(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0125030
(43) 공개일자 2015년11월09일

(51) 국제특허분류(Int. Cl.)
C12Q 1/68 (2006.01)
(21) 출원번호 10-2014-0051226
(22) 출원일자 2014년04월29일
심사청구일자 2015년02월04일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김기열
서울 서대문구 연세로 50, 치과대학 구강종양연구소 (신촌동, 연세대학교)
차인호
서울특별시 종로구 통일로 246-20 103-1503(무악동, 무악현대아파트)
장향란
서울 서대문구 연세로 50, 치과대학 구강종양연구소 (신촌동, 연세대학교)
(74) 대리인
특허법인다나

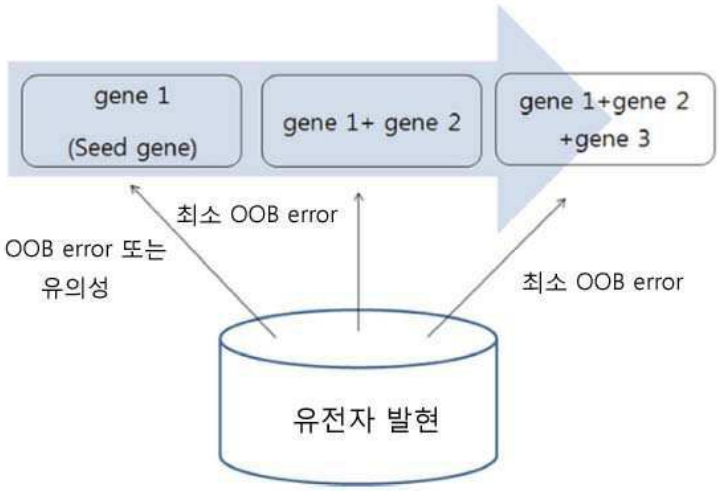
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **림프절 전이 또는 구강암 진단용 유전자 발굴 방법**

(57) 요 약

본 발명은 림프절 전이 또는 구강암 진단용 유전자 발굴 방법에 관한 것으로, 보다 상세하게는, 림프절 전이 또는 구강암 진단에 특이적인 유전자조합을 복합적으로 사용하여 진단 능력을 높일 수 있다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	10041653
부처명	산업통상자원부
연구관리전문기관	연세대학교 산학협력단
연구사업명	바이오의료기기산업융합원천기술개발
연구과제명	위암구강암 예후 예측 및 치료지침을 위한 mRNA-miRNA 복합진단시스템 개발
기 여 율	1/1
주관기관	연세대학교 산학협력단
연구기간	2012.06.01 ~ 2015.05.31

명 세 서

청구범위

청구항 1

구강암 진단 예측을 위한 유전자조합 발굴 방법에 있어서,

MMP1, SOCS3, ACOX1, RUNX2, MTERFD2, FAP, MTAP, LAMB1, RYK, ESD, FCER1A, C10orf128 및 MARVELD3로 구성되는 유전자군에서 선택되는 3종의 유전자를 포함하여 구성되는 적어도 하나 이상의 유전자조합을 포함하는 유전자조합 후보군을 구성하는 단계;

상기 유전자조합 후보군을 구성하는 유전자조합 후보에 대하여, 하기 식 1에 따라 각 유전자의 발현을 계산하고, 상기 유전자조합 후보군의 민감도, 특이도 및 정확도를 검증하는 단계; 및

상기 단계에서 검증된 유전자조합 후보군에 대하여, 구강암 발병 환자군과 정상군에서 상기 유전자조합 후보군의 발현 패턴을 비교하여 오분류율(OOB error)이 낮은 유전자조합을 구강암 진단 예측 능력이 우수한 유전자조합으로 판정하는 단계를 포함하는, 구강암 진단 예측을 위한 유전자조합 발굴 방법:

[식 1]

유전자조합의 발현= $w_1g_1 + w_2g_2 + w_3g_3$

상기 식에서,

w_1, w_2, w_3 는 유전자들의 가중치이며,

g_1, g_2, g_3는 유전자들의 발현양을 나타낸다.

청구항 2

제1항에 있어서,

유전자조합 후보군을 구성하는 단계는 유전자군을 구성하는 개별 유전자에 대하여, 오분류율(OOB error)과 독립적인 t-검정을 실행하여 예측력과 유의성여부를 판단하고, 예측력이 가장 높은 유전자와 유의성이 높은 유전자를 각각 채택하고, 나머지 2종의 유전자들을 임의로 채택하여 3종의 유전자조합을 구성하는 것인, 구강암 진단 예측을 위한 유전자조합 발굴 방법.

청구항 3

제1항 또는 제2항에 있어서,

유전자조합은 MMP1+SOCS3+ACOX1 조합, 또는 MMP1+RUNX2+MTERFD2 조합 중 어느 하나인 구강암 진단 예측을 위한 유전자조합 발굴 방법.

청구항 4

구강암과 전암병소(dysplasia)의 분류 진단 예측을 위한 유전자조합 발굴 방법에 있어서,

PTPRJ, NEK6, SLC44A1, FAM176A 및 KLHL8로 구성되는 유전자군에서 선택되는 3종의 유전자를 포함하여 구성되는 적어도 하나 이상의 유전자조합을 포함하는 유전자조합 후보군을 구성하는 단계;

상기 유전자조합 후보군을 구성하는 유전자조합 후보에 대하여, 하기 식 1에 따라 각 유전자의 발현을 계산하고, 상기 유전자조합 후보군의 민감도, 특이도, 정확도를 검증하는 단계; 및

상기 단계에서 검증된 유전자조합 후보군에 대하여, 구강암 발병 환자군과 전암병소군에서 상기 유전자조합 후보군의 발현 패턴을 비교하여 오분류율(OOB error)이 낮은 유전자조합을 구강암과 전암병소의 분류 진단 예측 능력이 우수한 유전자조합으로 판정하는 단계를 포함하는, 구강암과 전암병소의 분류 진단 예측을 위한 유전자조합 발굴 방법:

[식 1]

유전자조합의 발현= $w_1g_1 + w_2g_2 + w_3g_3$

상기 식에서,

w_1, w_2, w_3 는 유전자들의 가중치이며,

g_1, g_2, g_3는 유전자들의 발현양을 나타낸다.

청구항 5

제4항에 있어서,

유전자조합 후보군을 구성하는 단계는 유전자군을 구성하는 개별 유전자에 대하여, 오분류율(OOB error)과 독립적인 t-검정을 실행하여 예측력과 유의성여부를 판단하고, 예측력이 가장 높은 유전자와 유의성이 높은 유전자를 각각 채택하고, 나머지 2종의 유전자들을 임의로 채택하여 3종의 유전자조합을 구성하는 것인, 구강암과 전암병소의 분류 진단 예측을 위한 유전자조합 발굴 방법.

청구항 6

제4항 또는 제5항에 있어서,

유전자조합은 EVA1A+NEK6+KLHL8 조합을 포함하는 구강암과 전암병소의 분류 진단 예측을 위한 유전자조합 발굴 방법.

청구항 7

구강암으로 진단된 대상의 암세포를 포함하는 생물학적 샘플에서,

APLP2, MFN2, IGF1, PTP4A1, RELA, ZFAND3, SLC7A6, TAF6, HEATR1, CTTN, CDC27, LGMN, RNF139, RAB3GAP2, ZMYND10, ARL3, COMMD8, CEP68 및 L3MBTL로 이루어진 군에서 선택된 2 이상의 유전자 발현 패턴을 측정하는 단계;

상기 단계에서 결정된 유전자 발현 패턴과 림프절 전이 여부에 대해 기설정된 유전자 발현 패턴을 비교하여 림프절 전이 여부를 진단 예측하는 단계를 포함하는, 구강암으로 진단된 대상에서 림프절 전이를 진단 예측하는 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 림프절 전이 또는 구강암 진단에 특이적인 유전자조합을 복합적으로 사용하여 진단 능력을 높일 수 있는 림프절 전이 또는 구강암 진단용 유전자 발굴 방법에 관한 것이다.

배 경 기 술

[0002] 림프절 전이는 구강암의 중요한 예후인자이며, 림프절 전이가 있는 환자들은 림프절 전이가 없는 환자들에 비해 예후가 안 좋은 것으로 알려져 있다. 림프절 전이가 있는 환자들의 5년 생존율은 25-40%인데 반해, 림프절 전이

가 없는 환자들의 생존율은 90% 정도에 이른다.

[0003] 림프절 전이는 치료방법을 결정하는데 중요한 요인인데, 림프절 전이를 정확하게 진단하는 것이 어려워서 많은 환자들이 적절하지 못한 치료를 받기도 한다. 그러나 림프절 전이를 예측하는 바이오마커의 부족으로 부정확한 진단으로 환자들의 예후가 안 좋은 경우가 발생하고 있다. 따라서 림프절 전이를 정확하게 진단할 수 있는 바이오마커의 선별이 필요하다.

[0004] 또한, 유전체학의 추세는 의미 있는 개별적인 유전자 선별보다는 연관성이 있는 유전자 셋(set)을 선별하려는 경향을 보여오고 있다. 더군다나 유전자의 조합이 특이 암환자의 분류를 더 정확히 한다고 알려졌다. 따라서 구강암의 조기진단을 위해서도 유전자조합에 의한 바이오마커 선별이 필요하다.

선행기술문헌

비특허문헌

[0005] (비특허문헌 0001) Chen et al. Biomarkers Prev. 17 (2008) pp. 2152-2162

(비특허문헌 0002) Ye et al. BMC Genomics 9(2008) pp. 69

(비특허문헌 0003) Pyeon et al. Cancer Res, 67(2007) 4605-4619

발명의 내용

해결하려는 과제

[0006] 본 발명의 목적은 구강암 진단에 특이적인 바이오마커를 복합적으로 사용함으로써 유전자조합에 의한 구강암 진단 능력을 높인 구강암 진단용 유전자 발굴 방법을 제공하는 것이다.

[0007] 본 발명의 다른 목적은 구강암과 전암병소 진단에 특이적인 바이오마커를 복합적으로 사용함으로써 유전자조합에 의한 구강암과 전암병소 진단 능력을 높인 구강암과 전암병소 진단용 유전자 발굴 방법을 제공하는 것이다.

[0008] 본 발명의 또 다른 목적은 림프절 전이 진단에 특이적인 바이오마커를 이용하여 구강암에서 림프절 전이를 진단 예측하는 방법을 제공하는 것이다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위하여, 본 발명은 구강암 진단 예측을 위한 유전자조합 발굴 방법에 있어서,

[0010] MMP1, SOCS3, ACOX1, RUNX2, MTERFD2, FAP, MTAP, LAMB1, RYK, ESD, FCER1A, C10orf128 및 MARVELD3로 구성되는 유전자군에서 선택되는 3종의 유전자를 포함하여 구성되는 적어도 하나 이상의 유전자조합을 포함하는 유전자조합 후보군을 구성하는 단계;

[0011] 상기 유전자조합 후보군을 구성하는 유전자조합 후보에 대하여, 하기 식 1에 따라 각 유전자의 발현을 계산하고, 상기 유전자조합 후보군의 민감도, 특이도 및 정확도를 검증하는 단계; 및

[0012] 상기 단계에서 검증된 유전자조합 후보군에 대하여, 구강암 발병 환자군과 정상군에서 상기 유전자조합 후보군의 발현 패턴을 비교하여 오분류율(OOB error)이 낮은 유전자조합을 구강암 진단 예측 능력이 우수한 유전자조합으로 판정하는 단계를 포함하는, 구강암 진단 예측을 위한 유전자조합 발굴 방법을 제공한다:

[0013] [식 1]

[0014] 유전자조합의 발현= $w_1g_1 + w_2g_2 + w_3g_3$

[0015] 상기 식에서,

[0016] w_1, w_2, w_3 는 유전자들의 가중치이며,

[0017] g_1, g_2, g_3는 유전자들의 발현량을 나타낸다.

[0018] 본 발명은 또한 구강암과 전암병소(dysplasia)의 분류 진단 예측을 위한 유전자조합 발굴 방법에 있어서,

[0019] PTPRJ, NEK6, SLC44A1, FAM176A 및 KLHL8로 구성되는 유전자군에서 선택되는 3종의 유전자를 포함하여 구성되는 적어도 하나 이상의 유전자조합을 포함하는 유전자조합 후보군을 구성하는 단계;

[0020] 상기 유전자조합 후보군을 구성하는 유전자조합 후보에 대하여, 상기 식 1에 따라 각 유전자의 발현을 계산하고, 상기 유전자조합 후보군의 민감도, 특이도, 정확도를 검증하는 단계; 및

[0021] 상기 단계에서 검증된 유전자조합 후보군에 대하여, 구강암 발병 환자군과 전암병소군에서 상기 유전자조합 후보군의 발현 패턴을 비교하여 오분류율(OOB error)이 낮은 유전자조합을 구강암과 전암병소의 분류 진단 예측 능력이 우수한 유전자조합으로 판정하는 단계를 포함하는, 구강암과 전암병소의 분류 진단 예측을 위한 유전자조합 발굴 방법을 제공한다.

[0022] 본 발명은 또한 구강암으로 진단된 대상의 암세포를 포함하는 생물학적 샘플에서,

[0023] APLP2, MFN2, IGF1, PTP4A1, RELA, ZFAND3, SLC7A6, TAF6, HEATR1, CTTN, CDC27, LGMN, RNF139, RAB3GAP2, ZMYND10, ARL3, COMMD8, CEP68 및 L3MBTL로 이루어진 군에서 선택된 2 이상의 유전자 발현 패턴을 측정하는 단계;

[0024] 상기 단계에서 결정된 유전자 발현 패턴과 림프절 전이 여부에 대해 기설정된 유전자 발현 패턴을 비교하여 림프절 전이 여부를 진단 예측하는 단계를 포함하는, 구강암으로 진단된 대상에서 림프절 전이를 진단 예측하는 방법을 제공한다.

발명의 효과

[0025] 본 발명은 구강암 진단에 특이적인 바이오마커를 복합적으로 사용하여 유전자조합을 통해 구강암, 구강암과 전암병소의 진단 예측 능력을 높일 수 있다.

[0026] 또한, 본 발명은 구강암 진단 환자에서 바이오마커 발현 패턴을 비교함으로써 림프절 전이 여부를 진단 예측할 수 있다.

도면의 간단한 설명

[0027] 도 1은 구강암 진단 예측을 위한 유전자조합을 생성하기 위한 유전자 셋의 선별 과정을 도시한 것이다.

도 2는 구강암 진단 예측을 위한 유전자의 유의미성과 예측력 간의 관계를 도시한 것이다.

도 3은 구강암 진단 예측을 위한 유전자조합의 구성 유전자 개수에 따른 오분류율을 나타낸 것이다.

도 4는 구강암 진단 예측을 위한 데이터 셋 GSE6791에서 유전자조합 3개의 발현패턴 비교 결과로, (A)는 MMP1, SOCS3, ACOX1 조합, (B)는 FAP, MTAP, C10orf128 조합, (C)는 MMP1, RUNX2, MTERFD2 조합의 결과이다. 여기서, CN은 자궁경부암의 정상군, CC는 자궁경부암의 암환자, HNN은 두경부암의 정상군, HCC은 두경부암의 암환자를 의미한다.

도 5는 구강암 진단 예측을 위한 조합된 유전자를 이용한 전암병소단계와 암환자간의 발현패턴 비교 결과이다.

도 6은 림프절 전이 진단 예측을 위한 공용 데이터베이스에서 확보된 데이터 셋의 처리 과정을 도시한 것이다.

도 7은 림프절 전이 여부에 따른 유전자 발현 패턴 비교 결과이다.

도 8은 림프절 전이 진단 예측을 위한 2개의 데이터 셋에서 선별된 유전자 셋의 발현패턴을 비교한 결과이다.

도 9는 림프절 전이 진단 예측을 위한 병합된 데이터 셋에서 선별한 유전자를 검증용 데이터에서 발현패턴을 확

인한 결과이다.

도 10은 림프절 전이 진단 예측을 위한 각 데이터 셋에서 선별된 유전자의 분류정확성을 비교하기 위한 OOB error 비교 결과이다.

도 11은 림프절 전이 진단 예측을 위한 각 데이터 셋에서 선별된 유전자 셋의 다른 데이터 셋에서의 일치도 비교 결과이다.

도 12는 림프절 전이 진단 예측을 위한 병합된 데이터 셋에서 선별된 유전자들의 발현패턴을 보여주는 결과이다.

도 13은 림프절 전이 진단 예측을 위한 MFN2와 CTTN, ZFAND3와 SLC7A6과의 상관관계, MAPK7과 ARL3, LGMN과의 상관관계를 보여주는 유전자 발현패턴이다.

도 14 내지 17은 림프절 전이 진단 예측을 위한 선별된 유전자들의 전체 염색체 상에서의 분포된 발현패턴을 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0028] 이하, 본 발명의 구성을 구체적으로 설명한다.

[0029] 본 발명은 구강암 진단 예측을 위한 유전자조합 발굴 방법에 있어서,

[0030] MMP1, SOCS3, ACOX1, RUNX2, MTERFD2, FAP, MTAP, LAMB1, RYK, ESD, FCER1A, C10orf128 및 MARVELD3로 구성되는 유전자군에서 선택되는 3종의 유전자를 포함하여 구성되는 적어도 하나 이상의 유전자조합을 포함하는 유전자조합 후보군을 구성하는 단계;

[0031] 상기 유전자조합 후보군을 구성하는 유전자조합 후보에 대하여, 하기 식 1에 따라 각 유전자의 발현을 계산하고, 상기 유전자조합 후보군의 민감도, 특이도 및 정확도를 검증하는 단계; 및

[0032] 상기 단계에서 검증된 유전자조합 후보군에 대하여, 구강암 발병 환자군과 정상군에서 상기 유전자조합 후보군의 발현 패턴을 비교하여 오분류율(OOB error)이 낮은 유전자조합을 구강암 진단 예측 능력이 우수한 유전자조합으로 판정하는 단계를 포함하는, 구강암 진단 예측을 위한 유전자조합 발굴 방법에 관한 것이다:

[0033] [식 1]

[0034] 유전자조합의 발현= $w_1g_1 + w_2g_2 + w_3g_3$

[0035] 상기 식에서,

[0036] w_1, w_2, w_3 는 유전자들의 가중치이며,

[0037] g_1, g_2, g_3는 유전자들의 발현량을 나타낸다.

[0039] 본 발명은 또한 구강암과 전암병소(dysplasia)의 분류 진단 예측을 위한 유전자조합 발굴 방법에 있어서,

[0040] PTPRJ, NEK6, SLC44A1, FAM176A 및 KLHL8로 구성되는 유전자군에서 선택되는 3종의 유전자를 포함하여 구성되는 적어도 하나 이상의 유전자조합을 포함하는 유전자조합 후보군을 구성하는 단계;

[0041] 상기 유전자조합 후보군을 구성하는 유전자조합 후보에 대하여, 상기 식 1에 따라 각 유전자의 발현을 계산하고, 상기 유전자조합 후보군의 민감도, 특이도, 정확도를 검증하는 단계; 및

[0042] 상기 단계에서 검증된 유전자조합 후보군에 대하여, 구강암 발병 환자군과 전암병소군에서 상기 유전자조합 후보군의 발현 패턴을 비교하여 오분류율(OOB error)이 낮은 유전자조합을 구강암과 전암병소의 분류 진단 예측 능력이 우수한 유전자조합으로 판정하는 단계를 포함하는, 구강암과 전암병소의 분류 진단 예측을 위한 유전자조합 발굴 방법을 제공한다.

[0043] 본 발명의 구강암 진단 예측을 위한 유전자조합 발굴 방법은 구강암의 조기진단의 진단 능력을 높이기 위한 것으로, 우선, 구강암의 판별력을 향상시킬 수 있는 조합형 바이오마커를 선별하였다. 가장 정확한 분류력을 유지하기 위해서 몇 개의 유전자를 조합하는 것이 가장 적절한지 결정하는 과정을 진행하고, 정해진 유전자 개수만

큼 유의한 유전자를 선별하는 단계에서, 최초의 유전자는 가장 유의한 유전자를 선택하였고 추가되는 유전자는 조합하였을 때 가장 분류의 오류를 작게 하는 유전자를 선별하여 유전자 셋을 구성하였다. 미리 정해진 개수의 유전자 셋이 확보되면, 유전자조합을 정의하기 위한 각 유전자의 가중치를 계산하였다. 이를 위하여 주성분 분석(PCA, Principal Component Analysis)을 사용하였다. 이 과정을 진행하기 위해 공용 마이크로어레이 데이터베이스로부터 세 개의 데이터 셋을 확보하여 사용하였다. 이들은 유전자조합의 선별과 선별된 유전자조합의 검증을 위하여 사용하였다. 검증을 위한 척도로는 오분류율을 나타내는 OOB error(out of bag error)를 사용하였다. 개별적으로 유의미한 유전자만이 분류의 정확성이 좋은 것은 아니므로, 본 발명자들은 유의미한 유전자와 유의미하지 않은 유전자를 조합함으로써 높은 분류의 정확성을 보여주는 유전자조합을 찾을 수 있었다.

[0044] 본 발명의 구강암 진단 예측을 위한 유전자조합 발굴 방법을 단계별로 설명하면 다음과 같다.

[0045] 제1단계는 공용 마이크로어레이 데이터베이스에서 3개의 데이터 셋을 확보하고, 두 개의 데이터 셋은 유전자조합을 발굴하는데 사용하고, 나머지 데이터 셋은 발굴한 유전자조합을 검증하는데 사용한다.

[0046] 유전자조합을 발굴하기 전에 조합할 유전자의 적정 수를 결정하기 위해 1 내지 5개의 유전자를 임의로 샘플링하여 다음과 같이 오분류율(OOB error)을 계산한다.

[0047] (1) 동일한 데이터 셋에서 반복을 허용하여 n 개의 bootstrap samples {$B1$,$B2$,...,Bn} 을 추출한다.

[0048] (2) 추출된 표본 중 Bk 를 사용하여 Tree classifier Tk 를 생성하고 이를 사용하여 Bk 이외의 표본들을 prediction 한다. 이것을 out-of-bag(OOB) 표본(samples) 이라고 한다.

[0049] (3) 최종적인 예측력은 모든 bootstrap samples에 대한 out-of-bag estimators의 평균으로 하고, 이를 overall classification error(OOB error) 라고 한다.

[0050] 본 발명의 일 구현예에 따르면, 최적의 유전자조합 수는 3개인 것으로 확인되었다.

[0051] 유전자조합의 적정 수가 결정되면, 주성분분석(Principal Component Analysis, PCA)에 의해 각 유전자의 가중치를 계산한다. PCA는 많은 요인을 포함하는 복잡한 데이터 셋에서 타당한 정보를 찾아내기 위해 사용할 수 있는 비모수적인 통계적 분석 방법으로, 상대적으로 적은 데이터 셋에서도 사용이 가능하다. 또한, PCA는 연관성이 있는 요인들을 서로 독립적인 요인으로 바꾸기 위하여 직교변환을 사용하는 방법으로, 직교변환에 의해 새로 생성된 독립적인 요인을 주성분(Principal component, PC)라고 하며, 이것이 조합된 유전자가 된다.

[0052] 조합된 유전자의 발현은 상기 식 1과 같다.

[0053] 조합된 유전자의 효율성은 민감도(sensitivity), 특이도(specificity), 정ㅎ확도(accuracy)에 의해 측정하고, 이는 프로그램 R(version 2.13.0)으로 작성할 수 있다.

[0054] 유전자의 조합은 개별 유전자들에 대해 예측력과 유의성 여부를 확인하여 결정하고, 예측력은 OOB error로, 유의성 여부는 독립적인 t-검정을 실행하여 확인한다.

[0055] 확인 결과, 예측력이 가장 높은 유전자와 유의성 여부가 가장 높은 유전자를 제1유전자로 선정하고, 나머지 2개의 유전자에 대해서는 임의로 선택하여 3종의 유전자로 구성된 유전자조합 후보를 결정한다.

[0056] 각 유전자조합 후보들은 상술한 데이터베이스에서 확보한 검증용 데이터 셋에 대하여 민감도, 특이도, 정확도를 계산하여, 구강암 환자군와 정상군을, 전암병소와 구강암을 분류할 수 있는 유전자조합을 검증한다.

[0057] 상기에서 검증된 유전자조합 후보군에 대하여 구강암 발병 환자군과 정상군, 구강암 발병 환자군과 전암병소군에서, 상기 유전자조합 후보군의 발현 패턴을 비교하여 오분류율(OOB error)이 낮은 유전자조합을 구강암 진단 예측 능력이 우수한 유전자조합 또는 구강암과 전암병소의 분류 진단 예측 능력이 우수한 유전자조합이라 판정한다.

[0058] 본 발명에 따르면, 구강암 진단 예측 능력이 우수한 유전자조합은 MMP1+SOCS3+ACOX1 조합, 또는 MMP1+RUNX2+MTERFD2 조합일 수 있다.

[0059] 본 발명에 따르면, 또는 구강암과 전암병소의 분류 진단 예측 능력이 우수한 유전자조합은 EVA1A+NEK6+KLHL8 조합일 수 있다.

[0060] 본 발명은 또한 구강암으로 진단된 대상의 암세포를 포함하는 생물학적 샘플에서,

[0061] APLP2, MFN2, IGF1, PTP4A1, RELA, ZFAND3, SLC7A6, TAF6, HEATR1, CTTN, CDC27, LGMN, RNF139, RAB3GAP2, ZMYND10, ARL3, COMMD8, CEP68 및 L3MBTL로 이루어진 군에서 선택된 2 이상의 유전자 발현 패턴을 측정하는 단계;

[0062] 상기 단계에서 결정된 유전자 발현 패턴과 림프절 전이 여부에 대해 기설정된 유전자 발현 패턴을 비교하여 림프절 전이 여부를 진단 예측하는 단계를 포함하는, 구강암으로 진단된 대상에서 림프절 전이를 진단 예측하는 방법에 관한 것이다.

[0063] 본 발명의 구강암으로 진단된 대상에서 림프절 전이를 진단 예측하는 방법에 따르면, 림프절 전이를 정확하게 진단하기 위해, 공용데이터베이스에서 림프절 전이와 관련된 마이크로어레이 데이터를 확보하여 림프절 전이 여부를 예측하는데 활용할 수 있는 유전자를 선별하였다. 유전자 선별을 하기 위한 충분한 표본수를 확보하기 위하여 확보한 데이터 셋을 병합하는 방법도 사용하였다. 선별된 유전자 셋은 공용데이터베이스에서 확보한 검증용 데이터 셋을 사용하여 검증하였다. 검증에는 OOB error를 사용하였다. 병합한 데이터 셋에서 선별한 림프절 전이 관련 유전자 셋은 검증용 데이터 셋에서 림프절 전이 여부를 더욱 정확하게 분류함을 확인하였다.

[0064] 본 발명은 기존의 림프절 전이 예측을 위한 연구 결과의 일치도가 상당히 낮았다. 이것은 각 연구의 결과는 실험조건에 강하게 의존적이기 때문이다. 본 발명은 공용 마이크로어레이 데이터베이스에서 동일한 목적으로 실험된 여러 개의 마이크로어레이 데이터를 확보하고, 이를 병합한 데이터를 사용함으로써 좀 더 안정적인 유전자 셋을 선별할 수 있었다.

[0065] 본 발명의 구강암으로 진단된 대상에서 림프절 전이를 진단 예측하는 방법을 단계별로 설명하면 다음과 같다.

[0066] 우선, 공용 마이크로어레이 데이터베이스에서 3개의 데이터 셋을 확보하고 림프절 전이와 관련된 유전자를 선별하고, 선별된 유전자를 검증하는데 사용하였다.

[0067] 독립적인 두 개의 데이터 셋으로부터 각각 림프절 전이와 관련된 유전자를 선별하기 위하여 Mann-Whitney U test를 사용하고, , 이들을 병합한 후 림프절 전이 관련 유전자를 선별하기 위하여 유전자 발현정보를 범주화한 값을 사용한다. 선별된 유전자 셋을 검증하기 위한 척도로는 OOB error와 다른 선별된 유전자들의 다른 셋에서의 일치도를 사용한다.

[0068] 본 발명의 일 구현예에 따르면, 2개의 데이터 셋에서 선별된 림프절 전이 관련 유전자들의 발현 패턴을 비교한 결과, 공통적으로 선별할 수 있는 유전자군이 발견되지 않아, 데이터 셋을 조합하여 유의한 유전자를 선별하였다.

[0069] 병합한 데이터 셋에서 림프절 전이 관련 유전자를 선별하고, 발현 패턴은 클러스터와 TreeView를 사용하여 탐색한다.

[0070] 선별된 유전자의 분류 정확성은 OOB error를 사용하여 비교하여 분류의 정확도와 안정성을 확인한다.

[0071] 상기에서 정확도와 안정성이 우수한 유전자를 상술한 유전자 셋에서 OOB error를 사용하여 오류율이 낮은 유전자를 선정한다.

[0072] 상기에서 선별된 유전자들은 염색체상에서 발현 패턴을 탐색하고, 상술한 데이터 셋에서 검증한다.

[0073] 본 발명에 따르면, APLP2, MFN2, IGF1, PTP4A1, RELA, ZFAND3, SLC7A6, TAF6, HEATR1, CTTN, CDC27, LGMN, RNF139, RAB3GAP2, ZMYND10, ARL3, COMMD8, CEP68 및 L3MBTL로 이루어진 군에서 선택된 유전자들의 발현 패턴을 비교하여 구강암에서 림프절 전이 여부를 진단 예측할 수 있다.

[0074] 이하, 본 발명을 실시예에 의해 상세히 설명한다. 단, 하기 실시예는 본 발명을 예시하는 것일 뿐, 본 발명의 내용이 하기 실시예에 한정되는 것은 아니다.

[0075] **<실시예 1> 구강암 진단**

[0076] (데이터 확보)

[0077] 본 연구를 위하여 공용 마이크로어레이 데이터베이스로부터 세 개의 데이터 셋을 확보하였다. 사용한 데이터 셋은 표 1에 정리하였다. GSE30784 는 Chen *et al.* (*Biomarkers Prev.* 17 (2008) pp. 2152-2162)에 의해 발표되

었고, 이것은 유전자조합을 찾는데 사용하였다. 다른 두 개의 데이터 셋은 발굴한 유전자조합을 검증하는데 사용되었으며, Ye *et al.*(GSE9844)(*BMC Genomics* 9(2008) pp. 69)과 Pyeon *et al.*(GSE6791) (Cancer Res, 67(2007) 4605-4619)에 의해 발표된 데이터이다.

표 1

유전자조합을 찾기 위하여 사용한 데이터 셋의 요약

	GSE30784 N=229	GSE9844 N=38	GSE6791 N=75
Age	19-39 24 40-49 42 50-59 67 60-68 96	Mean 56.40(std* 12.22) Median 57 Range 37~82	Mean 54.41 Range 18~88
Gender	F 67 M 162	F 9 M 29	F 48 M 34 NA 3
	Cancer 167 Normal 45 Dysplasia 17	OSCC 26 Normal 12	Cervical cancer (CC) 20 Cervical normal (CN) 8 HN cancer (HNC) 42 HN normal (HNN) 5
N stage		Negative 15 Positive 11	
Platform	GPL570 [HG-133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array 54675 probes	GPL570 [HG-133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array 54675 probes	GPL570 [HG-133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array 54675 probes

*std: standard deviation

[0078]

[0079] (통계적 방법)

[0080] 유전자조합을 발굴하기 전에 조합할 유전자의 적절한 개수를 결정하기 위하여 1개의 유전자부터 5개 유전자까지 임의로 10,000번을 샘플링하여 오분류율(OOB error)을 계산하였다. 10,000번 반복하여 유전자를 샘플링한 것은 selection bias 를 최소화하기 위한 것이다. OOB error의 계산은 다음의 단계로 이루어졌다.

[0081] (1) 동일한 데이터 셋에서 반복을 허용하여 n 개의 bootstrap samples $\{B1, B2, \ldots, Bn\}$ 을 추출한다.

[0082] (2) 추출된 표본 중 Bk 를 사용하여 Tree classifier Tk 를 생성하고 이를 사용하여 Bk 이외의 표본들을 prediction 한다. 이것을 out-of-bag(OOB) 표본(samples) 이라고 한다.

[0083] (3) 최종적인 예측력은 모든 bootstrap samples 에 대한 out-of-bag estimators 의 평균으로 하고, 이를 overall classification error(OOB error) 라고 한다.

[0084] 최적의 유전자 개수가 정해지고 유전자 셋이 선별되면, 주성분분석(Principal Component Analysis, PCA)에 의해 각 유전자의 가중치를 계산하였다. PCA는 많은 요인을 포함하는 복잡한 데이터 셋에서 타당한 정보를 찾아내기 위해 사용할 수 있는 비모수적인 통계적 분석 방법이다, 상대적으로 적은 데이터 셋에서도 사용이 가능하다.

[0085] PCA는 연관성이 있는 요인들을 서로 독립적인 요인으로 바꾸기 위하여 직교변환을 사용하는 방법이다. 직교변환에 의해 새로 생성된 독립적인 요인을 주성분(Principal component, PC)라고 하며, 이것이 조합된 유전자가 된다. 조합된 유전자의 발현량은 선택된 유전자들에 의한 함수식의 형태로 계산이 되면 공식은 다음과 같다:

[0086] [식 1]

[0087] PC 또는 유전자조합의 발현 = $w_1g_1 + w_2g_2 + w_3g_3$

[0088] 상기 식에서,

[0089] w_1, w_2, w_3 는 유전자들의 가중치이며,

[0090] g_1, g_2, g_3는 유전자들의 발현양을 나타낸다.

[0091] 조합된 유전자의 효율성은 민감도(sensitivity), 특이도(specificity), 정확도((accuracy)에 의해 측정하였으며, 모든 프로그램은 R(version 2.13.0)으로 작성하였다.

[0092] 도 1은 유전자조합을 생성하기 위한 유전자 셋의 선별 과정을 도시하고 있다.

[0093] (유전자의 유의성과 예측력의 관계)

[0094] 유의미한 유전자가 예측력이 높은지를 알아보기 위하여 데이터 셋의 모든 유전자의 유의성과 예측력을 계산하여 두 값 간의 관계를 살펴보았다. 모든 유전자 셋 중에서 임의로 1000개의 유전자를 선별하여 같은 작업을 실행하였다.

[0095] 도 2에 나타난 바와 같이, 전체 유전자 셋과 임의로 선별한 1000개의 유전자 셋에서 유의미한 모든 유전자가 예측력이 높은 것은 아니라는 것을 확인했다. 전체적으로 볼 때, 유전자의 유의성과 예측력은 연관성이 없는 것으로 나타났다. 그러므로, 예측력을 최대화하는 유전자조합을 유의한 유전자와 무의미한 유전자를 포함하여 선정할 필요성이 있다.

[0096] (유전자조합을 구성하기 위한 적절한 유전자 개수)

[0097] 유전자조합에 적절한 유전자 개수를 결정하기 위하여 유전자 개수를 1개부터 5개까지, 각 경우에 대하여 10000번씩 반복추출한 유전자 셋의 오분류율(OOB error) 를 계산하여 가장 작은 오류율을 보여주는 수를 적절한 유전자 개수로 정하였다.

[0098] 도 3에 나타난 바와 같이, 유전자의 개수를 증가시킬수록 오분류율은 감소하는 것을 확인할 수 있었으며, 유전자 개수가 3보다 큰 경우에는 오분류율은 조금 감소하나 변이가 증가하는 경향을 보였다. 따라서 적절한 유전자의 개수를 3으로 결정하였다.

[0099] (유전자 셋을 구성하는 초기 유전자 선별: 유의성과 예측도를 고려함)

[0100] 초기 유전자(seed genes)는 두 가지 기준 -유의성과 예측력-에 의해 선별하였다. 유의성과 예측력에 의해 선별된 20개의 유전자는 표 2 및 3에 정리하였다.

표 2

OOB error rate 및 유의성 여부에 따라 선별된 유전자들(정상 대 구강암)

Selected genes by OOB error rate		*Selected genes by significance*	
Gene name	**description**	**Gene name**	**description**
MMP1	matrix metallopeptidase 1	FAP	FAP fibroblast activation protein, alpha
PLAUR	plasminogen activator, urokinase receptor	ADAM12	ADAM12 ADAM metallopeptidase domain 12
COL1A1	collagen, type I, alpha 1	MMP10	MMP10 matrix metallopeptidase 10 (stromelysin 2)
PLAU	plasminogen activator, urokinase	NOX4	NOX4 NADPH oxidase 4
TMEM184B		SH3BGRL2	SH3BGRL2 SH3 domain binding glutamic acid-rich protein like 2
COL1A1	collagen, type I, alpha 1	HOXC4	HOXC4 homeobox C4
COL1A2	collagen, type I, alpha 2	PADI1	peptidyl arginine deiminase, type 1
COL1A1	collagen, type I, alpha 1	SCIN	scinderin
NCOA1	nuclear receptor coactivator 1	PXDN	peroxidasin homolog (Drosophila)
CAB39L	calcium binding protein 39-like	NUCB2	nucleobindin 2
RFK	riboflavin kinase	CRISP3	cysteine-rich secretory protein 3
SERPINH1	serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1)	PTHLH	parathyroid hormone-like hormone
COL4A2	collagen, type IV, alpha 2	C22orf32	
COL3A1	collagen, type III, alpha 1	CGNL1	cingulin-like 1
LOC441178		LAMC2	laminin, gamma 2
SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	COL5A2	collagen, type V, alpha 2
SCEL	sciellin	C22orf32	
BLNK	B-cell linker	GALNT12	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 12 (GalNAc-T12)
KRP4	cyclin-dependent kinase inhibitor 4	ANKRD20A1	ankyrin repeat domain 20 family, member A1
BRP44L	brain protein 44-like	HMGA2	high mobility group AT-hook 2

[0101]

표 3

OOB error rate 및 유의성 여부에 따라 선별된 유전자들(전암병소 대 구강암)

Selected genes by OOB error rate		*Selected genes by significance*	
Gene name	**description**	**Gene name**	**description**
PTPRJ	protein tyrosine phosphatase, receptor type	FAM176A	
TNXB	tenascin XB	MMP11	matrix metallopeptidase 11 (stromelysin 3)
MAST2	microtubule associated serine/threonine kinase 2	ADORA3	adenosine A3 receptor
CGNL1	cingulin-like 1	PPP4R4	
PCM1	pericentriolar material 1	EIF5A2	eukaryotic translation initiation factor 5A2
TBP	box binding protein	ZNF114	
GSTO1	glutathione S-transferase omega 1	CST1	cystatin SN
MOCS2	molybdenum cofactor synthesis 2	ZNF114	
RANGAP1	Ran GTPase activating protein 1	SLC6A15	solute carrier family 6 (neutral amino acid transporter), member 15
KCMF1	potassium channel modulatory factor 1	COL22A1	collagen, type XXII, alpha 1
BNC2	basonuclin 2	NRP2	neuropilin 2
SMEK1	MEK homolog 1, suppressor of mek1 (Dictyostelium)	MMP13	matrix metallopeptidase 13 (collagenase 3)
ZBTB44		OLR1	oxidized low density lipoprotein (lectin-like) receptor 1
PRUNE	prune homolog (Drosophila)	INHBA	inhibin, beta A
GK5	glycerol kinase 5 (putative)	NEBL	
LOC284551		CAPNS2	calpain, small subunit 2
COPS3	COP9 constitutive photomorphogenic homolog subunit 3 (Arabidopsis)	NEBL	
THAP9		TRAF4	TNF receptor-associated factor 4
ADORA2B	adenosine A2b receptor	GRP	gastrin-releasing peptide
LRRC48		TNFSF4	tumor necrosis factor (ligand) superfamily, member 4

[0102]

[0103] 예측력은 OOB error로 대신하였고, 유의성여부는 독립적인 t-검정을 실행하였다. 이 단계에서는 모든 유전자를 대상으로 분석하였으며, 암환자와 정상군을 분류하는 경우에는 MMP1 은 가장 예측력이 높은 유전자였고 FAP는 가장 유의성이 높은 유전자였다.

[0104] 전암단계의 환자와 암환자를 분류하는 경우에는 PTPRJ 와 EVA1A가 각각 예측력이 높고, 유의성이 높은 유전자로 선별되었다. 이 두 개의 초기 유전자(seed gene)로부터 시작하여 유전자조합을 정의하였다.

[0105] (유전자조합의 발굴 및 검증)

[0106] 초기 유전자로부터 유전자조합을 정의하고 PCA에 의하여 각 유전자의 가중치를 계산하였다. 각 유전자조합의 민감도, 특이도, 정확도를 계산하여 표 4에 정리하였다.

표 4

정의한 유전자조합을 데이터 셋 GSE30784을 사용하여 검증한 결과

Classification group	*Gene name*	*Weights by PCA*	*sensitivity*	*specificity*	*accuracy*
Normal & OSCC	MMP1, SOCS3, ACOX1	0.916, 0.839, -0.859	0.982	0.933	0.972
	MMP1, RUNX2, MTERFD2	-0.907, -0.811, 0.811	0.988	0.956	0.981
	FAP, MTAP, LAMB1	0.951, 0.309, 0.941	0.976	0.911	0.962
	FAP, MTAP, LAMB1	-0.947, -0.306, -0.935	0.970	0.889	0.953
	FAP, MTAP, RYK	0.810, 0.441, 0.703	0.964	0.867	0.953
	FAP, MTAP, ESD	-0.759, -0.771, -0.028	0.946	0.8	0.915
	FAP, MTAP, FCER1A	-0.855, -0.404, 0.830	0.982	0.933	0.972
	FAP, MTAP, C10orf128	0.816, 0.426, -0.700	0.982	0.933	0.971
	FAP, MTAP, MARVELD3	0.727, 0.736, -0.354	0.940	0.778	0.906
Dysplasia & OSCC	PTPRJ, NEK6, SLC44A1	-0.732, -0.746, 0.553	0.952	0.529	0.913
	EVA1A, NEK6, KLHL8	0.867, 0.889, 0.293	0.985	0.588	0.924

[0107]

[0108] MMP1, RUNX2 과 MTERFD2으로 구성된 유전자조합이 암환자군과 정상군을 분류하는데 가장 효과가 좋았다(accuracy=0.981). EVA1A, NEK6 과 KLHL8으로 구성된 유전자조합은 전암병소와 암을 분류하는데 가장 효과가 좋은 유전자조합이었다(accuracy=0.924). 전암병소와 암을 분류하는 경우에는. 전암병소를 암으로 분류하려는 경

향이 나타나서 정확도는 높았지만 특이도는 낮게 나타났다.

[0109] 선별된 유전자조합의 기존의 연구 결과와 비교하였다 (Chen et al., 2008; Ye et al., 2008).

표 5

조합된 유전자 셋의 효과를 기존 연구와 비교한 결과

[0110]

Data	*Gene sets identified in previous study*	*Weights from logistic regression*	*sensitivity*	*specificity*	*accuracy*
GSE30784	LAMC2, COL4A1	7.8739, 7.6269	0.976	0.911	0.962
	COL1A1, PADI1	2.4377, -2.8841	0.976	0.911	0.962
GSE9844	LAMC2, COL4A1	7.8739, 7.6269	0.923	0.833	0.895
	COL1A1, PADI1	2.4377, -2.8841	0.846	0.667	0.789
Data	*Gene sets identified in our study*	*Weights by PCA*	*sensitivity*	*specificity*	*accuracy*
GSE30784	MMP1, SOCS3, ACOX1	0.916, 0.839, -0.859	0.982	0.933	0.972
	MMP1, RUNX2, MTERFD2	-0.907, -0.811, 0.811	0.988	0.956	0.981
	FAP, MTAP, C10orf128	0.816, 0.426, -0.700	0.982	0.933	0.971
GSE9844	MMP1, SOCS3, ACOX1	0.916, 0.839, -0.859	0.9612	0.917	0.947
	MMP1, RUNX2, MTERFD2	-0.907, -0.811, 0.811	0.9612	0.917	0.947
	FAP, MTAP, C10orf128	0.816, 0.426, -0.700	0.9612	0.917	0.947

[0111] 조합된 유전자 셋의 효과를 Chen et al.(2008) 연구 결과와 비교하기 위하여 검증용 데이터 셋 GSE30784 와 GSE9844를 사용하였다. 본 연구에서 선별한 유전자조합들이 두 데이터 셋 모두에서 민감도, 특이도와 정확도가 더 높았다.

[0112] 또한, 조합된 유전자 셋의 발현패턴을 GSE6791에서 탐색한 결과. 도 4에 나타난 바와 같이, 3개의 유전자조합 모두 자궁경부암의 암환자와 정상군, 두경부암의 암환자와 정상군 간에 명확한 발현의 차이를 확인하였다.

[0113] 검증용 데이터 셋 GSE30784에서 전암병소와 암환자군 간의 조합된 유전자 발현패턴을 살펴보았다.

[0114] 도 5에 나타난 바와 같이, 가장 유의하거나 정확성이 높은 유전자 1개만을 사용하는 것보다 유전자조합을 사용하는 것이 분류의 정확성이 높게 나타났다.

[0115] 1개의 유의하거나 정확성이 높은 유전자는 전암단계와 암환자를 정확하게 분류하지 못하였다. 유의성이 높은 유전자 대부분이 50% 정도의 예측력밖에 보여주지 않았다. 그러나 2개 이상의 유전자를 조합하면서 예측력은 크게 향상되었다.

[0116] 검증용으로 사용한 데이터 셋에서 전암병소와 암환자의 표본수가 심하게 불균형이어서 전암병소의 표본수에 맞춰서 100번 반복추출하여 매번 OOB error를 계산하고, 이 값들의 분포를 그림에 표시하였다.

[0117] **<실시예 2> 림프절 전이 진단**

[0118] (데이터 준비)

[0119] 공용 마이크로어레이 데이터베이스로부터 데이터를 확보하였고, 이들은 림프절 전이와 관련된 유전자를 선별하고, 선별된 유전자를 검증하는데 사용되었다. 확보된 데이터 셋들은 다른 실험환경에서 마이크로어레이 실험에 의한 것으로 유전자 발현량의 스케일이 달랐다.

표 6

본 연구에 사용된 데이터 셋의 요약

[0120]

Data set	Platform	Number of experiments	Range of expressions	Original size of data set (number of genes)
GSE3524 (Torunes, et al., 2004)	affy U133A array	7N+, 7N-	0.00 – 000.00	4N, 16T (14119 genes)
GSE2280 (O'Donnell, et al., 2005)	affy U133A array	14N+, 8N-	0 - 00000.00	5N, 22T (22283 genes)
GSE10121 (Sticht, et al., 2008)	DKFZ/Operon Human Oligo Set v4	23N+, 12N-	-0.000 ~ 0.000	6N, 35T (33485 genes)

[0121] 검증용으로 사용한 데이터 셋은 결측치가 포함이 되어 있어서 SKNN (sequential k nearest neighbor) 방법에 의하여 결측치를 추정해서 사용하였다.

[0122] (통계적 방법)

[0123] 독립적인 2개의 데이터 셋으로부터 각각 림프절 전이와 관련된 유전자를 선별하기 위하여 Mann-Whitney U test를 사용하였고, 이들을 병합한 후 림프절 전이 관련 유전자를 선별하기 위하여 유전자 발현정보를 범주화한 값을 사용하였다. 선별된 유전자 셋을 검증하기 위한 척도로는 OOB error 와 다른 선별된 유전자들의 다른 셋에서의 일치도를 사용하였다. OOB error의 계산은 상기 실시예 1과 같이 수행하였다.

[0124] 도 6에는 공용 데이터베이스에서 확보된 데이터 셋의 처리 가정을 도시하였다.

[0125] (림프절 전이 여부에 따른 유전자 발현 패턴 비교)

[0126] 림프절 전이가 있는 환자의 유전자 발현량이 림프절 전이가 없는 환자에 비해 변이가 크게 나타난다(도 7).

[0127] (2개의 데이터 셋에서 선별한 유전자 셋의 발현패턴)

[0128] 2개의 데이터 셋에서 림프절 전이와 관련된 유전자를 선별하고 발현패턴을 비교하였다. 유전자 선별은 Mann-Whitney U test를 사용하고, 그림으로 표현하기 위하여 무료 프로그램인 클러스터와 TreeView를 사용하여 실행하였다.

[0129] 도 8에 나타난 바와 같이, 각 데이터 셋에서 선별된 유전자에 의해 서로 다른 실험군은 잘 분리되는 경향을 보였다. 2개의 데이터 셋에서 공통으로 선별된 유전자는 FLJ12529였으며, 이것은 N+군에서 과발현하는 양상을 보였다.

[0130] 각 데이터 셋에서 선별한 유전자 셋의 공통부분이 거의 일치하는 않는다는 것은 유의한 유전자 선별이 사용하는 데이터 셋에 의존적임을 나타낸다. 따라서 좀 더 안정적인 유전자 셋을 선별하기 위하여 데이터 셋을 조합한 후 유의한 유전자를 선별하였다.

[0131] (병합된 데이터 셋에서 선별한 유전자를 검증용 데이터에서 발현패턴 확인)

[0132] 병합한 데이터에서 림프절 관련 유전자를 선별하고, 발현패턴을 클러스터(cluster)와 TreeView를 사용하여 탐색했다.

[0133] 도 9에 나타난 바와 같이, 선별된 유전자 셋에 의해 N+군의 일부분은 잘 분류되었으나, 다른 부분은 N-군과 섞이는 경향을 보였다. 이것은 N+군은 발현량의 변동이 크게 나타났으므로 일부분은 N- 군과 유사한 발현 패턴을 갖는 것으로 해석된다.

[0134] 데이터 셋 A 와 데이터 셋 A, B를 병합한 데이터 셋 AB에서 선별한 유전자 셋의 공통 유전자는 PTPN14, HEATR1, GCS1, VPS24, LANCL1, LRP12, F8 였으며, 데이터 셋 B 와 병합된 데이터 셋에서 선별한 유전자 셋의 공통 유전

자는 CEP68, ZMYND10, TIGD6, PTP4A1, CHRNA10, COL2A1, GPR68, RAB11FIP3, SLC7A6, CAMTA2, MRM1, HELZ, FBXL12, BBC3였다. 병합한 데이터 셋 AB에서 선별된 유전자 목록은 표 7과 8에 정리하였다.

표 7

Symbol	Name
POLS	Data not found
HEXIM1	hexamethylene bis-acetamide inducible 1
FHL2	four and a half LIM domains 2
SIAH1	seven in absentia homolog 1 (Drosophila)
GLS	glutaminase
OCRL	oculocerebrorenal syndrome of Lowe
TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa
SLC7A6	solute carrier family 7 (cationic amino acid transporter, y+ system), member 6
WBP4	WW domain binding protein 4 (formin binding protein 21)
TERF2	telomeric repeat binding factor 2
HELZ	helicase with zinc finger
RAB11FIP3	RAB11 family interacting protein 3 (class II)
RNGTT	RNA guanylyltransferase and 5'-phosphatase
SLC26A1	solute carrier family 26 (sulfate transporter), member 1
SERPINI1	serpin peptidase inhibitor, clade I (neuroserpin), member 1
ATP2B4	ATPase, Ca++ transporting, plasma membrane 4
PTPN14	protein tyrosine phosphatase, non-receptor type 14
AIFM1	apoptosis-inducing factor, mitochondrion-associated, 1
ACAN	aggrecan
MAP2K6	mitogen-activated protein kinase kinase 6
F8	coagulation factor VIII, procoagulant component
ABI2	abl-interactor 2
KRT85	keratin 85
CEP68	centrosomal protein 68kDa
TGDS	TDP-glucose 4,6-dehydratase
APLP2	amyloid beta (A4) precursor-like protein 2
SUPT6H	suppressor of Ty 6 homolog (S. cerevisiae)
ERI3	ERI1 exoribonuclease family member 3
SEC22B	SEC22 vesicle trafficking protein homolog B (S. cerevisiae) (gene/pseudogene)
RGS16	regulator of G-protein signaling 16
RNF139	ring finger protein 139
ZC3H13	zinc finger CCCH-type containing 13
CNNM2	In multiple Geneids
WWP2	WW domain containing E3 ubiquitin protein ligase 2
PAIP1	poly(A) binding protein interacting protein 1
L3MBTL	Data not found
GCS1	Data not found
TPM1	tropomyosin 1 (alpha)
GPR68	G protein-coupled receptor 68
APLP2	amyloid beta (A4) precursor-like protein 2
IGF1	insulin-like growth factor 1 (somatomedin C)
FTHP1	Data not found
NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
BBC3	BCL2 binding component 3
ATP9A	ATPase, class II, type 9A
GCOM1	GRINL1A complex locus
GRINL1A	glutamate receptor, ionotropic, N-methyl D-aspartate-like 1A
MIA3	melanoma inhibitory activity family, member 3
SNRNP27	small nuclear ribonucleoprotein 27kDa (U4/U6.U5)
SPRED2	sprouty-related, EVH1 domain containing 2
KIAA0892	Data not found
RRAS2	related RAS viral (r-ras) oncogene homolog 2
PTPN11	protein tyrosine phosphatase, non-receptor type 11
JMJD6	jumonji domain containing 6
CAMTA2	calmodulin binding transcription activator 2
TAF4	TAF4 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 135kDa
ARL3	ADP-ribosylation factor-like 3
COL2A1	collagen, type II, alpha 1
ZNF638	zinc finger protein 638
RBM9	Data not found
CTTN	cortactin
DST	dystonin

[0135]

표 8

[0136]

RNASEH2B	ribonuclease H2, subunit B
TMEM111	transmembrane protein 111
ERC1	ELKS/RAB6-interacting/CAST family member 1
ICALM	Data not found
MFN2	mitofusin 2
ZMYND10	zinc finger, MYND-type containing 10
VPS24	vacuolar protein sorting 24 homolog (S. cerevisiae)
FLJ12529	Data not found
CDC27	cell division cycle 27 homolog (S. cerevisiae)
ZNF22	zinc finger protein 22 (KOX 15)
ZFAND3	zinc finger, AN1-type domain 3
COMMD8	COMM domain containing 8
HEATR2	HEAT repeat containing 2
GLT25D1	glycosyltransferase 25 domain containing 1
FAM13B1	Data not found
HEATR1	HEAT repeat containing 1
TFB2M	transcription factor B2, mitochondrial
OPN3	opsin 3
SLC41A3	solute carrier family 41, member 3
MYO19	myosin XIX
LRP12	low density lipoprotein receptor-related protein 12
MRM1	mitochondrial rRNA methyltransferase 1 homolog (S. cerevisiae)
FBXL12	F-box and leucine-rich repeat protein 12
CHRNA10	cholinergic receptor, nicotinic, alpha 10
C20orf30	chromosome 20 open reading frame 30
ROBO4	roundabout homolog 4, magic roundabout (Drosophila)
FLJ20433	Data not found
TIGD6	tigger transposable element derived 6
TCF7L1	In multiple Geneids
TM7SF4	transmembrane 7 superfamily member 4
GPR27	G protein-coupled receptor 27
B4GALT5	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 5
KIAA1219	Data not found
FAM89B	family with sequence similarity 89, member B
MYO7A	myosin VIIA
MAPK7	mitogen-activated protein kinase 7
GCAT	glycine C-acetyltransferase
PRR14	proline rich 14
RP5-1077B9.4	Data not found
FLJ21865	Data not found

[0137] (OOB error 비교)

[0138] 각 데이터 셋에서 선별된 유전자의 분류정확성을 OOB error를 사용하여 비교하였다.

[0139] 도 10에 나타난 바와 같이, 데이터 셋 A 와 데이터 셋 B에서 선별된 유전자들에 비해 병합한 데이터 셋 AB에서 선별된 유전자들이 분류의 정확성이 높고 안정적이었다.

[0140] (검증용 데이터 셋을 이용한 분류의 정확성 비교)

[0141] 검증용 데이터 셋은 결측치가 많이 포함되어 있어서 SKNN(sequential k nearest neighbors)에 의해 결측치를 추정해서 사용했다. 병합된 데이터 셋에서 선별한 유전자 셋이 오류율(OOB error)이 가장 낮았다. 이것은 각각 독립적인 데이터 셋에서 선별한 유전자 셋의 오류율과 유의한 차이를 보였다; 데이터 셋 A, 데이터 셋 B와의 유의성은 p=5.009e-07, p=1.872e-05였다.

[0142] (각 데이터 셋에서 선별된 유전자 셋의 다른 데이터 셋에서의 일치도 비교)

[0143] 선별된 유의한 유전자들의 발현패턴을 서로 다른 데이터 셋과 병합된 데이터 셋에서 비교하였다. 이것은 잘 선별된 유전자 셋은 다른 데이터 셋에서도 발현이 유사한 패턴으로 유지되고 있을 것이라는 가정에 의한 것이다.

[0144] 도 11에 나타난 바와 같이, 유전자 셋 AB는 병합된 데이터 셋 AB에서 선별된 유전자들의 발현과 다른 데이터 셋에서의 발현패턴의 상관계수를 보여준다. 이값이 클수록 유전자 셋이 잘 선별되었음을 말해준다. 병합된 데이터 셋에서 선별한 유전자들이 가장 상관계수가 높았으며, 이에 대한 p값은 가장 유의미하게 나타났다.

[0145] 도 12에서 다크그레이는 N-, 라이트그레이는 N+ 를 나타낸다. N+군의 발현은 변동이 N-군에 비해 크게 나타났다. 유전자에 따라 N+군의 일부가 N-의 발현패턴과 유사한 경우도 보였다. IGF1을 제외한 유전자들은 N+군에서 과발현되는 경향을 보였다. 반면에 IGF1 은 N-군에서 과발현되는 패턴을 보여주고 있다.

[0146] 두 개의 유전자 간의 발현패턴을 살펴본 결과, 도 13에서와 같이, 강한 양의 상관관계와 강한 음의 상관관계를 갖는 유전자들의 발현패턴을 보여준다. MFN2의 발현은 CTTN, ZFAND3, SLC7A6와 강한 양의 상관관계를 보였고, MAPK7 은 ARL3, LGMN 과 강한 음의 상관관계를 보였다. 그러나, 이러한 상관관계는 N+군에서 나타났으며, N-군에서는 나타나지 않았다.

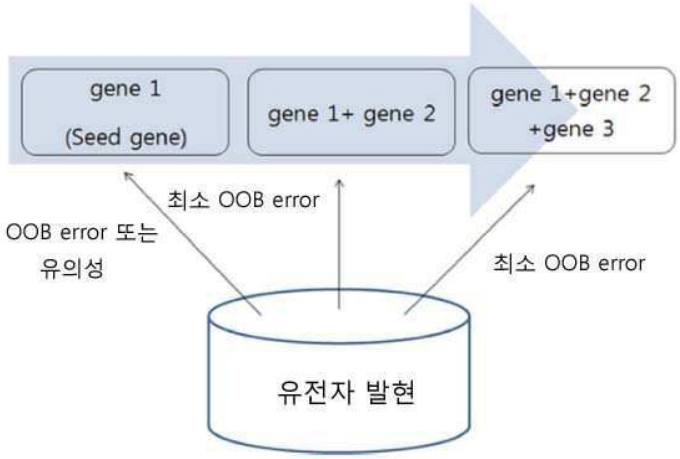
[0147] (염색체상에서 선별된 유전자의 발현패턴 확인)

[0148] 병합한 데이터 셋에서 선별한 유전자들의 발현패턴을 염색체상에서 탐색하였다. 이를 위하여 사전연구로부터 확보한 aCGH 데이터 셋을 사용하였다. 선별된 유전자들은 전체 염색체에 걸쳐 분포하고 있었으며, 림프절 전이 여부를 명확하게 분류하고 있었다. 도 14 내지 17은 선별된 유전자 중 5번, 8번, X 염색체상에 분포한 유전자들의 발현패턴을 보여준다. 이것은 Agilent CGH analytical software (Agilent Technologies) 를 사용하여 작성하였다.

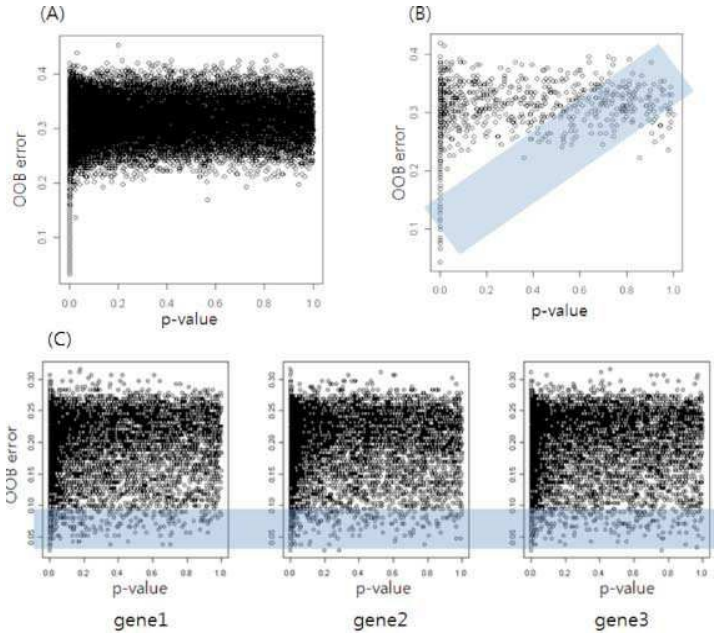
[0149] 도 14 내지 17에서와 같이, 염색체 5번 상의 NR3C1은 N+군에서 과발현하였으며, PAIP1은 하향발현되었다. 8번 염색체와 X 염색체상의 LPR12와 F8은 N+군에서 과발현하였으며, 이 유전자들은 CNV region으로 보고된 영역이다 (http://projects.tcag.ca/variation/, Human genome assembly build 36-hg18 version).

도면

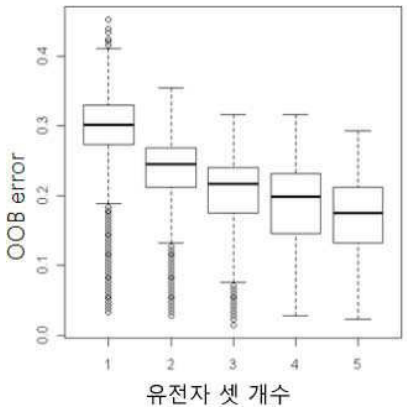
도면1



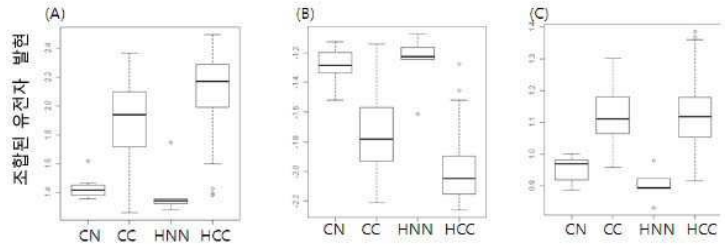
도면2


(A)
(B)
(C)
OOB error
p-value
gene1
gene2
gene3

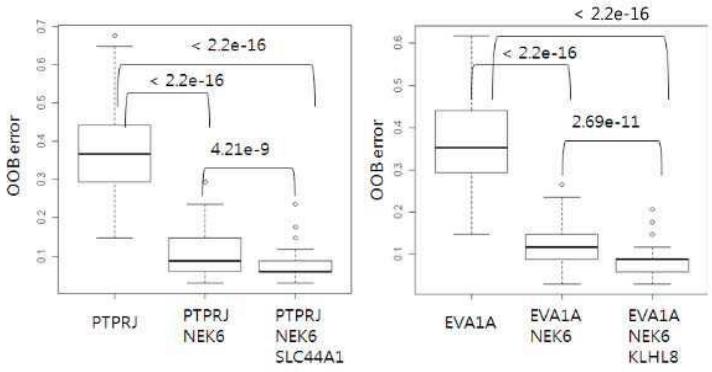
도면3


OOB error
유전자 셋 개수

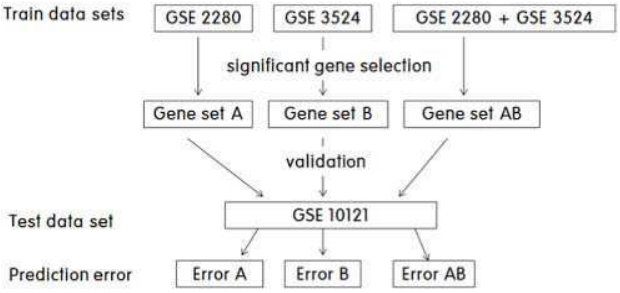
도면4


(A)
(B)
(C)
조합된 유전자 발현
CN
CC
HNN
HCC

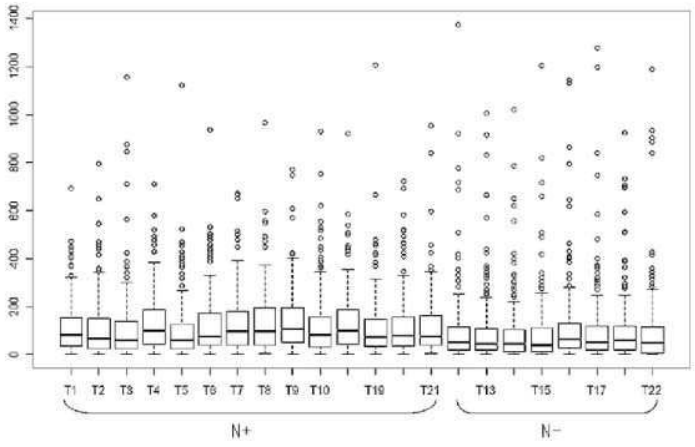
도면5


OOB error
< 2.2e-16
< 2.2e-16
4.21e-9
PTPRJ
PTPRJ NEK6
PTPRJ NEK6 SLC44A1
< 2.2e-16
< 2.2e-16
2.69e-11
EVA1A
EVA1A NEK6
EVA1A NEK6 KLHL8

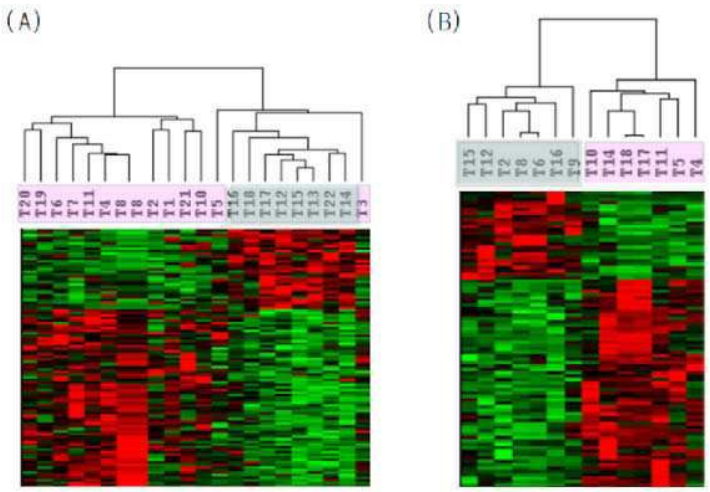
도면6


Train data sets
GSE 2280
GSE 3524
GSE 2280 + GSE 3524
significant gene selection
Gene set A
Gene set B
Gene set AB
validation
Test data set
GSE 10121
Prediction error
Error A
Error B
Error AB

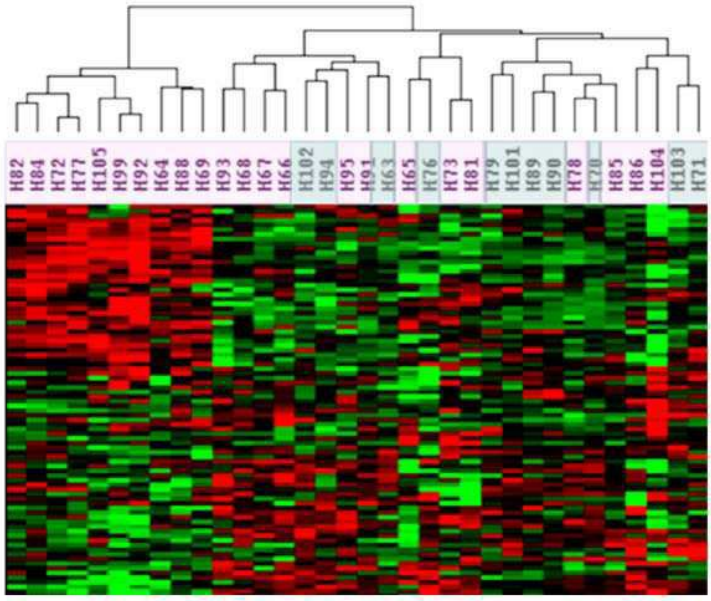
도면7


T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T19 T21 T13 T15 T17 T22
N+
N-

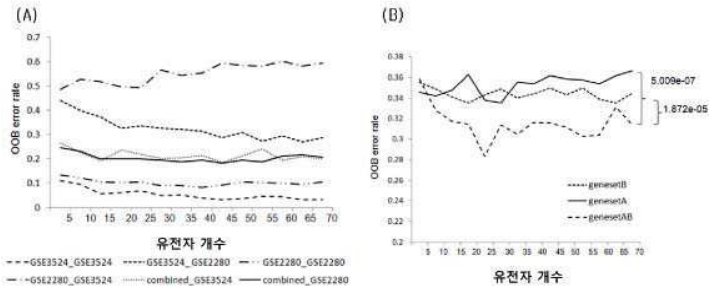
도면8


(A)
(B)

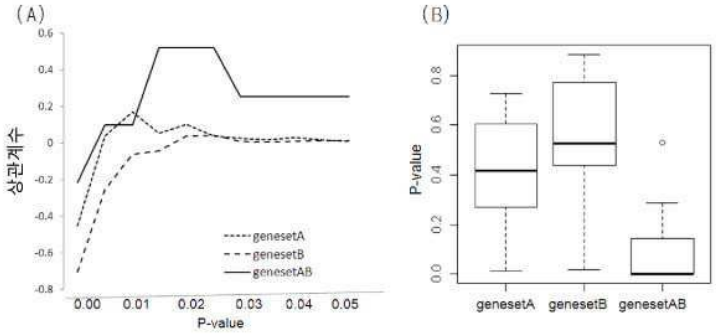
도면9



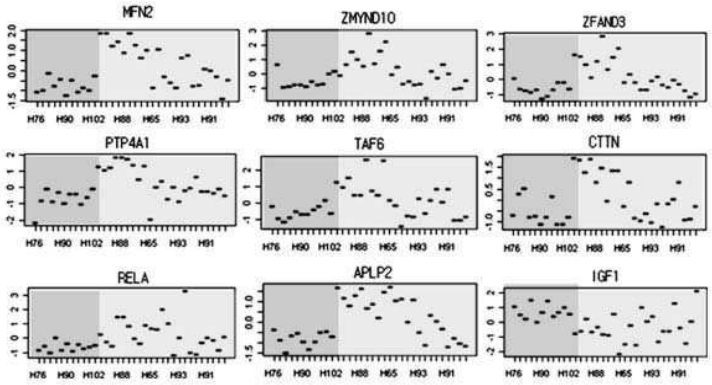
도면10


(A)
(B)
OOB error rate
유전자 개수
GSE3524_GSE3524
GSE3524_GSE2280
GSE2280_GSE2280
GSE2280_GSE3524
combined_GSE3524
combined_GSE2280
5.009e-07
1.872e-05
genesetB
genesetA
genesetAB

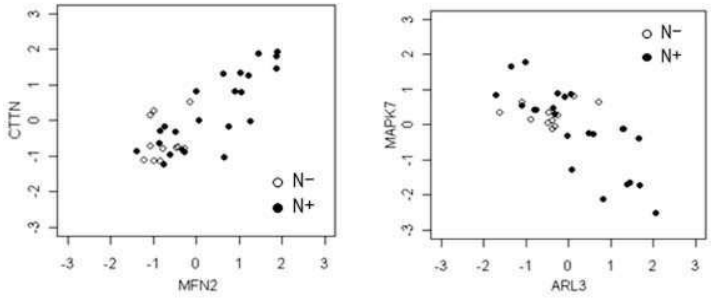
도면11



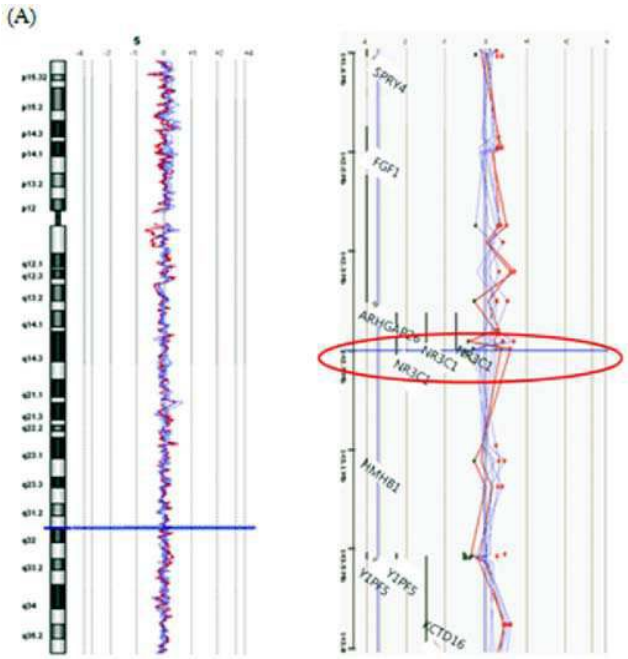
도면12



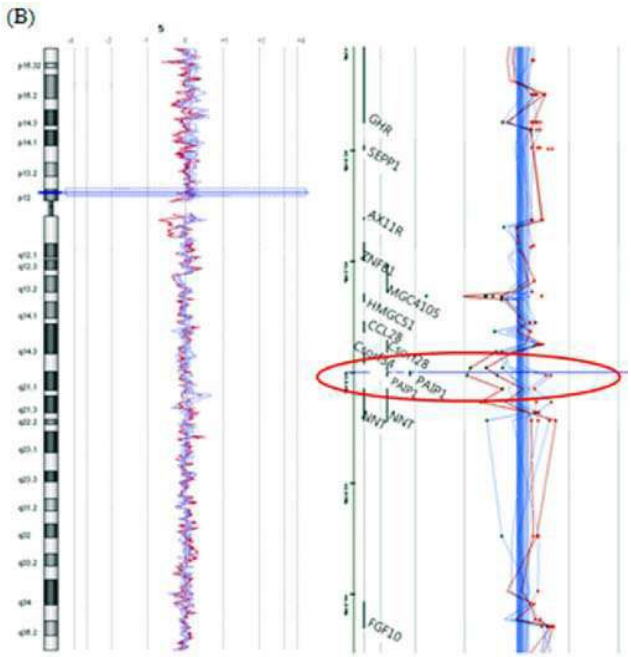
도면13



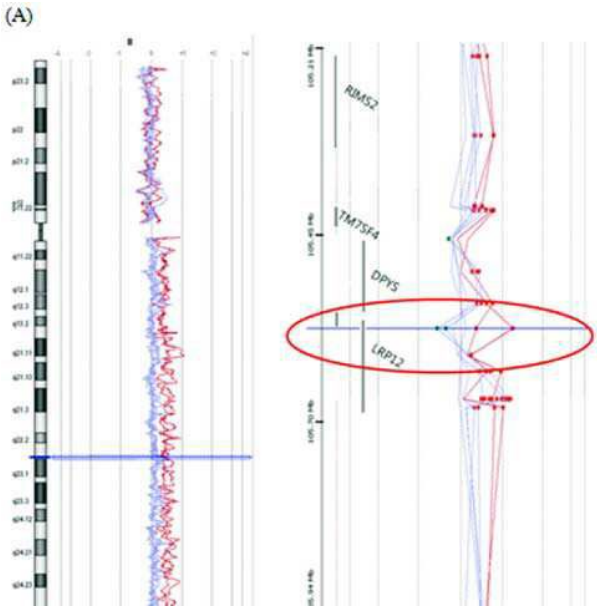
도면14


(A)
SPRY4
FGF1
ARHGAP5
NR3C1
NR3C1
NR3C1
MMHB1
YIPF3
YIPF5
KCTD16

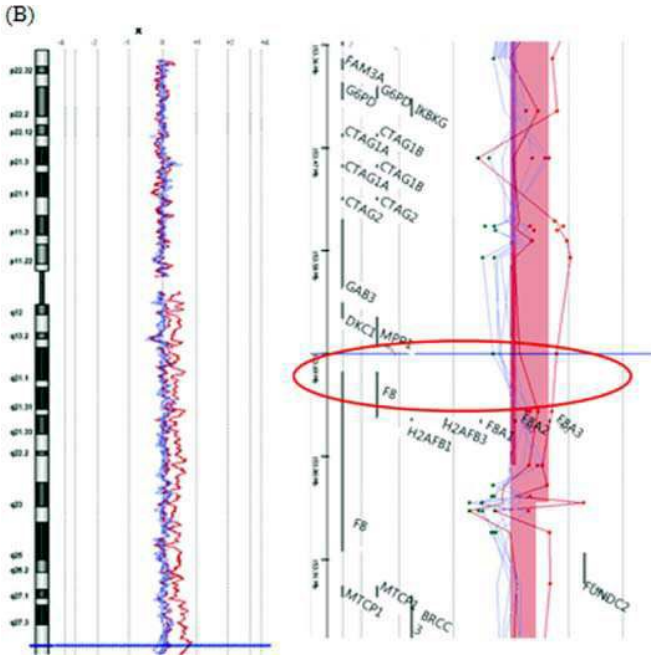
도면15


(B)
GHR
SEPP1
AXL1R
DNFB1
MGC4105
HMGCS1
CCL28
PAIP1
PAIP1
MVK
MVK
FGF10

도면16


(A)
RIMS2
TM7SF4
DPYS
LRP12

도면17


(B)
FAM3A
G6PD
IKBKG
CTAG1B
CTAG1A
CTAG2
GAB3
DKC1
MPP1
FB
H2AFB3
H2AFB1
FBA1
FBA2
FBA3
MTCP1
BRCC3
FUNDC2