의하게 차이를 보였다.

Mass (m/z)	Controls	Breast cancer patients	p-value
3260	2.34 ± 2.54	7.02 ± 4.14	0.000002
4207	8.95 ± 7.79	17.67 ± 6.61	0.000018
2101	1.19 ± 1.69	3.52 ± 2.66	0.000153
2550	3.92 ± 4.44	8.44 ± 4.79	0.000354
5001	9.53 ± 12.50	0.82 ± 2.03	0.000392
4150	4.88 ± 5.87	1.00 ± 1.60	0.000928

Table 4. m/z values and TIC normalized intensity values between the healthy controls (n = 30) and the breast cancer patients (n = 30)

추가적으로 구별되는 스펙트럼을 갖는 6가지 m/z 값의 TIC 정규 화된 값을 확인해서 비교하였다(표 5). 각 60개의 샘플에서 각 m/z 값의 강도 값의 분포를 확인하고 구별되 는 정도를 비교하여 볼 수 있었다(그림6). 뚜렷한 cut off intensity 값으로 두 군의 sample이 나 완전히 구별되는 m/z 값도 (그림 6. A, D) 있고 일정 범위의 intensity를 공 유하는 값도 있었다. (그림 6. B, C, E, F)

Table 5.	Additional	m/z	values	and	TIC	normalized	intensity	values	between	the	healthy
controls (n = 30) and	the b	preast ca	ancer	patie	ents $(n = 30)$					

Mass (m/z)	Controls	Breast cancer patients
2161	25.1 ± 10.58	77.6 ± 10.71
4153	86.6 ± 15.08	51.1 ± 23.03
4600	8.4 ± 3.58	16.8 ± 3.60
7618	19.0 ± 5.50	5.8 ± 2.25
8503	4.6 ± 2.57	3.6± 1.51
9575	17.2 ± 6.79	11.6 ± 2.07



Fig. 6. Scatter plot of m/z values and TIC normalized intensity values between the healthy controls (n = 30) and the breast cancer patients (n = 30) A: 2161m/z B:4153 m/z C: 4600 m/z D: 7618 m/z E: 8503 m/z F: 9575 m/z

본 연구에서는 표 4의 질량 값의 실제 단백질 또는 펩타이드를 확인할 수 없지만 heatmap, PCA plots, ROC 곡선 및 NosID 결과 (그림 2, 4, 5, 표 2, 3)를 근거로, m/z 마 커는 높은 수준의 암 진단 능력을 입증했다. 인간 혈청을 정제하고 농축한 다음 MALDI-TOF MS로 샘플을 분석함으로써 관찰된 질량 값에 대해 건강한 대조군과 유방암 환자 사이에 상당한 차이가 있었고 이러한 차이는 앞서 설명한 대로 heatmap, PCA plots, ROC 곡선으로 시각화 할 수 있었다. 또한 표 4에서 두 임상 그룹 간의 TIC 정규 화 값의 명확한 차이를 수치 및 통계적으로 잘 관찰할 수 있었다. 이러한 차이는 표 4 의 값이 모두 0.001 미만의 p-값으로 통계적으로 유의하였다. 그러나 이러한 차이가 일 반적으로 다소 큰 데이터 변동을 표시하는 MALDI-TOF MS 시스템의 특성 때문인지 완 전히 확인되지는 않았다. 이러한 큰 변동으로 인해 데이터가 왜곡되거나 어떤 방식으로 든 영향을 받을 수 있으며 통계 분석에서 변동으로 인한 차이가 발생할 수 있다. 그럼 에도 불구하고 위의 표와 그림들은 유방암 환자를 건강한 대조군과 구별하는 데 사용할 수 있는 일부 마커 또는 단백질이 있음을 보여주었다.

획득한 데이터의 품질은 p-값 가중치 방법과 Random Forest model을 기반으로 하는 두 가지 알고리즘을 사용하는 진단 소프트웨어(NosID)로 추가 분석 및 강화되었다(표 2, 표 3). 두 가지 다른 통계 방법은 모두 높은 특이도, 민감도 및 정확도 값을 보여 MALDI-TOF MS에 의한 정제된 혈청 단백질 분석이 유방암을 진단에 도움이 될 수 있음 을 나타낸다. 더 많은 샘플이 데이터베이스에 추가되면 프로그램의 진단 성능이 더욱 향 상될 것이다.

이 연구에서 입증된 뚜렷한 결과에도 불구하고, 암 진단을 위해 혈청 단백질을 분석하 기 위해 MALDI-TOF MS를 사용 여부를 결정하기 위해서는 추가 연구가 필요하다. 첫째, 통계적으로 유의하게 관찰된 m/z 값(표 4)의 펩타이드 또는 단백질 여부를 확인해야 한 다. 표 4에 나열된 마커는 혈청 단백질이거나 인간 혈청 단백질 단편일 수 있으며 이를 분석하기 위해 LC-MS/MS 와 같은 더 민감한 기기나 추가 분석이 없다면 표 4의 마커의 정체를 결정하기 어려울 것이다.

19

Mass (m/z)	Possible Proteins	UniProt Access ID/Reference
3260	Epidermal Growth Factor receptor tyrosine kinase domain	A0A1Y0B9G4
	Fibrinogen α-chain fragment	P02671 (FIBA_HUMAN)
4207	Vitamin D binding protein fragment	A0A1B1CYC8

Table 6. Human serum protein biomarkers (Table 5) and their possible protein identities based on the cancer biomarker.

UniProtKB 데이터베이스에서 인간 혈청 단백질을 검색하여 분자량 및 가능한 단백질 이름 목록을 얻고 이 목록의 무게를 표 4의 6개 마커와 비교하였다. 모두 확인된 것은 아니지만 m/z 3,260 및 4,207은 각각 표피표피 성장 인자 수용체 Epidermal growth factor receptor (EGFR) tyrosine kinase 도메인 (UniProtKB 항목 ID A0A1Y0B9G4), 피브리노 젠 알파 체인 단편 (Fibrinogen α-chain fragment) (UniProtKB 항목 ID P02671) 및 비타민 D 결합 단백질 단편 (UniProtKB 항목 ID A0A1B1CYC8)일 수 있다는 것을 확인할 수 있 다(표 5). MALDI-TOF MS/MS 또는 LC/MS/MS와 같은 기술이 이 연구에서 수행되지 않아 서 표 4에 있는 마커의 실제 ID를 설정할 수 없었기 때문에 정확한 확인은 어렵다. 그럼 에도 불구하고, 다른 연구에서 EGFR tyrosine kinase 가 유방암과 관련이 있다는 사실에 주목하는 것은 흥미로운 일이다[17]. 이에 따르면, EGFR tyrosine kinases는 HER1의 촉매 하는 도메인에 위치하고 있으며 EGFR 패밀리를 구성하고 있는 것들은 높은 비율로 유방 암에서 과발현 되었으며 이들 분자는 원형질막과 연관되어 있고 유사분열 형성에 관여하 기 때문에 유방암의 발달과 진행을 촉진하기 위해 서로 협력하여 기능하는 것으로 예상 되어진다. G. L. Hortin의 연구에 따르면 m/z 3,260은 인간 혈청 및 혈장에서 MALDI-TOF 또는 Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (SELDI-TOF MS)를 사용하여 유방암의 바이오 마커로 확인된 피브리노젠 알파 체인 단편 (Fibrinogen α-chain fragment) 일 수 있다[15]. 또한, 또 다른 연구에서는 비타민 D 결합 단백질 Vitamin D Binding Protein (DBP)이 비타민 D 대사의 중요한 매개체이며 총 25(OH)D 레벨과 관련하여 칼슘 흡수와 부갑상선 호르몬 레벨과 연관되어 있다. 더 높은 골밀도, 더 높은 칼슘 레벨, 더 높은 부갑상선 호르몬 레벨과 관련하여 DBP 가 낮게 발 현될 경우 유방암 위험률이 높아 지는 것을 확인 할 수 있었으며 DBP가 유방암 병인에 역할을 할 수 있다고 밝혔다 [18-20].

추가로 확인된 6개의 m/z 값을 기준으로 가장 가까운 혈청내 물질을 기존의 암과 관

련된 바이오 마커들 중에서 찾아 나열해 보았다(표 7). [21-22]

Mass (m/z)	Possible Proteins	
2161	apolipoprotein C-3	
4153	complement C5	
4600	complement C4-B	
7618	alpha 2 macroglobulin	
8503	tropomyosin alpha -4 chain	
9575	-	

Table 7. Human serum protein biomarkers (Table 5) and their possible protein identities based on the cancer biomarker.

더 많은 임상 샘플을 가지고 추가적으로 시행한 연구들을 통해 그림 2 및 표 4,5에 서 확인된 마커들의 확인이 필요하다. 이 연구에 사용된 샘플과 향후 연구에 사용된 샘 플 간에 겹치는 m/z 값이 많이 있는 경우 이러한 m/z 값은 더 뛰어난 유방암 바이오 마 커가 될 것이다. 또한 샘플 준비 단계의 변동을 최소화하기 위한 새롭고 개선된 방법이 요구된다. 특히 데이터 변동이 적은 것이 보다 정확하고 신뢰할 수 있는 암 진단 도구를 얻기 위해 중요하기 때문이다. 실제로 변화에 취약할 수 있는 단계 중 하나는 Peptide Cleanup C18 pipet tip을 사용할 때의 혈청 단백질 정제 단계이다. 왜냐하면 이 과정에는 샘플과 다른 용액을 반복적으로 pipetting하는 많은 과정이 포함되기 때문인데 이러한 과정에서 흡인과 분주 속도가 샘플마다 다를 수 있기 때문이다. 따라서 이러한 프로세스 를 균일화 할 수 있는 더 많은 연구가 향후에 도움이 될 수 있으며, 이는 데이터를 보다 균일하고 정확하게 만들 것이다. 또한 자동화된 샘플 준비 시스템을 사용하면 시스템이 인적 오류 및 변형을 제거하므로 보다 정확한 데이터 수집에 더 유리할 것이다. MALDI-TOF MS를 통해 얻은 질량 정로 얻어 Heatmap을 통해 건강한 대조군 30명과 유방암 환자 30명의 통계적으로 유의한 m/z 피크의 패턴 차이를 보여줄 수 있었고 m/z 마커에 대한 건강한 대조군과 유방암 환자 사이의 정규화 된 강도 값은 통계적으로 유의한 차이를 보여주었다. 본 연구의 진단 능력을 분석하기 위한 ROC 곡선을 그렸으며 0.932의 AUC 값으로 검증되었다. TIC 정규화 강도 값에서 유의하게 차이를 보이는 6개의 질량 값을 확인하였다. 더 많은 임상 샘플을 통해 샘플 간에 더 많이 겹치는 m/z 값을 확인하고 이 프로세스를 좀더 자동화하고 균일화 된 시스템을 사용하면 인적 오류 및 변형을 제거하므로 보다 정확한 데이터 수집에 더 유리할 것이다. 따라서 향후 유방암의 진단에 유용할 것으로 생각된다.

참 고 문 헌

[1] Breast Cancer, World Health Organization. Available online:

https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[2] A. A. Abdelati, R. A. Elnemr, N. S. Kandil, F. I. Dwedar, and R. A. Ghazala, "Serum Peptidomic Profile as a Novel Biomarker for Rheumatoid Arthritis", *Int J Rheumatol.*, vol. 2020, 10 pages, 2020.

[3] M. Pietrowka and P. Widlak, "MALDI-MS-Based Profiling of Serum Proteome: Detection of Changes Related to Progression of Cancer and Response to Anticancer Treatment", *Int J Proteomics*, vol. 2012, 10 pages, 2012.

[4] T. Cabezon, I. Gromova, P. Gromov, R. Serizawa, V. T. Wielenga, N. Kroman, J. E. Celis, and J. M. A. Moreira, "Proteomic Profiling of Triple-negative Breast Carcinomas in Combination With a Three-tier Orthogonal Technology Approach Identifies Mage-A4 as Potential Therapeutic Target in Estrogen Receptor Negative Breast Cancer", *Mol Cell Proteomics*, 12(2): 381-394, 2013.

[5] J. S. Biscardi, R. C. Ishizawar, C. M. Silva, and S. J. Parsons, "Tyrosine kinase signaling in breast cancer Epidermal growth factor receptor and c-Src interactions in breast cancer", *Breast Cancer Res*, 2: 203–210, 2000.

[6] E. Tagliabue, S. Raimondi, and S. Gandini, "Meta-analysis of Vitamin D-Binding Protein and Cancer Risk", *Cancer Epidemiol Biomarkers Prev*, 24(11): 1758-1765, 2015.

[7] D. E. Misek and E. H. Kim, "Protein Biomarkers for the Early Detection of

Breast Cancer", International Journal of Proteomics, vol 2011, 9 pages, 2011.

[8] C. Mueller, A. Haymond, J. B. Davis, A. Williams, and V. Espina, "Protein biomarkers for subtyping breast cancer and implications for future research", Expert Rev Proteomics, 15(2): 131–152, 2018.

[9] L. Chung, K. Moore, L. Phillips, F. M. Boyle, D. J. Marsh, and R. C. Baxter, "Novel serum protein biomarker panel revealed by mass spectrometry and its prognostic value in breast cancer", Breast Cancer Res., 16(3): R63, 2014.

[10] S. Miah, C. A. S. Banks, M. K. Adams, L. Florens, K. E. Lukong, and M. P. Washburn, "Advancement of mass spectrometry-based proteomics technologies to explore triple negative breast cancer", Mol Biosyst., 13(1): 42–55, 2016.

[11] T. Cabezon, I. Gromova, P. Gromov, R. Serizawa, V. T. Wielenga, N. Kroman, J. E. Celis, and J. M. A. Moreira, "Proteomic Profiling of Triple-negative Breast Carcinomas in Combination With a Three-tier Orthogonal Technology Approach Identifies Mage-A4 as Potential Therapeutic Target in Estrogen Receptor Negative Breast Cancer", Mol Cell Proteomics, 12(2): 381-394, 2013.

[12] D. Ding, M. Chen, X. Xiao, P. Cao, and S. Li, "Novel serum peptide model revealed by MALDI-TOF-MS and its diagnostic value in early bladder cancer", *In J Biol Markers*, 8 pages, 2020.

[13] D. Song, L. Yue, H. Li et al., "Diagnostic and prognostic role of serum protein peak at 6449 m/z in gastric adenocarcinoma based on mass spectrometry", Br J Cancer, 114(8): 929-938, 2016. [14] H. Park, K. Jang, H. Park, W. Song, Y. Jeong, D. Ahn, S. Kim, Y. Yang, and Y. Kim, "MALDI-TOF MS-based total serum protein fingerprinting for liver cancer diagnosis", *Analyst*, 144(7): 2231-2238, 2019.

 [15] G. L. Hortin, "The MALDI-TOF Mass Spectrometric View of the Plasma Proteome and Peptidome", *Clin Chem*, 52(7): 1223–1237, 2006.

[16] Y. Jin and T. Manabe, "Direct targeting of human plasma for matrix-assisted laser desorption/ionization and analysis of plasma proteins by time of flight-mass spectrometry", *Electrophoresis*, 26(14): 2823-2834, 2005.

[17] J. S. Biscardi, R. C. Ishizawar, C. M. Silva, and S. J. Parsons, "Tyrosine kinase signaling in breast cancer Epidermal growth factor receptor and c-Src interactions in breast cancer", *Breast Cancer Res*, 2: 203–210, 2000.

[18] E. Tagliabue, S. Raimondi, and S. Gandini, "Meta-analysis of Vitamin D-Binding Protein and Cancer Risk", *Cancer Epidemiol Biomarkers Prev*, 24(11): 1758-1765, 2015.

[19] D. E. Misek and E. H. Kim, "Protein Biomarkers for the Early Detection of Breast Cancer", *International Journal of Proteomics*, vol 2011, 9 pages, 2011.

[20] T. M Pawlik, D. H. Hawke, "Proteomic analysis of nipple aspirate fluid from women with early-stage breast cancer using isotope-coded affinity tags and tandem mass spectrometry reveals differential expression of vitamin D binding protein ", *BMC Cancer*, 6:68, 2006 [21] A. Tessitore, A. Gaggiano, "Serum Biomarkers Identification by Mass Spectrometry in High-Mortality Tumors" *Int J Proteomics.* 2013; 2013: 125858.

[22] G. Arapidi, M. Osetrova, "Peptidomics dataset: Blood plasma and serum samples of healthy donors fractionated on a set of chromatography sorbents" *Data Brief.* 2018 Jun; 18: 1204–1211.

Abstract

SOOJEONG CHOI

Department of surgery, Seoul Medical center Ulsan University college of Medicine, Republic of Korea

Purpose

The main objective of this study was to develop novel methods for the screening of breast cancer by analyzing various human blood serum proteins and peptides by MALDI-TOF MS system and discover any meaningful protein biomarkers.

Subjects and methods

Thirty healthy controls and 30 breast cancer serum samples were obtained from Asan Medical Center (Korea). The serum samples were first diluted, after which the proteins were purified and concentrated by using Peptide Cleanup C18 pipette tips (Agilent Technologies). The proteins were then analyzed by MALDI-TOF MS in linear positive mode, and the acquired data were statistically analyzed by heatmap, PCA plotting, and ROC curve analysis methods. An In - house diagnostic software was also used for diagnosis.

Results

From the analyses, six protein biomarkers were discovered, which would become the basis for future breast cancer diagnosis. Further, more studies are needed to validate such biomarkers, and more clinical samples are needed to further verify their diagnostic performance.

Conclusion

The more m/z values overlap between the samples used in future studies, the more promising breast cancer biomarkers will be. Furthermore, a novel or improved method to

minimize any variations in sample preparation steps is strongly encouraged, especially because low data variations are crucial for obtaining more accurate and reliable cancer screening.

Keywords

Breast neoplasms, tumor marker, heatmap, MALDI-TOF MS, peptide