(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0148943
(43) 공개일자 2023년10월26일

(51) 국제특허분류(Int. Cl.)
C12Q 1/6886 (2018.01)
(52) CPC특허분류
C12Q 1/6886 (2022.01)
C12Q 2537/165 (2013.01)
(21) 출원번호 10-2022-0047949
(22) 출원일자 2022년04월19일
심사청구일자 2022년04월19일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
한경숙
서울특별시 서초구 사임당로 137, 3동 309호(서초동, 신동아아파트)
인서뢰
인천광역시 미추홀구 재넘이길 86(용현동)
이욱
인천광역시 미추홀구 한나루로463번길 66, 201호(용현동)
(74) 대리인
특허법인태백

전체 청구항 수 : 총 5 항

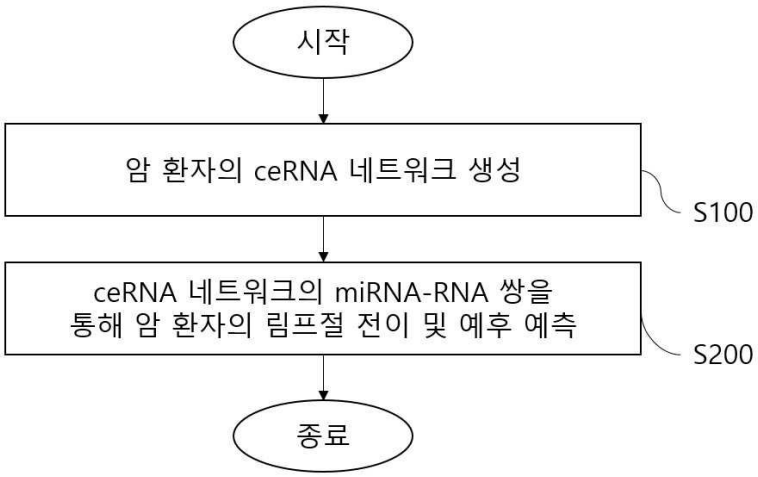
(54) 발명의 명칭 **miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법**

(57) 요 약

본 발명은 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법에 관한 것이다. 본 발명에 따르면, 암 환자의 ceRNA 네트워크를 생성하는 과정; 및 상기 ceRNA 네트워크의 miRNA-RNA 쌍을 통해 암 환자의 림프절 전이 및 예후를 예측하는 과정을 포함한다.

이와 같이 본 발명에 따르면, 종래의 임상적 진단과는 달리 림프절 전이가 완전히 진행되기 전에 림프절 전이를 예측할 수 있으며, 유전자 발현 데이터만을 이용하는 종래 방법에 비해 림프절 전이 예측에 있어서 훨씬 우수한 성능을 보여준다. 더 나아가, 주요 miRNA-RNA 쌍을 이용하여 다양한 유형의 암에서 개별 환자의 예후를 예측할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
C12Q 2600/118 (2013.01)
C12Q 2600/178 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711128786
과제번호 2020R1A2B5B01096299
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 바이오 빅데이터로부터 개인별 분자관계 네트워크 추론과 암의 조기 진단과 예후 예측을 위한 네트워크 바이오마커 발굴
기 여 율 1/1
과제수행기관명 인하대학교
연구기간 2021.03.01 ~ 2022.02.28

명 세 서

청구범위

청구항 1

암 환자의 ceRNA 네트워크를 생성하는 과정; 및

상기 ceRNA 네트워크의 miRNA-RNA 쌍을 통해 암 환자의 림프절 전이 및 예후를 예측하는 과정을 포함하며,

상기 RNA는 mRNA, lncRNA 및 유사유전자(pseudogene) 중 어느 하나인 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법.

청구항 2

제 1 항에 있어서,

상기 암 환자의 ceRNA 네트워크를 생성하는 과정은,

복수의 종양 샘플 및 정상 샘플에서 복수의 miRNA-RNA 쌍을 생성하고 MIC(maximal information coefficient)를 산출하는 단계;

상기 복수의 miRNA-RNA 쌍 중에서 miRNA-RNA 쌍의 MIC가 기 설정된 제1 임계값 이상인 쌍들만 선별하는 단계;

복수의 정상 샘플에서 상기 선별된 miRNA-RNA 쌍의 제1 PCC(Pearson CoΔrrelation Coefficient)를 산출하는 단계;

상기 복수의 정상 샘플이 단일 종양 샘플에 의해 교란된 샘플에서 miRNA-RNA 쌍의 제2 PCC를 산출하는 단계;

상기 제1 PCC 및 제2 PCC의 차이값인 ΔPCC를 산출하는 단계;

상기 ΔPCC 기반의 Wilcox test를 통해 p-값이 기 설정된 제2 임계값 미만인 쌍들만 선별하는 단계; 및

선별된 miRNA-RNA 쌍들을 이용하여 ceRNA 네트워크를 생성하는 단계를 포함하는 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법.

청구항 3

제 2 항에 있어서,

상기 암 환자의 ceRNA 네트워크를 생성하는 과정은,

대상 암의 모든 종양 샘플들에 대한 |ΔPCC| 산출 결과를 기초로 개인별 ceRNA 네트워크를 생성하는 단계를 더 포함하며,

상기 기초로 개인별 ceRNA 네트워크를 생성하는 단계는,

상기 ceRNA 네트워크의 모든 miRNA-RNA쌍들에 대해 대상 암의 모든 종양 샘플들에서 |ΔPCC|를 산출하고, 산출된 |ΔPCC|가 모든 |ΔPCC|의 중앙값보다 큰 miRNA-RNA쌍만 남겨서 개인별 ceRNA 네트워크를 생성하는 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법.

청구항 4

제 1 항에 있어서,

상기 ceRNA 네트워크의 miRNA-RNA 쌍을 통해 암 환자의 림프절 전이 및 예후를 예측하는 과정은,

상기 ceRNA 네트워크의 miRNA-RNA 쌍의 ΔPCC에 대한 PCA(Principal component analysis) 수행을 통해 특징 벡터의 차원을 줄이는 단계;

상기 miRNA-RNA 쌍의 ΔPCC를 이용하여 림프절 전이 예측을 위한 복수의 기저모델 및 이들을 조합한 앙상블 모델 중 적어도 하나를 구축하는 단계; 및

상기 miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트(log-rank test) 수행을 통해 암 환자의 예후를 예측하는 단계를 포함하는 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법.

청구항 5

제 4 항에 있어서,

상기 암 환자의 예후를 예측하는 단계는,

상기 miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트를 수행하여 얻은 p-값이 가장 작은 상위 n개(여기서, n은 기 설정된 자연수)의 miRNA-RNA 쌍의 ΔPCC를 이용하여 암 환자의 예후를 예측하는 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 miRNA가 매개하는 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이와 예후를 예측하는 방법에 관한 것으로서, 더욱 상세하게 설명하면 암 환자의 경쟁적 내인성 RNA(competitive endogenous RNA; ceRNA) 상호작용을 도출하고 miRNA가 매개하는 RNA들의 상호작용의 차등적 상관관계를 기반으로 암 환자의 림프절 전이와 예후를 예측할 수 있는 방법에 관한 것이다.

배 경 기 술

[0002] 원발성 종양에서 신체의 다른 부분으로 암세포가 퍼지는 것을 전이라고 한다. 전이 동안 암세포는 혈류나 림프계를 통해 다른 부위로 이동한다. 종양 전이의 단계 중 하나인 림프절 전이는 암 환자에서 흔히 관찰된다. 림프절 전이 자체가 환자의 생명을 직접적으로 위협하는 것은 아니지만 악성 종양은 림프절 전이를 통해 신체의 다른 부위로 전이될 수 있다. 많은 연구를 통하여 림프절 전이가 있는 환자의 예후가 림프절 전이가 없는 환자의 예후보다 나쁘다고 알려져 있다. 또한, 림프절 전이 여부는 암 환자의 효과적인 치료 방법을 결정하는 중요한 요소이다.

[0003] 림프절 전이는 일반적으로 임상 검사에서 얻은 영상을 기반으로 감지된다. 최근에는 초음파 영상에서 임상의가 림프절 전이를 감지하는 데 도움을 주기 위해 컨볼루션 뉴럴 네트워크(CNN)와 같은 딥 러닝 방법이 사용되고 있다. 이처럼 초음파 영상에서 림프절 전이를 감지하는 것은 전이를 진단하는 직접적이고 정확한 방법이지만, 림프절 전이가 발생한 후에 진단이 이루어진다는 한계가 있다.

[0004] 한편, 다양한 연구에서 림프절 전이 과정에서의 비정상적인 유전자 발현이 보고된 바 있다. 그러나 암은 복잡하고 이질적인 질병이므로 개별 유전자의 비정상적인 발현으로는 암의 발생과 전이를 충분히 설명할 수 없다. 즉, 암의 발생과 전이는 개별적인 유전자보다는 유전자 상호작용의 조절 장애로 보다 잘 설명된다.

[0005] 비특허문헌 1은 경쟁적 내인성 RNA(competitive endogenous RNA; ceRNA) 가설로 알려진 새로운 유전자 조절을 제안하였다. ceRNA 가설은 유사한 miRNA 반응 요소를 가진 RNA가 동일한 miRNA에 결합하기 위해 경쟁하여 서로를 간접적으로 조절함을 시사한다. 이러한 가설을 뒷받침하는 증거가 증가함에 따라, ceRNA 네트워크를 구축하기 위해 다양한 계산 방식이 제안되었다.

[0006] 그러나, 종래의 대부분의 방식은 ceRNA로서 mRNA 또는 lncRNA에만 초점을 맞추고 있으며, 실제 ceRNA 역할을 많이 하는 유사유전자(pseudogene)를 고려하지 않는다는 제한이 있다.

선행기술문헌

비특허문헌

[0007] (비특허문헌 0001) Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?. Cell. 2011;146(3):353-358.

(비특허문헌 0002) Zhang S, Zhang C, Du J, et al. Prediction of Lymph-Node Metastasis in Cancers Using Differentially Expressed mRNA and Non-coding RNA Signatures. Front Cell Dev Biol. 2021;9:605977.

발명의 내용

해결하려는 과제

[0008] 이와 같이 본 발명에 따르면, 암 환자의 ceRNA 상호작용을 도출하고 miRNA가 매개하는 RNA들의 상호작용의 차등적 상관관계를 기반으로 암 환자의 림프절 전이와 예후를 예측할 수 있는 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0009] 이러한 기술적 과제를 이루기 위한 본 발명의 실시예에 따르면 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법에 있어서, 암 환자의 ceRNA 네트워크를 생성하는 과정; 및 상기 ceRNA 네트워크의 miRNA-RNA 쌍을 통해 암 환자의 림프절 전이 및 예후를 예측하는 과정을 포함한다.

[0010] 상기 RNA는 mRNA, lncRNA 및 유사유전자(pseudogene) 중 어느 하나일 수 있다.

[0011] 상기 암 환자의 ceRNA 네트워크를 생성하는 과정은, 복수의 종양 샘플 및 정상 샘플에서 복수의 miRNA-RNA 쌍을 생성하고 MIC(maximal information coefficient)를 산출하는 단계; 상기 복수의 miRNA-RNA 쌍 중에서 miRNA-RNA 쌍의 MIC가 기 설정된 제1 임계값 이상인 쌍들만 선별하는 단계; 복수의 정상 샘플에서 상기 선별된 miRNA-RNA 쌍의 제1 PCC(Pearson Correlation Coefficient)를 산출하는 단계; 상기 복수의 정상 샘플이 단일 종양 샘플에 의해 교란된 샘플에서 miRNA-RNA 쌍의 제2 PCC를 산출하는 단계; 상기 제1 PCC 및 제2 PCC의 차이값인 ΔPCC를 산출하는 단계; 상기 ΔPCC 기반의 Wilcox test를 통해 p-값이 기 설정된 제2 임계값 미만인 쌍들만 선별하는 단계; 및 선별된 miRNA-RNA 쌍들을 이용하여 ceRNA 네트워크를 생성하는 단계를 포함할 수 있다.

[0012] 상기 암 환자의 ceRNA 네트워크를 생성하는 과정은, 대상 암의 모든 종양 샘플들에 대한 |ΔPCC| 산출 결과를 기초로 개인별 ceRNA 네트워크를 생성하는 단계를 더 포함할 수 있다.

[0013] 상기 기초로 개인별 ceRNA 네트워크를 생성하는 단계는, 상기 ceRNA 네트워크의 모든 miRNA-RNA쌍들에 대해 대상 암의 모든 종양 샘플들에서 |ΔPCC|를 산출하고, 산출된 |ΔPCC|가 모든 |ΔPCC|의 중앙값보다 큰 miRNA-RNA쌍만 남겨서 개인별 ceRNA 네트워크를 생성할 수 있다.

[0014] 상기 ceRNA 네트워크의 miRNA-RNA 쌍을 통해 암 환자의 림프절 전이 및 예후를 예측하는 과정은, 상기 ceRNA 네트워크의 miRNA-RNA 쌍의 ΔPCC에 대한 PCA(Principal component analysis) 수행을 통해 특징 벡터의 차원을 줄이는 단계; 상기 miRNA-RNA 쌍의 ΔPCC를 이용하여 림프절 전이 예측을 위한 복수의 기저모델 및 이들을 조합한 앙상블 모델 중 적어도 하나를 구축하는 단계; 및 상기 miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트(log-rank test) 수행을 통해 암 환자의 예후를 예측하는 단계를 포함할 수 있다.

[0015] 상기 암 환자의 예후를 예측하는 단계는, 상기 miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트를 수행하여 얻은 p-값이 가장 작은 상위 n개(여기서, n은 기 설정된 자연수)의 miRNA-RNA 쌍의 ΔPCC를 이용하여 암 환자의 예후를 예측할 수 있다.

발명의 효과

[0016] 이와 같이 본 발명에 따르면, 종래의 임상적 진단과는 달리 림프절 전이가 완전히 진행되기 전에 림프절 전이를 예측할 수 있으며, 유전자 발현 데이터만을 이용하는 종래 방법에 비해 림프절 전이 예측에 있어서 훨씬 우수한 성능을 보여준다. 더 나아가, 주요 miRNA-RNA 쌍을 이용하여 다양한 유형의 암에서 개별 환자의 예후를 예측할 수 있다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 일 실시예에 따른 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법의 흐름도이다.

도 2는 도 1에 도시된 암 환자들의 전형적인 ceRNA 네트워크 및 환자 개인별 ceRNA 네트워크 생성 과정의 상세 흐름도이다.

도 3은 도 1에 도시된 암 환자의 림프절 전이 및 예후 예측 과정의 상세 흐름도이다.

도 4는 본 발명의 일 실시예에 따른 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법의 실험 과정을 개략적으로 도시한 도면이다.

도 5는 본 발명의 일 실시예에 따라 생성된 BRCA에 대한 ceRNA 네트워크의 예를 도시하는 도면이다.

도 6a 및 도 6b는 본 발명의 일 실시예에 따라 생성된 2명의 LUAD 환자의 개인별 ceRNA 네트워크의 예를 도시하는 도면이다.

도 7은 본 발명의 일 실시예에 따라 7가지 유형의 암에서 miRNA-RNA 쌍의 ΔPCC에 대한 환자의 전체 생존율을 도시하는 도면이다.

도 8은 도 7의 miRNA-RNA 쌍에서 개별 RNA 발현 수준에 대한 환자의 전체 생존율을 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0018] 이하 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명하기로 한다. 이 과정에서 도면에 도시된 선들의 두께나 구성요소의 크기 등은 설명의 명료성과 편의상 과장되게 도시되어 있을 수 있다.

[0019] 또한 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서, 이는 사용자, 운용자의 의도 또는 관례에 따라 달라질 수 있다. 그러므로 이러한 용어들에 대한 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0020] 도 1은 본 발명의 일 실시예에 따른 miRNA를 매개로 한 RNA들의 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법의 흐름도이고, 도 2는 도 1에 도시된 암 환자들의 전형적인 ceRNA 네트워크 및 환자 개인별 ceRNA 네트워크 생성 과정의 상세 흐름도이며, 도 3은 도 1에 도시된 암 환자의 림프절 전이 및 예후 예측 과정의 상세 흐름도이다.

[0021] 도 1 내지 도 3을 참조하면, 본 발명의 일 실시예에 따른 miRNA 매개 RNA 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법은 크게 암 환자의 ceRNA 네트워크를 생성하는 과정(S100) 및 ceRNA 네트워크의 miRNA-RNA 쌍을 통해 암 환자의 림프절 전이 및 예후를 예측하는 과정(S200)을 포함할 수 있다.

[0022] 먼저, 암 환자의 ceRNA 네트워크를 생성하기 위해서, 종양 샘플 및 정상 샘플에서 복수의 miRNA-RNA 쌍을 생성하고 MIC(maximal information coefficient)를 산출할 수 있다(S110).

[0023] 여기서, RNA는 ceRNA 후보로서 mRNA, lncRNA 및 유사유전자(pseudogene)를 포함하며, 이에 따라 종양 샘플 및 정상 샘플에서 miRNA-mRNA 쌍, miRNA-lncRNA 쌍 및 miRNA-유사유전자 쌍을 생성할 수 있다.

[0024] 또한, 생성된 복수의 miRNA-RNA 쌍 각각에 대해 MIC를 계산할 수 있다. 일반적으로 피어슨 상관 계수(Pearson Correlation Coefficient, PCC)가 유전자 간 상호작용을 도출하는데 사용된다. 그러나, PCC는 선형 관련성만 측정할 수 있는데 반해 MIC는 두 변수 간의 선형 또는 비선형 관련성을 측정할 수 있고, MIC는 PCC에 비해 특이치에 덜 취약하다. 이에, 본 발명에 따르면 miRNA-RNA 상호작용을 선택하기 위해 MIC를 사용한다. 또한, miRNA와 ceRNA의 연관 강도는 종양 샘플과 정상 샘플 간에 상이하므로, 복수의 miRNA-RNA 쌍의 MIC는 종양 샘플과 정상 샘플에서 별도로 계산된다.

[0025] miRNA-RNA 쌍의 RNA는 X축과 Y축의 $n_X \times n_Y$ 빈으로 구분되는 2차원 공간에 산란되며, 여기서 X는 miRNA의 발현 수준을 나타내고 Y는 miRNA-RNA 쌍에서 mRNA, lncRNA 및 유사유전자 중 하나의 발현 수준을 나타낸다.

[0026] 각 빈의 산란점수에 근거하여, 하기의 수학식 1에 따라 상호 정보 I(X;Y)를 산출한다.

수학식 1

$$I(X,Y) = \sum_{X,Y} p(X,Y) \log_2 \frac{p(X,Y)}{p(X)p(Y)}$$

where X: miRNA

Y: mRNA, lncRNA, or pseudogene

[0027]

[0028] 이 과정은 하기의 수학식 2에 따라 최대 상호 정보를 MIC로 획득할 때까지 반복 수행된다.

수학식 2

$$MIC(X,Y) = \max_{n_X * n_Y < B} \frac{I(X,Y)}{\log_2 \min(n_X, n_Y)}$$

[0029]

[0030] 여기서, 파라미터 B는 특징 매트릭스의 양을 제어하는 것으로, B를 너무 높게 설정하면 랜덤 데이터에서도 0이 아닌 점수가 나올 수 있으며, B를 너무 낮게 설정하면 단순한 패턴만 검색된다. 본 발명에서는 B에 대한 기본 설정인 샘플 수의 0.6제곱을 사용한다.

[0031] 이후, 복수의 miRNA-RNA 쌍 중에서 miRNA-RNA 쌍의 MIC가 기 설정된 제1 임계값 이상인 쌍들만 선별할 수 있다(S120). 정상 샘플 및 종양 샘플의 miRNA-RNA 쌍의 MIC가 제1 임계값 미만인 쌍은 ceRNA 네트워크에 포함될 만큼 연관성이 강하지 않으므로 제거될 수 있다.

[0032] 여기서, 제1 임계값을 너무 작게 설정하면 miRNA-RNA 쌍이 증가하여 다수의 비논리적인 쌍이 포함될 수 있다. 반대로 제1 임계값을 너무 크게 설정하면 잠재적인 예측 유전자 쌍을 놓칠 수 있다. 일 실시예에 따르면, 제1 임계값은 범위 [0, 1]의 중앙값인 0.5로 설정될 수 있으나, 반드시 이로 제한되는 것은 아니다.

[0033] 이후, 정상 샘플에서 선별된 miRNA-RNA 쌍의 PCC(이하, '제1 PCC'라 한다)를 하기의 수학식 3에 따라 산출할 수 있다(S130).

수학식 3

$$PCC(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

where n: number of samples

X: miRNA

Y: mRNA, lncRNA, or pseudogene

[0034]

[0035] 이후, 단일 종양 샘플에 의해 교란된 샘플에서 miRNA-RNA 쌍의 PCC(이하, '제2 PCC'라 한다)를 산출할 수 있다(S140).

[0036] 일 실시예에 따르면, n개의 정상 샘플들에 임의의 암 환자의 종양 샘플 하나를 추가하여 교란된 n+1개의 샘플에서 복수의 miRNA-RNA 쌍의 PCC를 다시 산출할 수 있다.

[0037] 이후, 제1 PCC 및 제2 PCC의 차이값(이하, 'ΔPCC'라 한다)을 하기의 수학식 4에 따라 산출할 수 있다(S150). 여기서, $PCC_n(X,Y)$는 제1 PCC를 나타내고, $PCC_{n+1}(X,Y)$는 제2 PCC를 나타내며, n은 정상 샘플 수를 나타낸다.

수학식 4

[0038]
$$\Delta PCC(X,Y) = PCC_{n+1}(X,Y) - PCC_n(X,Y)$$

[0039] 이후, ΔPCC 기반의 Wilcox test를 통해 p-값(p-value)이 기 설정된 제2 임계값 미만인 쌍들만 선별하고(S160) 선별된 miRNA-RNA 쌍들을 이용하여 ceRNA 네트워크를 생성할 수 있다(S170). 일 실시예에 따르면, 제1 임계값은 0.01로 설정될 수 있으나, 반드시 이로 제한되는 것은 아니다. 필요에 따라, S160 단계 후에 주성분 분석(Principal component analysis, PCA)(S210)을 먼저 수행하여 miRNA-RNA 쌍의 수를 더욱 줄여서 특징 벡터의 차원을 축소한 후에 ceRNA 네트워크를 생성할 수도 있다.

[0040] 여기서, ceRNA 네트워크의 노드는 miRNA, mRNA, lncRNA 또는 유사유전자 중 하나를 나타내고, ceRNA 네트워크의 모든 에지는 miRNA와 다른 RNA 간의 상호작용을 나타내는 것으로, 보다 구체적으로는 단일 종양 샘플에 의해 교란된 샘플과 복수의 정상 샘플 사이의 miRNA-RNA 쌍의 PCC 차이값, 즉 ΔPCC를 나타낸다. 이 단계에서는 MIC는 사용하지 않는데, 이는 ΔPCC가 단일 종양 샘플에 의한 교란을 정량화하는데 보다 적합하고, ΔMIC는 정의상 의미가 없기 때문이다.

[0041] 이후, 대상 암의 모든 종양 샘플들에 대한 |ΔPCC| 산출 결과를 기초로 개인별 ceRNA 네트워크를 생성할 수 있다(S180).

[0042] 일 실시예에 따르면, ceRNA 네트워크의 모든 miRNA-RNA쌍들에 대해 대상 암의 모든 종양 샘플들에서 |ΔPCC|를 산출하고, 산출된 |ΔPCC|가 모든 |ΔPCC|의 중앙값보다 큰 miRNA-RNA쌍만 남겨서 개인별 ceRNA 네트워크를 생성할 수 있다. 여기서, 개인별 ceRNA 네트워크는 S170 단계에서 생성된 ceRNA 네트워크의 서브 네트워크로서, 개인별 ceRNA 네트워크의 에지는 대상 암과 동일한 유형의 암 환자들과 비교하여 유의한 변화를 보이는 miRNA-RNA 상호작용을 나타낸다.

[0043] 한편, ceRNA 네트워크의 miRNA-RNA 쌍을 통해 암 환자의 림프절 전이 및 예후를 예측하기 위해서, 먼저 ceRNA 네트워크의 miRNA-RNA 쌍들의 ΔPCC에 대한 PCA 수행을 통해 특징 벡터의 차원을 줄일 수 있다(S210).

[0044] 이후, miRNA-RNA 쌍들의 ΔPCC를 이용하여 림프절 전이 예측을 위한 복수의 기저모델 및 이들을 조합한 앙상블 모델 중 적어도 하나를 구축할 수 있다(S220).

[0045] 일 실시예에 따르면, 림프절 전이 예측을 위한 기저모델로서 서포트 벡터 머신(Support vector machine, SVM) 및 로지스틱 회귀(logistic regression, LR) 모델을 사용할 수 있으나, 예측 모델의 종류가 반드시 이로 제한되는 것은 아니다. 또한, 복수의 기저모델을 배깅(bagging), 부스팅(boosting) 및 스태킹(stacking) 등과 같은 앙상블 학습 기법에 의해 조합하여 앙상블 모델을 구축할 수 있다. 예측 모델을 구축하는 구체적인 방법은 도 4를 참조하여 구체적으로 후술한다.

[0046] 이후, miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트(log-rank test) 수행을 통해 암 환자의 예후를 예측할 수 있다(S230).

[0047] 일 실시예에 따르면, miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트를 수행하여 얻은 p-값이 가장 작은 상위 n개(여기서, n은 기 설정된 자연수)의 miRNA-RNA 쌍의 ΔPCC를 이용하여 암 환자의 예후를 예측할 수 있다.

[0048] 도 4는 본 발명의 일 실시예에 따른 miRNA 매개 RNA 상호작용을 기반으로 한 암 환자의 림프절 전이 및 예후 예측 방법의 실험 과정을 개략적으로 도시한 도면이다.

[0049] 도 4를 참조하면, 먼저, TCGA(The Cancer Genome Atlas) 데이터 포털에서 모든 유형의 고형암의 원발 종양 샘플에 대해 lncRNA, mRNA, 유사유전자 및 miRNA의 유전자 발현 프로파일과 임상 프로파일을 수집하였다. 또한, 각 암 유형의 정상 샘플도 TCGA 데이터 포털에서 획득하였다. 본 실험에서 사용된 모든 유전자 발현 프로파일은 RNA 시퀀싱(RNA-seq)에 의해 얻어졌다.

[0050] TCGA에는 샘플이 200개 이상인 18가지 유형의 고형암이 있다. 18가지 유형 중 6가지 유형은 종양 샘플의 림프절 전이 데이터가 불충분하여 제외하였다. 나머지 12가지 유형의 고형암에서는 30개 이상의 정상 샘플과 림프절 전이가 있는 50개 이상의 종양 샘플을 가진 유형의 고형암을 선정하였다. 이러한 기준에 부합하는 고형암은 유방암(breast invasive carcinoma, BRCA), 결장암(colon adenocarcinoma, COAD), 두경부암(head and neck squamous cell carcinoma, HNSC), 폐선암(lung adenocarcinoma, LUAD), 폐편평상피세포암(lung squamous cell carcinoma, LUSC), 위암(stomach adenocarcinoma, STAD) 및 갑상선암(thyroid carcinoma, THCA)의 7가지이다.

[0051] TCGA 데이터에는 암의 진행 정도(즉, 병기)에 대한 정보가 포함되어 있는데, 암의 병기를 국제 표준인 TNM 병기 분류법에 따라 표기한다. TNM은 세 가지 요소인 T (Tumor, 종양의 크기), N (Node, 전이 정도), M (Metastasis, 원격 전이 유무)를 고려하여 병기를 분류한다. 원격 장기 전이는 림프절 전이와 공존하는 경우가 많고 예측 평가를 어렵게 하므로 M 단계 1인 샘플은 제외하였다. 또한, TNM 병기분류법을 기반으로 종양 샘플을 하기와 같이 림프절 전이 유, 무로 클러스터링 하였다.

[0052] * 림프절 전이가 있는 샘플: T기 1~4, N기 1~3, M기 0의 종양 샘플

[0053] * 림프절 전이가 없는 샘플: T기 1~4, N기 0, M기 0의 종양 샘플

[0054] 표 1은 본 실험에서 사용된 7가지 유형의 암에서 정상 샘플, 종양 샘플, 림프절 전이가 있는 종양 샘플 및 림프절 전이가 없는 종양 샘플 수를 정리한 것이다.

표 1

[0055]

cancer	#normal samples	#tumor samples	#lymph node metastasis	#non-metastasis
BRCA	113	1102	447	457
COAD	41	478	107	242
HNSC	44	500	98	81
LUAD	59	533	123	231
LUSC	49	502	149	259
STAD	32	375	210	103
THCA	58	502	127	145

[0056] TCGA 데이터의 유전자 이름은 Ensembl ID로 표시된다. 따라서, Ensembl 프로젝트에서 주석 파일을 얻고 유전자의 이름과 RNA 유형(mRNA, lncRNA, 유사유전자 및 miRNA)을 결정하였다. 표 2는 유전자의 수와 RNA 유형을 정리한 것이다.

표 2

[0057]

cancer	#miRNAs	#mRNAs	#lncRNAs	#pseudogenes
BRCA	165	18,084	8,553	5,528
COAD	157	17,573	7,284	5,304
HNSC	95	18,018	7,427	4,643
LUAD	197	18,054	8,755	5,954
LUSC	161	18,227	8,706	5,680
STAD	379	18,617	10,354	9,039
THCA	153	17,568	7,342	4,753

[0058] 이후, 유전자 필터링을 수행하여 평균 카운트 수가 1 미만인 유전자를 걸러냈다. RNA-seq 데이터에서 카운트 수는 음이 아닌 정수 값이다. 발현되지 않은 유전자의 카운트 수는 0이므로 발현된 유전자의 카운트 수는 적어도

1개이다. 평균 카운트 수 < 1인 유전자는 대부분의 샘플에서 발현되지 않은 유전자이기 때문에 제거하였다. 그런 다음 M 값의 트리밍 평균(trimmed mean of M-values; TMM) 방법을 사용하여 유전자의 RNA-seq 데이터를 정규화하였다.

[0059] 이후, 도 2를 참조하여 상술한 바에 따라 암 환자의 개인별 ceRNA 네트워크를 생성하였다. 표 3은 도 2를 참조하여 상술한 각 선별 단계를 통해 선별된 miRNA-RNA 쌍의 수를 정리한 것이다.

표 3

[0060]

cancer	#features after MIC filtering	#features after Wilcox test	#features after PCA
BRCA	90,837	1,563	480
COAD	178,973	1,969	80
HNSC	67,020	800	100
LUAD	341,146	12,981	200
LUSC	165,765	2,436	200
STAD	976,763	17,445	60
THCA	38,077	3,397	150

[0061] 도 5는 본 발명의 일 실시예에 따라 생성된 BRCA에 대한 ceRNA 네트워크의 예를 도시하는 도면이다.

[0062] 도 5를 참조하면, ceRNA 네트워크는 BRCA에 대한 모든 miRNA-RNA 쌍으로 구성되며, ceRNA 네트워크에는 119개의 miRNA, 423개의 lncRNA, 380개의 mRNA 및 252개의 유사유전자 간의 1,563개의 miRNA-RNA 상호작용이 포함되어 있다.

[0063] miR-149를 중심으로 한 소규모 네트워크는 빨간색 상자로 둘러싸인 서브 네트워크를 확대한 것이다. miR-149는 ceRNA 네트워크에서 가장 빈번하게 ceRNA와 상호작용하는 miRNA로서, 하향 조절 시 유방암 전이를 촉진하는 것으로 알려져 있다.

[0064] 또한, ceRNA 네트워크는 유방암과 관련된 여러 유전자를 포함함을 확인할 수 있다. 예를 들어, ERBB4의 돌연변이는 유방암과 관련이 있는 것으로 알려져 있고, YWHAE의 과잉 발현은 유방암 세포의 증식, 이동 및 침입 능력을 증가시키며, KAT6A는 발암성 염색질 조절인자 TRIM24에 대한 SMAD3 결합을 촉진하고 종양 억제제 TRIM33과의 상호작용을 방해하여 유방암에서 종양 세포 전이를 초래한다.

[0065] 도 6a 및 도 6b는 본 발명의 일 실시예에 따라 생성된 2명의 LUAD 환자의 개인별 ceRNA 네트워크의 예를 도시하는 도면이다. 구체적으로, 도 6a 및 도 6b는 miR-581_LINK00628 쌍의 상이한 ΔPCC를 가진 2명의 LUAD 환자의 개인별 ceRNA 네트워크를 도시하는 것으로, 도 6a는 해당 쌍의 높은 ΔPCC 그룹에 포함된 LUAD 환자의 개인별 ceRNA 네트워크를 도시하고, 도 6b는 해당 쌍의 낮은 ΔPCC 그룹에 포함된 LUAD 환자의 개인별 ceRNA 네트워크를 도시한다.

[0066] 도 6a의 네트워크는 29개의 miRNA, 77개의 lncRNA, 47개의 mRNA 및 38개의 의사유전자 간의 210개의 miRNA-RNA 쌍으로 구성되며, 도 6b의 네트워크는 5개의 miRNA, 53개의 lncRNA, 30개의 mRNA 및 19개의 유사유전자 간의 111개의 miRNA 쌍으로 구성된다.

[0067] 도 6a의 네트워크는 도 6b의 네트워크에 비해 더 많은 RNA와 이들 사이의 상호작용을 포함하고 있는 것을 알 수 있다. 또한, 도 7을 참조하여 후술하는 바와 같이, miR-581_LINC00628 쌍의 ΔPCC가 높은 환자는 해당 쌍의 ΔPCC가 낮은 환자에 비해 생존율이 훨씬 낮다.

[0068] 이후, 도 3을 참조하여 상술한 바에 따라 ceRNA 네트워크의 miRNA-RNA 쌍들의 ΔPCC를 이용하여 7가지 유형의 암에 대해 림프절 전이 예측을 위한 모델을 구축하고 이를 이용하여 암 환자의 림프절 전이 및 예후를 예측하였다.

[0069] 본 실험에서는 림프절 전이 예측을 위한 기저모델로서 SVM 및 LR 모델을 선택하고, 앙상블 학습 기법 중에서 예측 정확도가 높고 비교적 오버피팅 위험이 낮은 것으로 알려진 스태킹 기법에 의해 기저모델을 결합하여 앙상블 모델을 구축하였다.

[0070] 일 예로, Python의 머신러닝 라이브러리인 scikit-learn 라이브러리의 스태킹 앙상블 학습을 사용하여 SVM 및 LR 모델을 결합하였다. 원본 훈련 세트를 사용하여 기본 학습자인 SVM 및 LR 모델을 훈련시키고, 이들 모델의 예측 결과를 보조 학습자인 LR 모델을 훈련하기 위한 특징으로 사용하였다.

[0071] TCGA에서 획득한 종양 샘플은 7:3의 비율로 훈련 및 테스트 세트로 구분하였고, 예측 모델의 파라미터는 훈련 세트의 그리드 검색을 통해 결정되었다. 예측 모델의 훈련 및 검증에 있어서, 림프절 전이가 있는 종양 샘플은 양성 사례로, 림프절 전이가 없는 종양 샘플은 음성 사례로 간주되었다.

[0072] 본 발명의 실시예에 대한 성능 평가를 위해, miRNA-RNA 쌍의 ΔPCC와 RNA 발현 수준의 두 가지 종류의 특징을 각각 사용한 스태킹 모델의 성능을 비교하였다. RNA 발현 수준 특징의 경우 정상 샘플과 종양 샘플 간 차등적 유전자 발현 분석과 림프절 전이 샘플과 비전이 샘플 간의 차등적 유전자 발현 분석 모두에서 p-값이 0.01 미만인 RNA를 사용하였다. 또한, 본 발명의 실시예에 따른 스태킹 모델의 성능은 민감도(sensitivity, SN), 특이도(specificity, SP), 정확도(accuracy, ACC), 양의 예측값(positive predictive value, PPV), 음의 예측값(negative predictive value, NPV) 및 곡선 아래 면적(area under the curve, AUC)의 척도를 사용하여 5배 교차 검증 및 독립 테스트에 의해 평가하였다.

[0073] 표 4 및 표 5는 각각 5배 교차 검증과 독립 테스트에서 ΔPCC와 RNA 발현 수준(EXP)을 각각 특징으로 이용한 스태킹 모델의 성능을 보여준다. 여기서, PC(principal component)는 주성분을 나타낸다.

표 4

[0074]

cancer	feature	#features	#PCs	SN	SP	ACC	PPV	NPV	AUC
BRCA	EXP	5,119	430	0.674	0.709	0.692	0.694	0.689	0.691
	ΔPCC	1,563	480	**0.773**	**0.806**	**0.790**	**0.796**	**0.784**	**0.789**
COAD	EXP	835	100	0.360	0.935	0.758	0.711	0.767	0.647
	ΔPCC	1,969	80	**0.760**	**0.965**	**0.902**	**0.905**	**0.901**	**0.862**
HNSC	EXP	292	10	0.750	0.684	0.720	0.739	0.696	0.717
	ΔPCC	800	100	**0.956**	**0.877**	**0.920**	**0.903**	**0.943**	**0.917**
LUAD	EXP	6,193	110	0.477	0.882	0.741	0.683	0.759	0.679
	ΔPCC	12,981	200	**0.593**	**0.944**	**0.822**	**0.850**	**0.813**	**0.769**
LUSC	EXP	1,371	190	0.644	0.867	0.786	0.736	0.809	0.756
	ΔPCC	2,436	200	**0.875**	**0.934**	**0.912**	**0.884**	**0.929**	**0.904**
STAD	EXP	476	120	0.905	0.472	0.763	0.778	0.708	0.688
	ΔPCC	17,445	60	**0.973**	**0.903**	**0.950**	**0.953**	**0.942**	**0.938**
THCA	EXP	4,205	30	0.663	0.663	0.663	0.634	0.691	0.663
	ΔPCC	3,397	150	**0.674**	**0.723**	**0.700**	**0.682**	**0.716**	**0.698**

표 5

[0075]

cancer	feature	#features	#PCs	SN	SP	ACC	PPV	NPV	AUC
BRCA	EXP	5,119	430	0.664	0.710	0.688	0.690	0.685	0.687
	ΔPCC	1,563	480	**0.776**	**0.826**	**0.802**	**0.813**	**0.792**	**0.801**
COAD	EXP	835	100	0.563	0.932	0.819	0.783	0.829	0.747
	ΔPCC	1,969	80	**0.906**	**0.986**	**0.962**	**0.967**	**0.960**	**0.946**
HNSC	EXP	292	10	0.867	0.792	0.833	0.839	0.826	0.829
	ΔPCC	800	100	**0.967**	**0.792**	**0.889**	**0.853**	**0.950**	**0.879**
LUAD	EXP	6,193	110	0.622	0.943	0.832	0.852	0.825	0.782
	ΔPCC	12,981	200	**0.784**	**0.971**	**0.907**	**0.936**	**0.895**	**0.878**
LUSC	EXP	1,371	190	0.533	0.808	0.707	0.615	0.750	0.671
	ΔPCC	2,436	200	**0.889**	**0.962**	**0.935**	**0.930**	**0.938**	**0.925**
STAD	EXP	476	120	0.937	0.452	0.777	0.776	0.778	0.694
	ΔPCC	17,445	60	**0.905**	**0.968**	**0.926**	**0.983**	**0.833**	**0.936**
THCA	EXP	4,205	30	**0.737**	0.796	0.768	0.757	**0.778**	**0.766**
	ΔPCC	3,397	150	0.658	**0.864**	**0.768**	**0.807**	0.745	0.761

[0076] 표 4 및 표 5를 참조하면, ΔPCC을 이용한 스태킹 모델은 독립 테스트에서 THCA을 제외하고, 그리고 5배 교차

검증과 독립 테스트 모두에서 RNA 발현 수준을 이용한 스태킹 모델보다 더 나은 성능을 보임을 알 수 있다. 이로부터 miRNA-RNA 쌍의 ΔPCC가 림프절 전이를 예측하는 데 있어서 유전자 발현 수준보다 더 강력한 특징임을 알 수 있으며, 더 나아가 이는 림프절 전이가 개별 유전자보다는 유전자 상호작용의 조절 장애와 관련이 있음을 뒷받침한다.

[0077] 또한, 동일한 데이터 세트 및 동일한 모델을 사용하여 비특허문헌 2(이후, 'Zhang'이라 한다)에 개시된 방법과 본 발명의 실시예의 성능을 비교하였다. 표 6은 본 발명과 Zhang에서 공통적인 4가지의 암 유형에 대한 성능을 보여준다.

표 6

cancer	method_feature	train_score	test_score
BRCA	our model_ΔPCC	**0.972**	**0.787**
	Zhang_mRNA	0.798	0.680
	Zhang_miRNA	0.764	0.737
	Zhang_lncRNA	0.793	0.696
COAD	our model_ΔPCC	**0.984**	**0.905**
	Zhang_mRNA	0.849	0.871
	Zhang_miRNA	0.902	0.886
	Zhang_lncRNA	0.869	0.871
LUAD	our model_ΔPCC	**0.996**	**0.850**
	Zhang_mRNA	0.808	0.849
	Zhang_miRNA	0.885	0.795
	Zhang_lncRNA	0.798	0.849
LUSC	our model_ΔPCC	**0.999**	**0.904**
	Zhang_mRNA	0.871	0.900
	Zhang_miRNA	0.939	0.847
	Zhang_lncRNA	0.861	0.900

[0078]

[0079] 표 6의 훈련 점수와 테스트 점수는 Zhang에서 사용된 scikit-learn 패키지를 사용하여 얻었다. 표 6을 참조하면, 성능 비교에 사용된 모든 암 유형에서 ΔPCC를 특징으로 사용한 본 발명의 실시예에 따른 모델이 mRNA, miRNA 및 lncRNA 발현 수준을 각각 사용한 Zhang의 4개의 SVM 모델에 비해 우수함을 알 수 있다. 이로부터 림프절 전이를 예측할 때 miRNA-RNA 쌍의 ΔPCC가 RNA 발현 데이터보다 훨씬 강력한 특징임을 알 수 있다.

[0080] 더 나아가, 본 발명의 실시예에 따라 획득한 miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트를 수행하여 암 환자의 전체 생존율을 분석하였다.

[0081] 표 7은 각 암 유형에서 로그 순위 테스트의 p-값이 가장 작은 상위 3개의 miRNA-RNA 쌍을 보여준다.

표 7

cancer	miRNA-RNA pair	type of RNA in the pair	p-value of miRNA-RNA pair	p-value of miRNA	p-value of RNA
BRCA	miR-26b_AC079414.1	lncRNA	2.270E-05	9.203E-01	5.896E-01
	miR-3192_PPDPFL	mRNA	6.320E-05	1.351E-03	1.346E-02
	miR-3192_AC013549.3	lncRNA	.260E-04	5.028E-01	1.346E-02
COAD	miR-604_AL162426.1	lncRNA	1.869E-04	4.365E-01	6.730E-01
	miR-3679_RPL26P29	pseudogene	3.122E-04	1.315E-02	8.171E-01
	miR-6835_AC037459.2	lncRNA	7.746E-04	9.815E-01	2.938E-02
HNSC	miR-4539_KRTAP10-2	mRNA	1.849E-04	3.033E-01	1.629E-03
	miR-6730_LINC01435	lncRNA	9.783E-04	1.038E-02	3.211E-03
	miR-5195_AL390067.1	lncRNA	1.070E-03	8.716E-02	3.435E-02
LUAD	miR-581_LINC00628	lncRNA	4.719E-07	1.925E-02	8.736E-01
	miR-7848_AC087588.2	lncRNA	2.220E-06	1.750E-05	7.506E-01
	miR-3680-1_AL138789.1	lncRNA	1.300E-05	2.386E-02	5.371E-01
LUSC	miR-548z_PNLIPRP2	pseudogene	1.175E-04	3.178E-01	6.640E-04
	miR-3972_CSAG4	pseudogene	1.485E-04	5.168E-01	4.740E-01
	miR-146b_PHETA2	mRNA	1.488E-04	4.779E-02	2.760E-01
STAD	miR-604_OLFML3	mRNA	1.000E-05	4.787E-02	4.921E-01
	miR-554_OR10A5	mRNA	4.040E-05	4.727E-03	5.852E-02
	miR-149_OR10A5	mRNA	1.689E-04	4.727E-03	8.850E-01
THCA	miR-5685_GADD45A	mRNA	3.489E-03	7.915E-01	2.587E-01
	miR-6784_AC093281.2	lncRNA	3.762E-03	5.934E-01	5.559E-02
	miR-8071-2_CFB	mRNA	3.991E-03	1.392E-02	9.494E-01

[0082]

[0083] 표 7을 참조하면, ΔPCC를 사용한 로그 순위 테스트의 p-값은 miRNA-RNA 쌍에 관여하는 개별 RNA의 p-값보다 훨씬 작은 것을 알 수 있다. 또한, p-값이 가장 작은 상위 3개의 miRNA-RNA 쌍에는 3개의 의사유전자(RPL26P29, PNLIRP2, CSAG4)가 포함되어 있다.

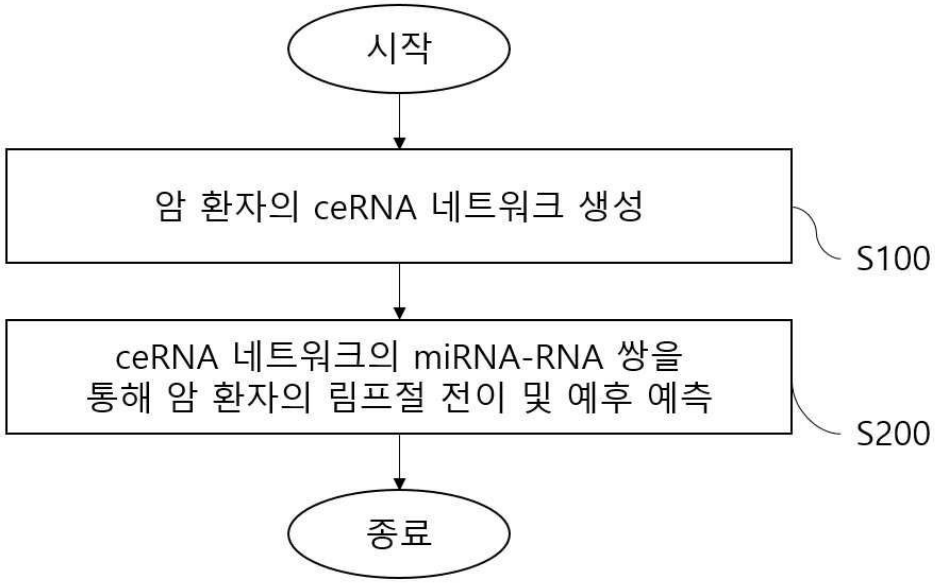
[0084] 도 7은 본 발명의 일 실시예에 따라 7가지 유형의 암에서 miRNA-RNA 쌍의 ΔPCC에 대한 환자의 전체 생존율을 도시하는 도면이고, 도 8은 도 7의 miRNA-RNA 쌍에서 개별 RNA 발현 수준에 대한 환자의 전체 생존율을 도시하는 도면이다. 도 8에서는 도 7의 miRNA-RNA 쌍(BRCA의 경우 miR-26B_AC079414.1 쌍, COAD의 경우 miR-604_AL162426.1 쌍, LUAD의 경우 miR-581_LINK00628 쌍)에 포함된 RNA가 비교 대상으로 선택되었다.

[0085] 도 7 및 도 8을 참조하면, 모든 암 유형에서 miRNA-RNA 쌍의 ΔPCC는 환자의 생존율을 예측하는 데 매우 강력한 반면에, 동일한 miRNA-RNA 쌍에 관련된 개별 RNA는 암 환자의 생존율을 예측하지 못함을 알 수 있다.

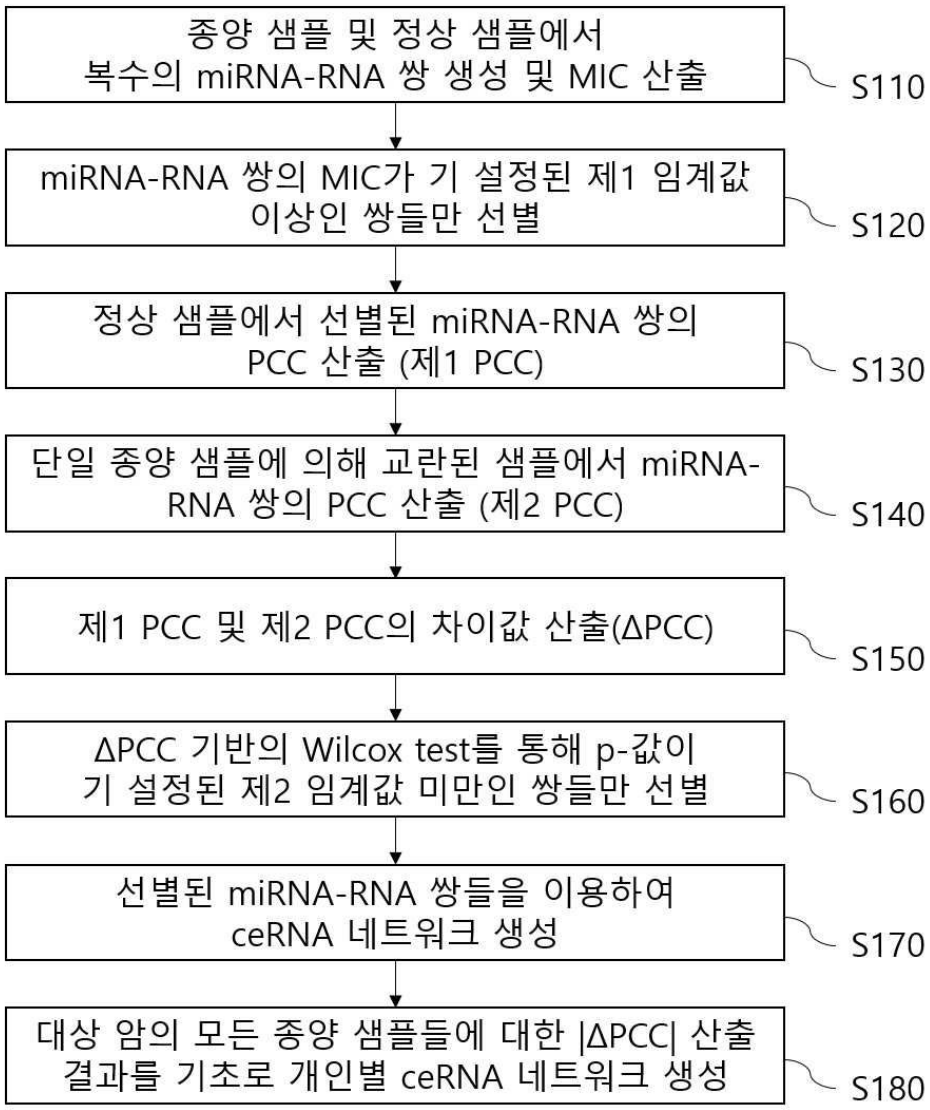
[0086] 본 발명은 도면에 도시된 실시예를 참고로 하여 설명되었으나 이는 예시적인 것에 불과하며, 당해 기술이 속하는 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 아래의 특허청구범위의 기술적 사상에 의하여 정해져야 할 것이다.

도면

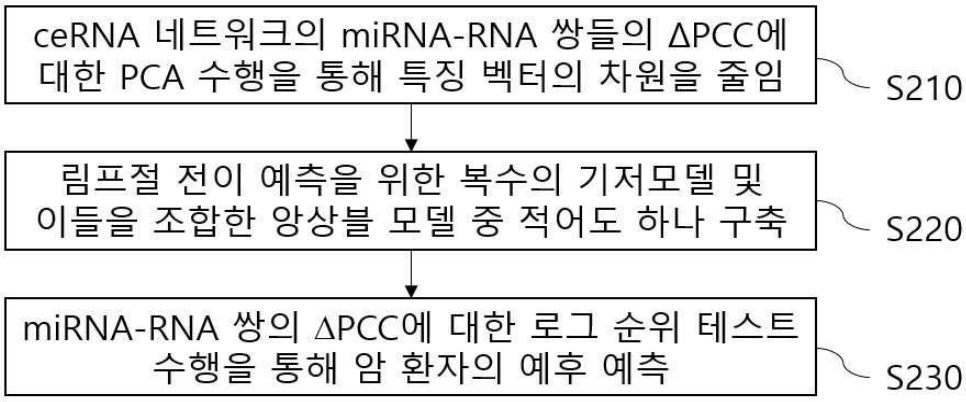
도면1


시작
암 환자의 ceRNA 네트워크 생성
S100
ceRNA 네트워크의 miRNA-RNA 쌍을
통해 암 환자의 림프절 전이 및 예후 예측
S200
종료

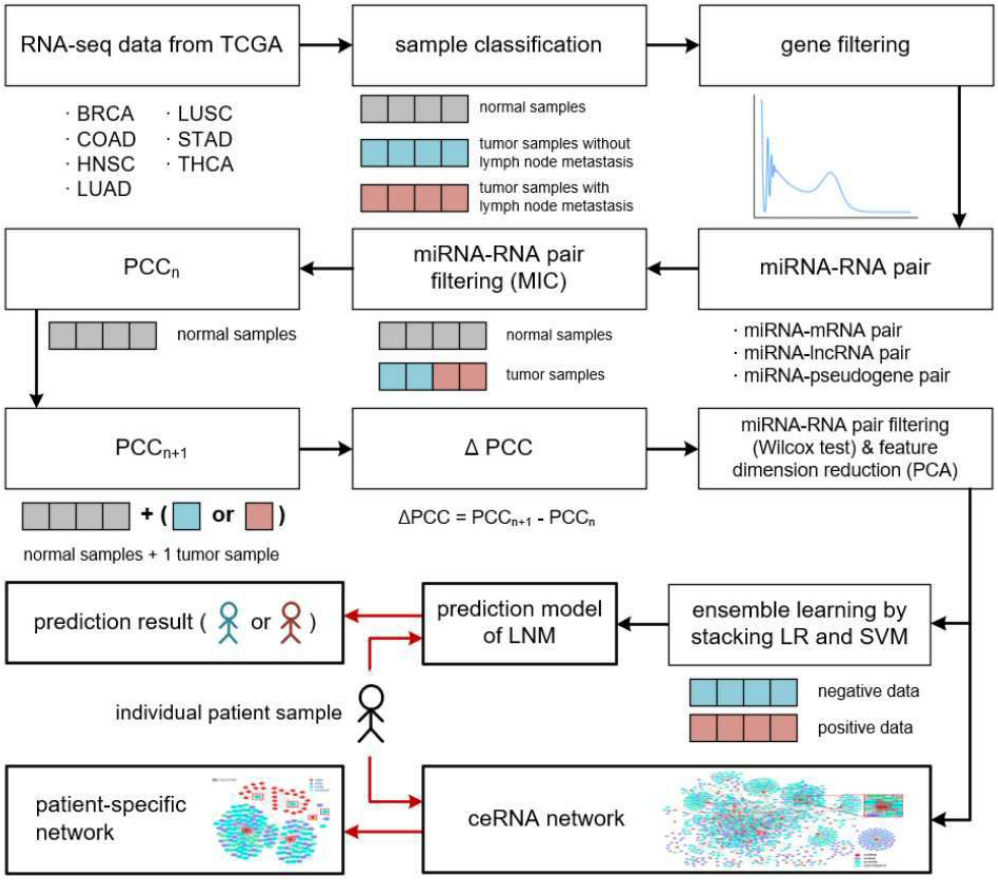
도면2


종양 샘플 및 정상 샘플에서
복수의 miRNA-RNA 쌍 생성 및 MIC 산출
S110
miRNA-RNA 쌍의 MIC가 기 설정된 제1 임계값
이상인 쌍들만 선별
S120
정상 샘플에서 선별된 miRNA-RNA 쌍의
PCC 산출 (제1 PCC)
S130
단일 종양 샘플에 의해 교란된 샘플에서 miRNA-
RNA 쌍의 PCC 산출 (제2 PCC)
S140
제1 PCC 및 제2 PCC의 차이값 산출(ΔPCC)
S150
ΔPCC 기반의 Wilcox test를 통해 p-값이
기 설정된 제2 임계값 미만인 쌍들만 선별
S160
선별된 miRNA-RNA 쌍들을 이용하여
ceRNA 네트워크 생성
S170
대상 암의 모든 종양 샘플들에 대한 |ΔPCC| 산출
결과를 기초로 개인별 ceRNA 네트워크 생성
S180

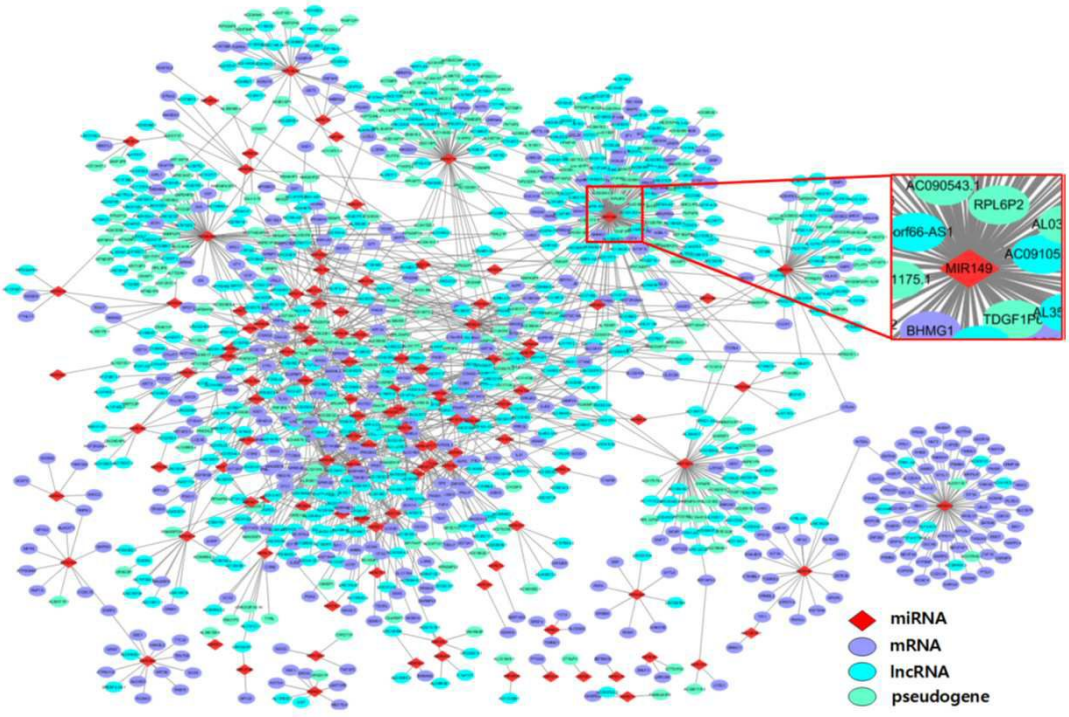
도면3


ceRNA 네트워크의 miRNA-RNA 쌍들의 ΔPCC에
대한 PCA 수행을 통해 특징 벡터의 차원을 줄임
S210
림프절 전이 예측을 위한 복수의 기저모델 및
이들을 조합한 앙상블 모델 중 적어도 하나 구축
S220
miRNA-RNA 쌍의 ΔPCC에 대한 로그 순위 테스트
수행을 통해 암 환자의 예후 예측
S230

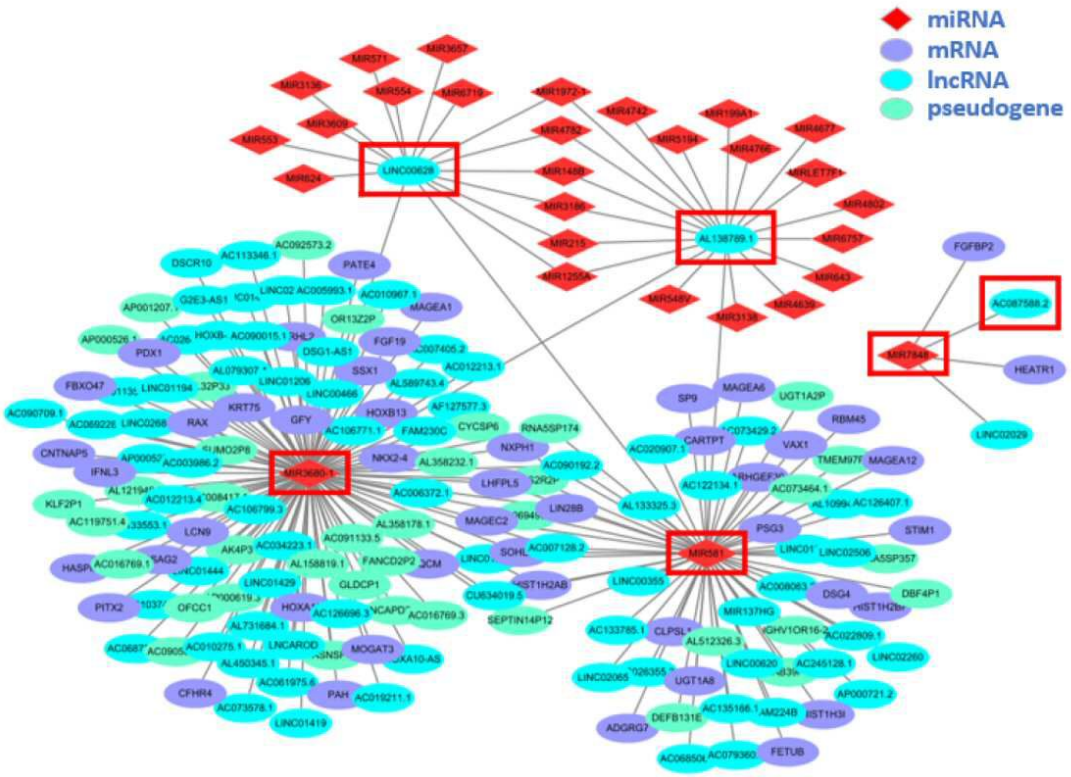
도면4


RNA-seq data from TCGA
sample classification
gene filtering
· BRCA · LUSC
· COAD · STAD
· HNSC · THCA
· LUAD
normal samples
tumor samples without lymph node metastasis
tumor samples with lymph node metastasis
PCCn
miRNA-RNA pair filtering (MIC)
miRNA-RNA pair
normal samples
normal samples
tumor samples
· miRNA-mRNA pair
· miRNA-lncRNA pair
· miRNA-pseudogene pair
PCCn+1
Δ PCC
miRNA-RNA pair filtering (Wilcox test) & feature dimension reduction (PCA)
+ (or)
ΔPCC = PCCn+1 - PCCn
normal samples + 1 tumor sample
prediction result (or)
prediction model of LNM
ensemble learning by stacking LR and SVM
negative data
positive data
individual patient sample
patient-specific network
ceRNA network

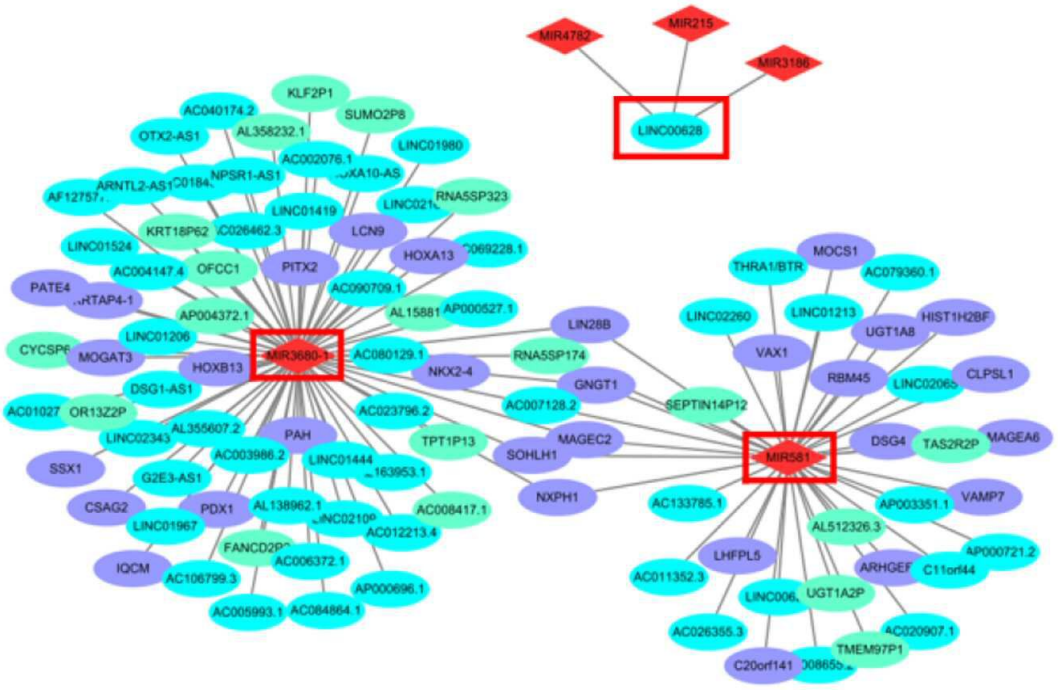
도면5


AC090543.1
RPL6P2
orf66-AS1
AL03
AC09105
MIR149
1175.1
TDGF1P2
BHMG1
miRNA
mRNA
lncRNA
pseudogene

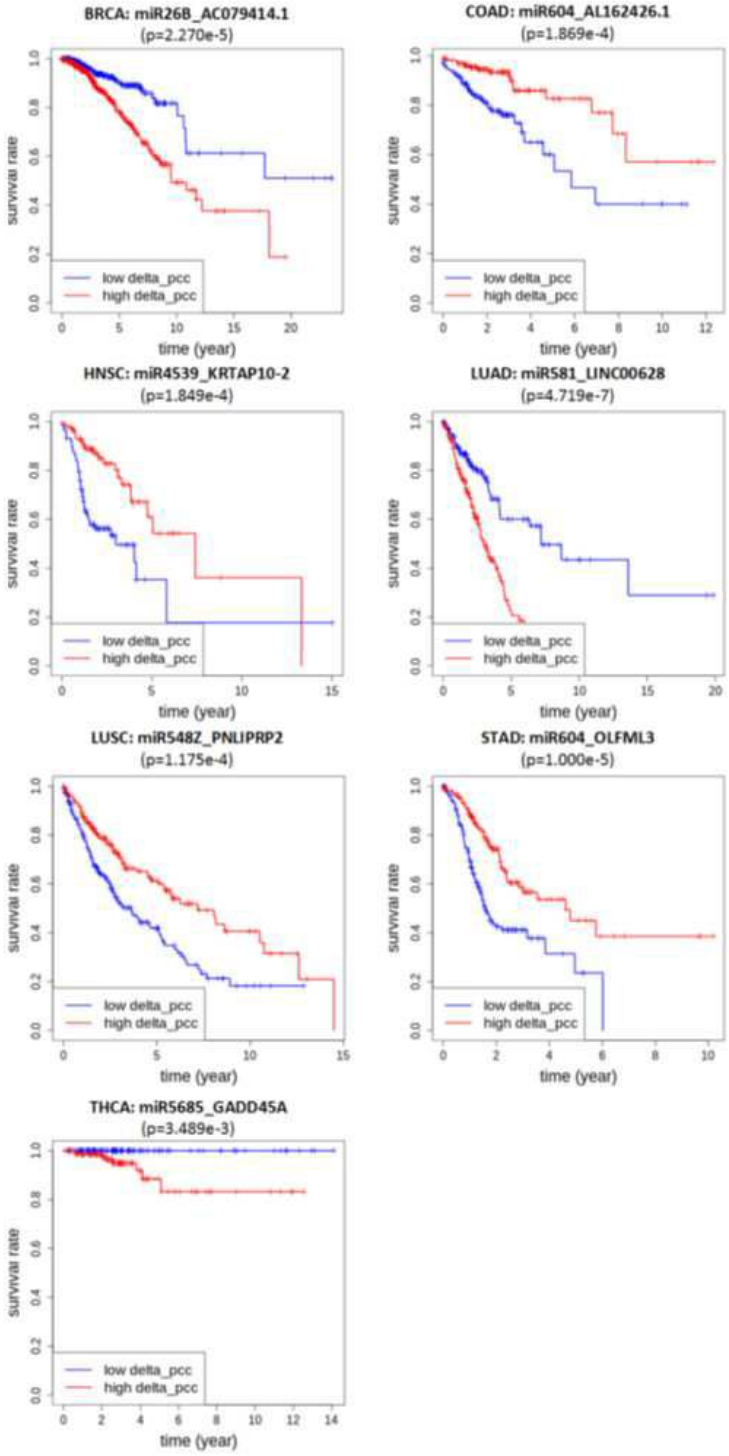
도면6a



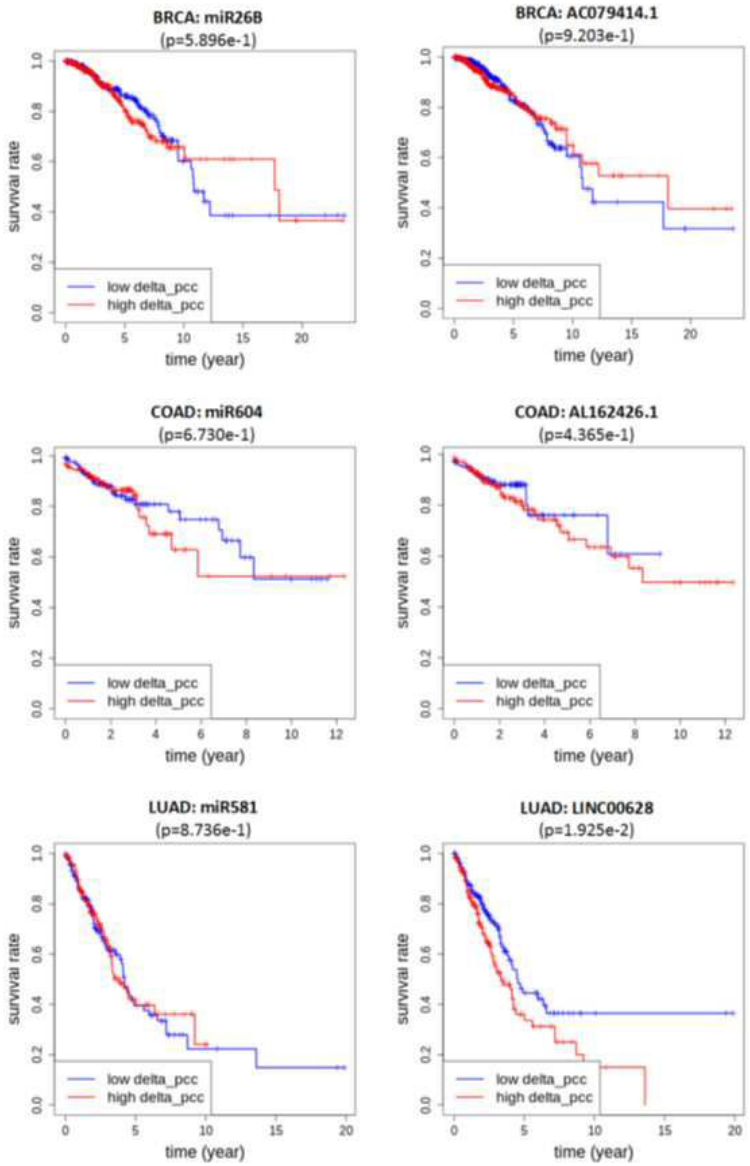
도면6b



도면7


BRCA: miR26B_AC079414.1
(p=2.270e-5)
COAD: miR604_AL162426.1
(p=1.869e-4)
HNSC: miR4539_KRTAP10-2
(p=1.849e-4)
LUAD: miR581_LINC00628
(p=4.719e-7)
LUSC: miR548Z_PNLIPRP2
(p=1.175e-4)
STAD: miR604_OLFML3
(p=1.000e-5)
THCA: miR5685_GADD45A
(p=3.489e-3)
survival rate
time (year)
low delta_pcc
high delta_pcc

도면8


BRCA: miR26B
(p=5.896e-1)
BRCA: AC079414.1
(p=9.203e-1)
COAD: miR604
(p=6.730e-1)
COAD: AL162426.1
(p=4.365e-1)
LUAD: miR581
(p=8.736e-1)
LUAD: LINC00628
(p=1.925e-2)
survival rate
time (year)
low delta_pcc
high delta_pcc