(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0142237
(43) 공개일자 2021년11월25일

(51) 국제특허분류(Int. Cl.)
G01N 33/574 (2006.01) *C12Q 1/6886* (2018.01)
(52) CPC특허분류
G01N 33/57415 (2019.01)
C12Q 1/6886 (2018.05)
(21) 출원번호 10-2020-0058809
(22) 출원일자 2020년05월18일
심사청구일자 없음

(71) 출원인
씨비에스바이오사이언스 주식회사
대전광역시 유성구 테크노11로 24,3층 (탑립동)
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
국립암센터
경기도 고양시 일산동구 일산로 323 (마두동)
(72) 발명자
이은숙
경기도 과천시 별양로 111, 506동 1404호
공선영
경기도 고양시 일산동구 강송로 195, 815동 1005호
(뒷면에 계속)
(74) 대리인
오국진

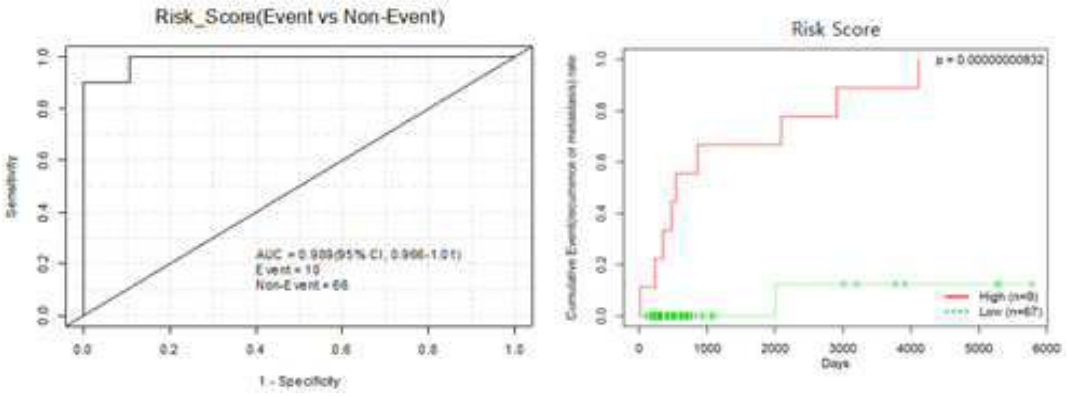
전체 청구항 수 : 총 2 항

(54) 발명의 명칭 **삼중음성유방암 환자에서 재발 및 전이 예후를 예측하기 위한 분석방법**

(57) 요 약

본 발명은 삼중음성유방암 환자에서 재발 및 전이의 예후를 나타내는 환자의 진단에 필요한 정보를 제공하기 위한 분석방법을 제공한다. 삼중음성유방암 환자의 종양조직에서 6개의 유전자 즉, DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 발현량을 조합하여 분석할 경우 높은 정확도로 재발 및 전이의 예후를 나타내는 삼중음성유방암 환자를 진단할 수 있다는 것이 본 발명에 의해 밝혀졌다. 따라서, 상기 유전자들의 조합은 삼중음성유방암 환자에서 재발 및 전이의 예후를 나타내는 환자를 선별할 수 있는 바이오마커로서 유용하게 사용될 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G01N 33/57484 (2013.01)
C12Q 2600/156 (2013.01)
G01N 2333/82 (2013.01)

(72) 발명자
정소연
경기도 고양시 일산서구 주화로 211, 104동 905호
이은경
경기도 고양시 일산서구 대화1로 70, 708동 803호
김민경
경기도 부천시 삼작로 349, 가동 104호
박경화
서울특별시 성북구 보국문로 116, 310동 1004호
박진영
서울특별시 마포구 독막로42길 2, 103동 502호
유윤석
대전광역시 서구 복수서로 15, 108동 2102호
서용배
부산광역시 남구 용호로269번길 29, 101동 104호
유원백
대전광역시 유성구 노은로 353, 301동 2005호
김창민
서울특별시 강서구 양천로62길 20

이 발명을 지원한 국가연구개발사업
과제고유번호 1465028053
과제번호 NCC1710172
부처명 보건복지부
과제관리(전문)기관명 국립암센터
연구사업명 기관고유연구사업
연구과제명 유전성 유암방/난소암의 고침투율 및 중간침투율 유전자 돌연변이 빈도 확인 및 원인 유전자에 따른 임상특성 파악
기 여 율 1/1
과제수행기관명 국립암센터
연구기간 2017.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

삼중음성유방암 환자에서 재발 및 전이의 예후를 나타내는 환자의 진단에 필요한 정보를 제공하기 위하여, 삼중음성유방암 환자로부터 체외로 분리된 종양조직 샘플에서 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 발현량을 각각 측정하는 단계를 포함하는 분석방법.

청구항 2

제1항에 있어서, 상기 측정이 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 mRNA 발현량을 각각 측정함으로써 수행되는 것을 특징으로 하는 분석방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 삼중음성유방암 환자에서 재발 및 전이 예후를 예측하기 위한 분석방법에 관한 것이다. 더욱 상세하게는, 삼중음성유방암 환자에서 재발 및 전이의 예후를 나타내는 환자의 진단에 필요한 정보를 제공하기 위한 분석방법에 관한 것이다.

배 경 기 술

[0002] 세계보건기구(WHO) 산하 국제암연구기구(IARC)가 2019년 3월에 발표한 결과에 의하면, 유방암은 세계 2위로 신규 발병이 높아 환자가 2백만명을 넘으며 62만명이 사망해 사망률 4위에 이르는 암 종이다. 삼중음성유방암(triple negative breast cancer)은 에스트로겐 수용체, 프로게스테론 수용체와 HER2의 발현이 모두 음성인 유방암의 아형이며 전체 유방암 환자의 10~15%를 차지한다. 재발 위험도가 높고 예후가 불량한 암으로, 전이가 흔하게 발생하여 폐, 간 및 뇌로의 전이가 주를 이룬다. 또한 변이(Driver mutation) 유전자가 연구된바 없어 삼중음성유방암 환자에 대한 표준 치료는 화학치료요법에 제한되어 있다.

[0003] 유전체 분석기술의 발전은 암의 개시(initiation)와 진행(progression)과 같은 암의 유전적 특성을 파악하는데 있어 큰 성과를 이루고 있다. 일부 연구결과를 보면 24%의 삼중음성유방암 환자에서 EGFR 유전자의 과발현이 확인되었고 화학치료요법 비반응과 연관성이 있으며 낮은 생존율이 나타났음이 확인되었다. 또한 최근의 연구에서 EGFR 발현은 단변량분석과 다변량분석에서 DFS(무병생존)에 대한 예측력을 보여주었다. 삼중음성유방암에서 VEGF 유전자의 과발현은 짧은 DFS, OS가 나타나고 있다. 아울러 VEGF 발현량은 종양의 크기, 병기 및 전이 부위와도 연관성이 있다는 연구가 보고되었다. 또한 CK-5, CK-14 및 CK-17등 유전자의 발현은 종양의 높은 병기, ER 음성, 짧은 DFS와 OS 등과 연관성이 높아 이런 유전자들의 발현은 예후 예측하는 바이오마커로 사용될 수 있다는 연구도 있다. 이 외에도 P53을 비롯한 다른 유전자들이 예후 예측 바이오마커로 제시가 되었지만 단일 유전자 관련 연구는 IHC 또는 PCR 등 기법을 통해 진행된 연구가 주를 이루고 있다.

[0004] 상기 유전자들은 이미 잘 알려진 유방암 혹은 삼중음성유방암과 관련된 유전자들의 일부에 불과하므로, 이들 유전자를 이용하여 유방암의 재발 및 전이 예후 진단을 하기는 매우 어려운 실정이다. 따라서 RNA 서열분석(RNA sequencing)을 통하여 정확도와 민감도가 높은 유방암 재발 및 전이 마커를 밝혀내는 것이 당업계에 요구되고 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명자들은 삼중음성유방암 환자의 재발 및 전이에 대한 예후를 예측할 수 있는, 임상적으로 적용 가능한 바이오머커("유전자 시그니쳐(gene signature)"로도 지칭됨)를 개발하기 위하여 다양한 연구를 수행하였다. 특히, 본 발명자들은 삼중음성유방암 환자의 종양조직 샘플을 사용하여 RNA를 추출하여 유전자 발현 분석을 수행하였다. 그 결과, 환자의 종양조직에서 특정 유전자들, 즉 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의

발현량 차이를 조합하여 분석할 경우 높은 정확도로 재발 및 전이에 대한 예후를 예측할 수 있다는 것을 발견하였다.

[0006] 따라서, 본 발명은 바이오마커로서 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자를 사용하는 것을 포함하는, 삼중음성유방암 환자에서 재발 및 전이 예후를 예측하기 위한 분석방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0007] 본 발명의 일 태양에 따라, 삼중음성유방암 환자에서 재발 및 전이의 예후를 나타내는 환자의 진단에 필요한 정보를 제공하기 위하여, 삼중음성유방암 환자로부터 체외로 분리된 종양조직 샘플에서 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 발현량을 각각 측정하는 단계를 포함하는 분석방법이 제공된다.

[0008] 본 발명의 분석방법에 있어서, 상기 유전자의 발현량측정은 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 mRNA 발현량을 각각 측정함으로써 수행될 수 있다.

발명의 효과

[0009] 삼중음성유방암 환자의 종양조직에서 특정 유전자들, 즉 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 발현량 차이를 조합하여 분석할 경우 높은 정확도로 재발 및 전이의 예후를 예측할 수 있다는 것이 본 발명에 의해 밝혀졌다. 따라서, 상기 유전자들의 조합은 삼중음성유방암 환자의 재발 및 전이의 예후를 확인할 수 있는 바이오마커로서 유용하게 사용될 수 있다.

도면의 간단한 설명

[0010] 도 1은 삼중음성유방암 환자의 재발 및 전이 판별을 하기 위한 예측 기준값을 얻기 위한 ROC 분석 및 예측 기준값을 활용한 생존분석(Kaplan-Meier)의 로그 순위 검정(log-rank test) 분석 결과를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0011] 삼중음성유방암은 재발 위험도가 높고 예후가 불량한 암으로, 전이가 흔하게 발생하여 폐, 간 및 뇌로의 전이가 주를 이룬다. 본 발명자들은 삼중음성유방암 환자 중 재발 및 전이 환자군(n=10), 재발 또는 전이가 없는 환자군(n=66)의 종양조직으로부터 다양한 유전자의 mRNA 발현량을 측정하였다. 또한 공지 데이터베이스인 GEO(Gene Expression Omnibus)에서 대조 데이터(삼중음성유방암 주변의 비종양조직, n=21)를 확보하여 분석을 수행하였다. 콕스 회귀분석을 사용하여 재발 및 전이 환자와 재발 및 전이가 없는 환자간에 상이한 발현을 나타내는 6개의 유전자(DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785)를 동정하였다(정확도(accuracy) = 98.68%, 특이도(specificity) = 100%, 민감도(sensitivity) = 90%). 상기 유전자 시그니쳐는 모든 병리학적 및 임상적 특징에 대하여 독립적이었다. 따라서, 상기 6개의 유전자 시그니쳐는 삼중음성유방암 환자의 재발 및 전이의 예후를 예측할 수 있는 바이오마커로서 사용될 수 있다.

[0012] 따라서, 본 발명은 삼중음성유방암 환자에서 재발 및 전이의 예후를 나타내는 환자의 진단에 필요한 정보를 제공하기 위하여, 삼중음성유방암 환자로부터 체외로 분리된 종양조직 샘플에서 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 발현량을 각각 측정하는 단계를 포함하는 분석방법을 제공한다.

[0013] 본 발명의 분석방법에 있어서, 상기 유전자의 발현량측정은 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 mRNA 발현량을 각각 측정함으로써 수행될 수 있다. 예를 들어, 상기 유전자의 발현량은 Illumina HiSeq 4000 sequencer(Illumina, Inc., San Diego, CA, USA)를 사용하여 얻어진 표준화된 발현량(normalized expressioin level)일 수 있으며, 상기 표준화(normalization)는 해당 Illumina HiSeq 4000 sequencer에서 제공되는 제조사 권장사항에 따라 수행된다.

[0014] 본 발명의 분석방법에서 바이오마커로 사용되는 상기 6개의 유전자 즉, DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자는 공지의 유전자로서, 그 서열은 진뱅크(GenBank) 등에 공지되어 있다. 예를 들어, DEPDC7(DEP domain-containing protein 7) 단백질의 NCBI 억세션 번호(NCBI accession number)는 NP_001070710, NP_631899 등이며, 이를 코딩하는 유전자 mRNA의 NCBI 억세션 번호는 NM_001077242, NM_139160 등이다. DIP2B(disco interacting protein 2 homolog B) 단백질의 NCBI 억세션 번호는 NP_775873 이며, 이를 코딩하는 유전자 mRNA의 NCBI 억세션 번호는 NM_173602 이다. LYST(lysosomal trafficking regulator) 단백질의 NCBI 억세션 번호는 XP_011542333, XP_011542334, XP_011542335, XP_011542336, XP_011542337, XP_011542338,

XP_011542339, XP_011542341, XP_016855639, XP_016855640 이며, 이를 코딩하는 유전자 mRNA의 NCBI 억세션 번호는 XM_011544031, XM_011544032, XM_011544033, XM_011544034, XM_011544035, XM_011544036, XM_011544037, XM_011544039, XM_017000150, XM_017000151 이다. SDS(serine dehydratase) 단백질의 NCBI 억세션 번호는 NP_006834 등이며, 이를 코딩하는 유전자 mRNA의 NCBI 억세션 번호는 NM_006843 등이다. SLC43A1(solute carrier family 43 member 1) 단백질의 NCBI 억세션 번호는 NP_001185739, NP_003618 등이며, 이를 코딩하는 유전자 mRNA의 NCBI 억세션 번호는 NM_001198810, NM_003627 등이다. ZNF785(zinc finger protein 785) 단백질의 NCBI 억세션 번호는 NP_689671 등이며, 이를 코딩하는 유전자 mRNA의 NCBI 억세션 번호는 NM_152458 등이다.

[0015] 본 발명에 따른 분석방법의 일 구현예에서, 삼중음성유방암 환자로부터 체외로 분리된 종양조직 샘플 중 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 유전자의 발현량을 측정하고, 하기 식에 따라 측정된 위험도 점수가 4.977464 보다 클 경우에는 해당 삼중음성유방암 환자는 재발 및 전이가 나타날 위험도가 높은 환자로 분류할 수 있으며, 4.977464 이하일 경우 재발 및 전이 위험도가 낮은 환자로 분류할 수 있다.

[0016] 위험도 점수 = $(-1.158489)*G_{DEPDC7} + (0.119946)*G_{DIP2B} + (0.242934)*G_{LYST} + (0.111303)*G_{SDS} + (-0.462352)*G_{SLC43A1} + (1.618724)*G_{ZNF785}$

[0017] 상기 식에서 G_{DEPDC7}, G_{DIP2B}, G_{LYST}, G_{SDS}, $G_{SLC43A1}$, G_{ZNF785} 는 각각 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785의 유전자 발현량을 나타낸다. 상기 각각의 유전자 발현량은 유전자 발현량 측정 기기인 Illumina HiSeq 4000 sequencer(Illumina, Inc., San Diego, CA, USA)를 사용하여 얻어진 표준화된 발현량(normalized expressioin level)을 나타낸다. 상기 표준화(normalization)는 해당 Illumina HiSeq 4000 sequencer에서 제공되는 제조사 권장사항에 따라 수행된다.

[0018] 이하, 본 발명을 실시예를 통하여 더욱 상세히 설명한다. 그러나 이들 실시예는 본 발명을 예시하기 위한 것으로, 본 발명의 범위가 이들 실시예에 한정되는 것은 아니다.

[0019] **1. 시험방법**

[0020] (1) RNA 추출

[0021] Allprep DNA/RNA Mini 키트(Qiagen, Hilden, Germany) 및 Trizol 시약(Invitrogen)을 사용하여 제조사 지침에 따라 액체질소에 보관되어 있는 TNBC 환자의 종양조직(n=76)으로부터 총 RNA를 추출하였다. 시험 참여 전에 환자로부터 서면 동의를 받았으며, 연구 프로토콜(NCCNCS13717, NCC2017-0127)은 국립암센터(National Cancer Center)의 임상시험심사위원회(institutional review board)의 승인을 받았다.

[0022] (2) 환자 및 반응 평가

[0023] TNBC 환자의 진단시 수술을 시행한 환자는 수술시 검체 또는 생검 검체를 보관하였고, 수술전 항암치료를 시행한 TNBC 환자 중 수술 시 잔여 암조직이 있는 경우에 검체를 수집하였다. TNBC 진단은 ASCO 가이드라인에 따라 기관에서 면역화학염색과 필요시 HER-2 SISH(Silver in situ hybridization) 검사를 통해 진단하였다.

[0024] (3) 유전자 발현 분석

[0025] 유전자 발현량은 마크로젠(서울, 한국)에 의뢰하여, Illumina HiSeq 4000 sequencer(Illumina, Inc., San Diego, CA, USA)를 사용하여 제조사 지침에 따라 측정하였다. 모든 샘플의 RNA의 RNA integrity number (RIN) 값이 8을 초과하였음을 확인 후, TruSeq Standed mRNA LT Sample Prep Kit (Illumina)을 이용하여 cDNA 라이브러리를 제작한 다음(Library protocol: TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E), Illumina HiSeq 4000을 사용하여 페어드-엔드 시퀀싱(paired-end sequencing)을 수행하였다. 결과로써 얻어진 FPKM(Fragments Per Kilobase of transcript per Million mapped reads)에 대하여 Illumina HiSeq 4000 sequencer에서 제공되는 제조사 권장사항에 따라 표준화(normalization) 작업이 수행된 최종 FPKM값을 마크로젠으로부터 전달 받아 유전자 발현 분석에 사용하였다.

[0026] (4) 공지 데이터(Public Data) 확보

[0027] 본 연구에서 삼중음성유방암 환자의 종양 조직에서 얻어진 데이터와의 비교를 위하여, 비종양 조직의 데이터로서 공지 데이터베이스인 GEO(Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/)를 활용하였다. GEO 데이터베이스에서 삼중음성유방암과 관련된 연구들 중 비종양 데이터 수가 많은 연구 데이터를 사용하였고(GEO accession id : GSE58135), GSE58135 데이터 중 21명의 비종양 RNA 시퀀싱 데이터를 이용하여 분석을 진

행하였다. 비종양 데이터는 Status가 Fail이거나 값이 $1.0x10^{-6}$ 미만인 데이터를 제외하였고, 종양 조직에서 생산한 데이터와 매칭하여 종양, 비종양에서 모두 값이 존재하는 10,856개의 유전자를 선정하였으며, 이후의 통계적 조합 유전자 분석에 사용하였다.

[0028] (5) 통계적 조합 유전자 분석(Statistical Combination Gene Analysis)

[0029] 본 연구에서의 모든 통계분석은 오픈 소스 통계 프로그램 환경 R 언어(open source statistical programming environment R language)(Version 3.4.3)를 사용하여 수행하였다. 윌콕슨 순위합 검정(Wilcoxon rank sum test)을 통하여 상이하게 발현되는 유전자(differentially expressed genes, DEGs)를 과발현 혹은 저발현된 것으로 분류하여(p <0.05 및 |폴드-변화(fold-change)| >1.5), 재발 및 전이 환자와 재발 또는 전이가 없는 환자와 비교하였다. 선발된 DEGs의 수를 조합하여 분석하였고, 유전자 조합(gene combinations)의 총 수는 다음 식을 사용하여 계산하였다:

[0030] $$\sum_{k=1}^{n} \frac{n!}{k!\,(n-k)!}$$

[0031] 상기 식에서, n은 선발된 DEGs 총 수이고, k는 조합하여 포함시킨 유전자의 수이다.

[0032] 콕스비례위험(Cox proportional hazard 또는 Cox regression) 분석을 수행하여 유전자 시그니처(gene signatures) 및 임상병리학적 특징의 관련성을 측정하였다(p<0.05).

[0033] 후보 유전자 시그니처(p < 0.05; AUC > 0.800; 민감도(sensitivity) > 80%; 및 특이도(specificity) > 80%)를 2-fold 교차 검증에 의해 계층화하여 최적의 유전자 조합(gene combination)을 동정하였다. 76명 종양 환자 군을 2개의 폴드(트레이닝 세트 및 시험 세트)로 나누어 트레이닝 세트에서의 기준 값을 테스트 세트에 적용하여 결과를 확인하였으며, 트레이닝 세트와 시험 세트의 환자군은 무작위로 나누었고 이를 300회 반복하여 테스트하였다. 정확도는 시험 세트에 대하여 p <0.05를 근거하여 계산하였다.

[0035] **2. 시험결과**

[0036] (1) 재발 및 전이 환자와 재발 또는 전이가 없는 환자 간의 상이한 유전자 발현의 분석

[0037] 삼중음성유방암 환자 76명의 종양 데이터와 GEO에서 확보한 21명의 비종양 데이터 간의 상이한 유전자 발현의 분석 결과, 9856/10856 유전자들이 통계적으로 유의하게 상이하게 발현되었다(p <0.05 및 |폴드-변화(fold-change)| >1.5). 이들 9856 유전자 중 재발 및 전이가 있는 환자(n=10)와 재발 또는 전이가 없는 환자(n=66)의 데이터 간의 상이한 유전자 발현의 분석 결과, 197개 유전자를 선별하였다(p <0.05 및 |폴드-변화(fold-change)| >1.5). 197개 유전자 중 콕스비례위험 분석을 통하여 유의하게 상이하게 발현된 52개 유전자를 선별하였다(p < 0.05)(표 1).

표 1

[0038] 재발 및 전이 환자와 재발 또는 전이가 없는 환자 간 상이하게 발현되는 유전자

Sr. No.	유전자	p-값
1	ANKRD23	8.48E-04
2	ANKRD36	6.75E-04
3	ANKRD36BP2	5.63E-04
4	C12orf65	2.89E-02
5	C2orf49	1.28E-02
6	CCDC107	4.54E-02
7	CCDC114	2.53E-03
8	CGREF1	1.66E-02
9	DEPDC7	1.97E-02
10	DGKH	3.93E-03
11	DIP2B	7.53E-03
12	EBLN2	2.79E-03
13	ELF3	4.41E-02
14	EMP1	1.02E-03

15	GADD45B	2.35E-02
16	GLS	4.59E-02
17	GRHL1	2.69E-02
18	ICAM1	4.74E-02
19	IMMP1L	3.92E-02
20	KCNC3	4.21E-02
21	KIAA1468	9.04E-03
22	KRT17	4.85E-02
23	LRIT3	4.93E-04
24	LSM11	7.98E-04
25	LYST	2.14E-04
26	MCM3AP.AS1	4.11E-04
27	MT2A	1.55E-03
28	MUM1	1.41E-02
29	NEURL2	1.85E-02
30	NOXA1	2.20E-04
31	OTUD3	2.35E-03
32	PGBD4	6.63E-05
33	RASGRP1	1.74E-04
34	RORA	4.56E-03
35	SDS	1.27E-03
36	SERINC5	5.13E-03
37	SLC22A20	6.17E-03
38	SLC27A5	4.32E-02
39	SLC43A1	4.27E-02
40	SLC45A4	1.47E-03
41	STAG3L3	1.95E-03
42	TMCC2	1.60E-03
43	TTLL4	1.53E-02
44	TYW5	9.10E-03
45	ZBED6	4.10E-02
46	ZNF217	2.77E-03
47	ZNF248	5.00E-04
48	ZNF320	2.65E-03
49	ZNF678	1.54E-02
50	ZNF70	3.81E-03
51	ZNF785	1.78E-05
52	ZNF808	2.91E-03

[0040] (2) 유전자 시그니쳐의 선별 및 콕스회귀 분석

[0041] 52개의 상이하게 발현되는 유전자들을 조합하여 콕스분석, ROC(Receiver Operating Characteristics) 분석, 2-폴드 교차검증을 수행하였다. 콕스분석은 재발 및 전이의 예후에 대한 조합 분석(위험점수)의 가중치를 계산하기 위하여 수행하였다. ROC 분석은 민감도[High Risk Event 환자 (기준값 위의 재발 및 전이환자) / 전체 Event 환자 (재발 및 전이가 있는 환자)]와 특이도 [Low Risk Non-Event (환자 재발 및 전이가 없는 환자) / 전체 Non-Event 환자 (재발 및 전이가 없는 환자)]로 각각 계산하였다. 2-폴드 교차검증은 상기 검증분석에서 예측모델의 성능을 평가하기 위하여 사용하였으며, 이는 검증에 사용되는 전체 데이터를 학습(training) 데이터와 검증(validation) 데이터로 나누어, 학습 데이터에 대한 최적의 예측모델을 만들고, 검증 데이터로 일반적 성능을 평가하는 방법이다. KM(Kaplan-Meier) 그래프, 로그 순위 검정(log-rank test), Cox 통계분석은 오픈 소스 통계 프로그램 환경 R 언어(open source statistical programming environment R language)(Version 3.4.3)에서 제공하는 survival Package를 사용하였다. ROC 통계분석은 R 언어(Version 3.4.3)에서 제공하는 verification Package를 사용하였다.

[0042] 상기 콕스분석 및 ROC분석을 통하여 유전자 시그니쳐 후보를 확인하였으며(표 2), 모든 후보 유전자 시그니쳐들의 2-폴드 교차검증 결과, 6개의 유전자 시그니쳐 즉, DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785의 조합이

최적의 유전자 조합(gene combination)임을 확인하였다. 상기 6개의 유전자 시그니쳐는 재발 및 전이 환자와 재발 및 전이가 없는 환자를 구분하는데 있어서 일관된 정확도(92.3%)를 나타내었다(트레이닝 코호트에서 민감도(sensitivity) = 90.0% 및 특이도(pecificity) = 100.0%)(표 2).

표 2

[0043] 유전자 시그니쳐 후보

Sr. No.	유전자 시그니쳐	콕스비례위험 회귀 p-값	교차검증 정확도 (%)	AUC*	민감도 (%)	특이도 (%)
1	DEPDC7_DIP2B_LYST_SDS_SLC43A1_ZNF785	3.94E-04	92.3	0.989	90.00	100.00
2	DEPDC7_IMMP1L_LYST_SLC43A1_SLC45A4	3.94E-04	92.0	0.986	90.00	100.00
3	GADD45B_GRHL1_IMMP1L_PGBD4_SLC43A1	4.39E-04	90.3	0.979	90.00	98.48
4	C2orf49_DIP2B_GADD45B_LYST_NEURL2_RORA_SLC22A20_SLC27A5_TYW5	2.74E-04	88.7	0.982	90.00	98.48
5	GADD45B_IMMP1L_PGBD4_SLC27A5_SLC43A1	3.94E-04	88.7	0.989	90.00	100.00
6	GADD45B_IMMP1L_PGBD4_SLC27A5_SLC43A1_TYW5	3.94E-04	88.7	0.989	90.00	100.00
7	GADD45B_IMMP1L_PGBD4_SLC22A20_SLC43A1	4.15E-04	88.3	0.980	90.00	98.48
8	DEPDC7_DIP2B_GADD45B_IMMP1L_LYST	4.15E-04	88.0	0.983	90.00	98.48
9	GADD45B_IMMP1L_LYST_OTUD3_ZNF320	4.39E-04	88.0	0.982	90.00	98.48
10	DIP2B_ICAM1_IMMP1L_PGBD4_RORA	4.39E-04	88.0	0.977	90.00	98.48

[0044] * AUC: 곡선하 면적(area under the curve)

[0046] 선별된 유전자 시그니쳐와 임상병리학적 특징과의 관련성을 검토하였다. 단변량 분석 결과, 선별된 유전자 시그니쳐가 재발 및 전이 예후와 유의하게 양성으로(significantly and positively) 관련되었다(표3).

표 3

[0047] 선별된 유전자 시그니쳐의 예측값의 단변량 로지스틱 회귀분석($p < 0.05$)

변수[1]	N[2]	Coef[3]	SE[4](Coef)	Z-스코어	p-값
후보 유전자					
DEPDC7_DIP2B_LYST_SDS_SLC43A1_ZNF785 (low vs. high)	76	3.764752	1.062185	3.544347	3.94E-04
임상병리학적 특징					
AGE (<55 years vs ≥55 years)	76	-0.618548	0.792647	-0.780357	4.35E-01
TNM Stage (012 vs 3)	76	0.780827	0.833248	0.937088	3.49E-01
TNM Stage (01 vs 23)	76	1.986070	1.082345	1.834968	6.65E-02

[0048] [1]약어: T; 종양, N; 절(node), M;전이

[0049] [2]N; 샘플의 수

[0050] [3]Coef; 계수(Coefficient)

[0051] [4]SE; 표준 편차(Standard error)

[0053] (3) 단변량 콕스 회귀분석을 통한 위험도 점수의 계산

[0054] 상기와 같은 방법으로 선별된 6개의 유전자, 즉 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785의 조합에 대하여 단변량 콕스 회귀분석(univariable cox regression)을 통하여 얻어진 각 유전자별 회귀 계수 값은 다음 표 4와 같다.

표 4

[0055]

유전자	회귀 계수 값
DEPDC7	-1.158489
DIP2B	0.119946
LYST	0.242934
SDS	0.111303
SLC43A1	-0.462352
ZNF785	1.618724

[0057] Illumina HiSeq 4000 sequencer(Illumina, Inc., San Diego, CA, USA)를 사용하여 얻어진 6개 유전자, 즉 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785 각각의 표준화된 발현량 및 각 유전자별 회귀 계수 값을 사용하여 다음 식에 따라, 위험도 점수를 계산하였다.

[0058] 위험도 점수 = $C_{DEPDC7}*G_{DEPDC7} + C_{DIP2B}*G_{DIP2B} + C_{LYST}*G_{LYST} + C_{SDS}*G_{SDS} + C_{SLC43A1}*G_{SLC43A1} + C_{ZNF785}*G_{ZNF785}$

[0059] 상기 식에서 $C_{유전자}$는 해당 유전자의 회귀 계수 값을 나타내며, $G_{유전자}$는 Illumina HiSeq 4000 sequencer(Illumina, Inc., San Diego, CA, USA)를 사용하여 얻어진 해당 유전자의 표준화된 발현량을 나타낸다. 따라서, 상기 표 4의 결과로부터, 위험도 점수는 하기 식에 따라 또한 계산될 수 있다.

[0060] 위험도 점수 = $(-1.158489)*G_{DEPDC7} + (0.119946)*G_{DIP2B} + (0.242934)*G_{LYST} + (0.111303)*G_{SDS} + (-0.462352)*G_{SLC43A1} + (1.618724)*G_{ZNF785}$

[0061] 상기와 같이 계산하여 얻어진 각각의 환자에 대한 위험도 점수는 다음 표 5 및 표 6과 같다.

표 5

[0062]

Sample ID	그룹	재발 또는 전이 여부 (1:발생) (0:해당없음)	조합 위험점수	민감도 (Sensitivity)	특이도 (specificity)	정확도 (accuracy)	Youden Index
Sample01	위험군	1	15.779260	0.100	1.000	0.882	0.10
Sample02		1	9.654481	0.200	1.000	0.895	0.20
Sample03		1	8.585630	0.300	1.000	0.908	0.30
Sample04		1	7.793085	0.400	1.000	0.921	0.40
Sample05		1	6.521226	0.500	1.000	0.934	0.50
Sample06		1	5.697706	0.600	1.000	0.947	0.60
Sample07		1	5.610815	0.700	1.000	0.961	0.70
Sample08		1	5.291414	0.800	1.000	0.974	0.80
Sample09		1	5.159479	0.900	1.000	0.987	0.90

Sample10	비	0	**4.977464**	0.900	0.985	0.974	0.88
Sample11	위	0	4.474678	0.900	0.970	0.961	0.87
Sample12	험	0	3.970314	0.900	0.955	0.947	0.85
Sample13	군	0	3.877342	0.900	0.939	0.934	0.84
Sample14		0	3.611965	0.900	0.924	0.921	0.82
Sample15		0	3.517072	0.900	0.909	0.908	0.81
Sample16		0	3.461817	0.900	0.894	0.895	0.79
Sample17		1	3.424566	1.000	0.894	0.908	0.89
Sample18		0	3.380364	1.000	0.879	0.895	0.88
Sample19		0	3.258880	1.000	0.864	0.882	0.86
Sample20		0	2.806072	1.000	0.848	0.868	0.85
Sample21		0	2.753257	1.000	0.833	0.855	0.83
Sample22		0	2.749770	1.000	0.818	0.842	0.82
Sample23		0	2.617917	1.000	0.803	0.829	0.80
Sample24		0	2.441097	1.000	0.788	0.816	0.79
Sample25		0	2.436208	1.000	0.773	0.803	0.77
Sample26		0	2.424576	1.000	0.758	0.789	0.76
Sample27		0	2.249013	1.000	0.742	0.776	0.74
Sample28		0	2.067505	1.000	0.727	0.763	0.73
Sample29		0	2.060774	1.000	0.712	0.750	0.71
Sample30		0	2.039521	1.000	0.697	0.737	0.70
Sample31		0	1.935868	1.000	0.682	0.724	0.68
Sample32		0	1.616827	1.000	0.667	0.711	0.67
Sample33		0	1.597566	1.000	0.652	0.697	0.65
Sample34		0	1.498501	1.000	0.636	0.684	0.64
Sample35		0	1.457272	1.000	0.621	0.671	0.62
Sample36		0	1.427002	1.000	0.606	0.658	0.61
Sample37		0	1.386498	1.000	0.591	0.645	0.59
Sample38		0	1.359689	1.000	0.576	0.632	0.58

표 6

[0064]

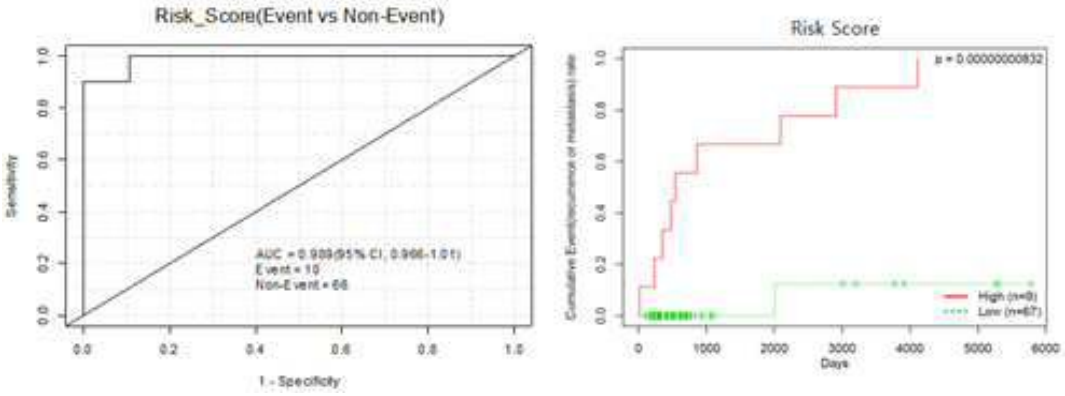
Sample ID	그룹	재발 또는 전이 여부 (1:발생) (0:해당없음)	조합 위험점수	민감도 (Sensitivity)	특이도 (specificity)	정확도 (accuracy)	Youden Index

Sample39	비	0	1.348544	1.000	0.561	0.618	0.56
Sample40	위	0	1.283606	1.000	0.545	0.605	0.55
Sample41	험	0	0.770510	1.000	0.530	0.592	0.53
Sample42	군	0	0.752583	1.000	0.515	0.579	0.52
Sample43		0	0.653110	1.000	0.500	0.566	0.50
Sample44		0	0.645426	1.000	0.485	0.553	0.48
Sample45		0	0.616810	1.000	0.470	0.539	0.47
Sample46		0	0.527976	1.000	0.455	0.526	0.45
Sample47		0	0.496256	1.000	0.439	0.513	0.44
Sample48		0	0.472278	1.000	0.424	0.500	0.42
Sample49		0	0.256824	1.000	0.409	0.487	0.41
Sample50		0	0.252923	1.000	0.394	0.474	0.39
Sample51		0	0.205003	1.000	0.379	0.461	0.38
Sample52		0	0.191312	1.000	0.364	0.447	0.36
Sample53		0	0.189092	1.000	0.348	0.434	0.35
Sample54		0	0.165910	1.000	0.333	0.421	0.33
Sample55		0	0.098706	1.000	0.318	0.408	0.32
Sample56		0	0.015363	1.000	0.303	0.395	0.30
Sample57		0	0.002531	1.000	0.288	0.382	0.29
Sample58		0	-0.189890	1.000	0.273	0.368	0.27
Sample59		0	-0.232356	1.000	0.258	0.355	0.26
Sample60		0	-0.376160	1.000	0.242	0.342	0.24
Sample61		0	-0.395924	1.000	0.227	0.329	0.23
Sample62		0	-0.421928	1.000	0.212	0.316	0.21
Sample63		0	-0.532388	1.000	0.197	0.303	0.20
Sample64		0	-0.919151	1.000	0.182	0.289	0.18
Sample65		0	-0.976731	1.000	0.167	0.276	0.17
Sample66		0	-1.338721	1.000	0.152	0.263	0.15
Sample67		0	-1.933180	1.000	0.136	0.250	0.14
Sample68		0	-2.599307	1.000	0.121	0.237	0.12
Sample69		0	-2.610693	1.000	0.106	0.224	0.11
Sample70		0	-2.620883	1.000	0.091	0.211	0.09
Sample71		0	-3.183215	1.000	0.076	0.197	0.08
Sample72		0	-3.495450	1.000	0.061	0.184	0.06
Sample73		0	-3.502283	1.000	0.045	0.171	0.05
Sample74		0	-6.009661	1.000	0.030	0.158	0.03
Sample75		0	-8.191465	1.000	0.015	0.145	0.02
Sample76		0	-16.059849	1.000	0.000	0.132	0.00

[0066] 표 5 및 표 6에 나타낸 바와 같이 조합위험점수를 내림차순으로 정렬하여 민감도, 특이도 및 정확도를 계산하였을 때, 아홉 번째에서 가장 좋은 Youden Index(Youden's J statistic)를 보였다. Youden Index는 ROC 커브에서 기준값(threshold)을 찾는 방법 중 하나로, '민감도 + 특이도 - 1'의 계산식을 통해 특이도와 민감도의 합이 최대가 되는 위치를 구할 수 있다. 즉, 삼중음성유방암 환자에서 DEPDC7, DIP2B, LYST, SDS, SLC43A1 및 ZNF785의 발현량을 각각 측정하고, 상기 식에 따라 얻어진 위험도 점수가 예측 기준 값인 4.977464를 초과할 경우에는 재발 및 전이의 예후를 나타내는 삼중음성유방암 환자로 판별될 수 있으며, 예측 기준 값인 4.977464 이하일 경우에는 재발 및 전이의 예후를 나타내지 않는 삼중음성유방암 환자로 판별될 수 있다. 예측 기준 값인 4.977464를 기준으로 high group과 low group으로 나누어 생존분석(Kaplan-Meier)을 수행하였다. High group과 low group의 차이에 따라서 로그 순위 검정(log-rank test)에 따라 통계적으로 유의한 생존 곡선을 나타내었다(도 1, $p < 0.0001$). 결과적으로, 표 2의 단변량 콕스 분석 결과와 도 1의 생존분석의 로그 순위 검정(log-rank test) 분석 결과로부터 통계적으로 유의성이 있음을 확인할 수 있다. 따라서, 본 발명에 의해 선별된 유전자 시그니쳐는 삼중음성유방암에서 재발 및 전이의 예후를 예측할 수 있는 바이오마커로서 유용하게 사용될 수 있다.

도면

도면1


Risk_Score(Event vs Non-Event)
Sensitivity
1 - Specificity
AUC = 0.989(95% CI, 0.966-1.01)
Event = 10
Non-Event = 66
Risk Score
p = 0.00000000832
High (n=9)
Low (n=67)
Days