

- (51) 국제특허분류(Int. Cl.)
G16B 40/00 (2019.01) *C12Q 1/6886* (2018.01)
G06N 3/04 (2023.01) *G06N 3/08* (2023.01)
G16B 20/20 (2019.01) *G16B 30/10* (2019.01)
G16H 10/20 (2018.01) *G16H 10/40* (2018.01)
G16H 10/60 (2018.01) *G16H 50/20* (2018.01)
G16H 50/30 (2018.01)

(52) CPC특허분류
G16B 40/00 (2019.02)
C12Q 1/6886 (2022.01)

(21) 출원번호 10-2023-0073754

(22) 출원일자 2023년06월08일
심사청구일자 2023년06월08일

(30) 우선권주장
1020220141324 2022년10월28일 대한민국(KR)

(71) 출원인
주식회사 클리노믹스
울산광역시 울주군 언양읍 유니스트길 50,110동
3층 307호
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50

(72) 발명자
김병철
울산광역시 울주군 삼남읍 도호1길 39-11, 202동
2401호(금아드림팰리스)
김여진
경상남도 양산시 물금읍 가촌서로 11, 102동 501
호(양산물금 한신 더휴 아파트)
(뒷면에 계속)

(74) 대리인
특허법인정진

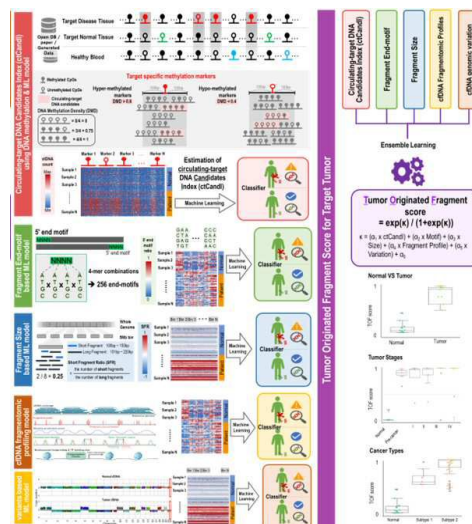
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 Cell-free DNA 를 이용한 건강 및 질병관리 시스템 및 방법

(57) 요약

본 발명은 체액 속의 cfDNA 단편이 가지는 후성 유전체 및 변이 정보와 cfDNA 단편 프로파일링을 활용하여 보다 효과적으로 건강상태를 파악하고 질병을 진단, 예측하기 위한 시스템 및 방법에 관한 것이다. 본 발명에 따르면 체액 속에 존재하는 cfDNA 단편이 갖는 cfDNA의 메틸화 정보를 이용하여, DNA의 메틸화 밀도(DNA Methylation Density; DMD)를 계산하고 ctCandi (circulating-target DNA Candidate index)를 구하여 타겟 조직 또는 세포 유래 cfDNA를 정량화 함으로써, 체액 시료만으로 건강/질병관리에 대한 정보를 제공할 수 있다. 또한, cfDNA의 메틸화 정보, 단편의 특성 및 변이정보를 이용하여, 암을 조기에 예측 및 진단할 수 있으며, 체액 시료만으로 노화와 같은 건강상태 및 질병에 대한 정보를 높은 민감도와 특이도로 모니터링 할 수 있다.

대표도 - 도8



(52) CPC특허분류

G06N 3/04 (2023.01)
G06N 3/08 (2023.01)
G16B 20/20 (2019.02)
G16B 30/10 (2019.02)
G16H 10/20 (2021.08)
G16H 10/40 (2021.08)
G16H 10/60 (2021.08)
G16H 50/20 (2018.01)
G16H 50/30 (2018.01)

(72) 발명자

전성원

울산광역시 남구 와와로 17(무거동)

전하현

울산광역시 울주군 삼남읍 도호1길 39-11, 201동
 3402호(금아드림팰리스)

엄효진

울산광역시 남구 장생포고래로179번길 4(장생포동)

김병철

경기도 용인시 기흥구 한보라2로 47-54(보라동)

박중화

울산광역시 울주군 언양읍 반천강변길 51, 101동
 1710호(현대아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1425156792
과제번호	P0016195
부처명	중소벤처기업부
과제관리(전문)기관명	한국산업기술진흥원
연구사업명	규제자유특구혁신사업육성(R&D)
연구과제명	지능형 오믹스 빅데이터 기반 질병 예측 및 진단 마커 개발 실증
기 여 율	1/1
과제수행기관명	(주)클리노믹스
연구기간	2021.01.01 ~ 2022.11.30

명세서

청구범위

청구항 1

- 1) 생물학적 시료 유래 cfDNA 로 이루어진 서열 DNA를 타겟 특이적 CpG 마커 사이트를 기준으로 정렬하는 단계;
- 2) 상기 타겟 특이적 CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 마커 영역으로 정의하고, 정렬된 DNA 중 상기 타겟 특이적 영역 내 존재하는 DNA의 메틸화 밀도(DNA Methylation Density; DMD)를 도출하는 단계;
- 3) 상기 타겟 특이적 영역에 존재하는 DNA 중 2) 단계에서 도출된 DMD 값이 역치 값을 초과하는 DNA 를 ctDNA(circulating-target DNA) 후보로 정의하고 ctDNA 후보 수를 계수하여 정규화하는 단계; 및
- 4) 각 타겟 특이적 영역에서 도출된 정규화된 ctDNA 후보 수의 평균을 ctCandi (circulating-target DNA Candidate index) 로 도출하는 단계; 를 포함하는, 타겟 유래 cfDNA 정량화 방법.

청구항 2

제1항에 있어서, 상기 DNA 메틸화 밀도 (DMD)는 DNA 한 분자 내에 존재하는 일부 또는 전체 영역에 대하여 메틸화된 CpG 의 수를 전체 CpG 수로 나눈 값인, 타겟 유래 cfDNA 정량화 방법.

청구항 3

제1항에 있어서, 상기 DNA 메틸화 밀도 (DMD) 도출을 위한 메틸화 정도 측정은 차세대 염기서열 분석법, 장서열 분석법, 패널, Chip 및 PCR로 이루어진 군에서 선택된 1종의 메틸화 측정 방법을 이용하여 측정되는 것인, 타겟 유래 cfDNA 정량화 방법.

청구항 4

제1항에 있어서, 상기 5) ctCandi (circulating-target DNA Candidate index) 값에 대하여 cfDNA 디콘볼루션 (deconvolution) 을 수행하는 단계; 를 더 포함하는 타겟 유래 cfDNA 정량화 방법.

청구항 5

제1항에 있어서, 상기 ctDNA는 체액 내 존재하는 cfDNA 중 타겟 조직 또는 세포에서 유리된 cfDNA인, 타겟 유래 cfDNA 정량화 방법.

청구항 6

제1항에 있어서, 상기 생물학적 시료는 액체 생검 시료인, 타겟 유래 cfDNA 정량화 방법.

청구항 7

제6항에 있어서, 상기 액체 생검시료는 혈액, 혈청, 혈장, 타액, 눈물, 소변, 대변, 질액, 소화액, 뇌척수액, 및 콧물로 이루어진 군에서 선택된 1종 이상인, 타겟 유래 cfDNA 정량화 방법.

청구항 8

- 1) 분석이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받는 단계;
 - 2) 상기 CpG 마커 사이트 정보를 기준으로 상기 입력된 서열 DNA 정보를 정렬하고, CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하며, 상기 타겟 특이적 영역 내 존재하는 서열 DNA 부위의 DNA 메틸화 밀도 (DNA Methylation density, DMD)를 계산하는 단계;
 - 3) 상기 타겟 특이적 영역에 존재하는 DNA 중 계산된 DMD 값이 역치값을 초과하는 DNA수를 ctDNA 후보 수로 정의하여 정규화하는 단계;
 - 4) 정규화된 ctDNA 후보 수의 평균을 계산하여 ctCandi (circulating-target DNA Candidate index) 로 도출하는 단계;
 - 5) 입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 학습된 머신러닝을 이용해 시료 제공자의 건강 상태에 대한 확률값을 산출하는 단계; 및
 - 6) 상기 산출된 확률값을 기준으로 상기 시료 제공자의 건강 상태를 식별하는 단계; 를 포함하고,
- 상기 학습된 머신러닝은 시료 제공자의 생체 정보와 도출된 ctCandi 정보를 입력으로 건강 상태에 대한 확률값을 출력하는 것인, 건강 상태를 식별하는 방법.

청구항 9

제8항에 있어서, 상기 머신러닝 알고리즘은 나이브 베이즈(Naive Bayes), KNN(K Nearest Neighbors), 랜덤 포레스트, 로지스틱 회귀 분석, 서포트 벡터 머신, 의사결정나무, 연관성 규칙 마이닝, 인공신경망, 선형 회귀, 순환 신경망 및 딥러닝 중 적어도 하나인, 건강 상태를 식별하는 방법.

청구항 10

분석이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받는 입력부;

상기 CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하며, 상기 입력된 서열 DNA 정보를 정렬하고, 상기 타겟 특이적 영역 내 존재하는 서열 DNA의 메틸화 밀도 (DNA Methylation density, DMD)를 계산하고,

상기 타겟 특이적 영역에 존재하는 서열 DNA 중 계산된 DMD 값이 역치값을 초과하는 DNA 수를 ctDNA 후보 수로 정의하여 정규화하며,

정규화된 ctDNA 후보 수의 평균을 계산하여 ctCandi (circulating-target DNA Candidate index) 로 도출하고,

입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 학습된 머신러닝을 이용해 시료 제공자의 건강 상태에 대한 확률값을 산출하며, 상기 학습된 머신러닝은 시료 제공자의 생체 정보와 도출된 ctCandi 정보를 입력으로 건강 상태에 대한 확률값을 출력하는 것이고, 상기 산출된 확률값을 기준으로 상기 시료 제공자의 건강 상태를 식별하는 프로세서; 및

상기 식별된 건강 상태 결과를 출력하는 출력부; 를 포함하는 건강 상태를 식별하는 장치.

청구항 11

제10항에 있어서, 상기 식별된 건강 상태 결과는 타겟 조직의 건강 상태, 질병 유무 진단 또는 질병의 진행 정도에 대한 정보인, 건강 상태를 식별하는 장치.

청구항 12

- 1) 진단이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받아 제8항에 따른 방법으로 산출된 시료 제공자의 건강상태에 대한 확률값을 데이터로 입력받는 단계;
- 2) cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델에서 도출된 스코어링 값과 질병 진단 정보의 상관관계에 대한 데이터를 입력받는 단계;
- 3) 상기 데이터를 정규화 또는 표준화하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어 값을 도출하는 단계; 및
- 4) 상기 TOF 스코어 값을 이용하여 상기 진단이 필요한 시료의 암 유무 또는 병기를 식별하는 단계; 를 포함하는 암을 진단하는 방법.

청구항 13

제12항에 있어서, 상기 cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델은

cfDNA 리드의 5' 말단의 4개 염기서열 모티프의 빈도를 표준화(Standardization)하고, 도출된 표준화 값과 암 진단에 대한 정보를 머신러닝 알고리즘을 통해 학습시켜, 상기 표준화된 값을 입력 데이터로 생성하는, 말단 모티프 특성을 이용한 스코어링 모델;

cfDNA 단편의 SFR(short fragment ratio) 값을 표준화(Standardization)하고, 도출된 표준화 값과 암 진단에 대한 정보를 머신러닝 알고리즘을 통해 학습시켜, 상기 표준화된 값을 입력 데이터로 생성하는, SFR 특성을 이용한 스코어링 모델;

Jagged end length, Preferred end coordinates, Oriented end density, Motif Diversity Score, Window Protection Score, cfDNA integrity 또는 Nucleosomal footprinting을 스코어링 하여 정규화 또는 표준화하고, 상기 도출된 값과 암 진단에 대한 정보를 머신러닝 알고리즘을 통해 학습시켜, 상기 도출된 값을 입력데이터로 생성하는, cfDNA 단편특성을 이용한 스코어링 모델 중 하나 이상인, 암을 진단하는 방법.

청구항 14

제13항에 있어서, 상기 TOF 스코어 값은 상기 1) 및 2) 단계에서 입력받은 데이터 정보를 입력으로 한 머신러닝 알고리즘 학습을 통해 획득된 값인, 암을 진단하는 방법.

청구항 15

제14항에 있어서, 상기 머신러닝 알고리즘은 앙상블 학습 모델인, 암을 진단하는 방법.

청구항 16

진단이 필요한 시료 유래 cfDNA 로 이루어진 서열 리드 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받아 제8항에 따른 방법으로 산출된 시료 제공자의 건강상태에 대한 확률값 데이터; 및 cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델에서 도출된 스코어링 값과 질병 진단 정보의 상관관계에 대한 데이터;를 입력받는 입력부;

상기 입력된 데이터들을 정규화 또는 표준화하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어 값을 도출하고 상기 TOF 스코어 값을 이용하여 상기 진단이 필요한 시료의 암 유무 또는 병기를 식별하는 프로세서; 및

상기 식별된 암 유무 또는 병기에 관한 결과를 출력하는 출력부; 를 포함하는 암 진단 장치.

발명의 설명

기술 분야

[0001] 본 발명은 체액 내 cell-free DNA 단편이 가지는 후성 유전체 및 변이 정보와 cfDNA 단편 프로파일링을 활용하여 질병을 보다 효과적으로 진단, 예측하고 건강상태를 관리하기 위한 시스템 및 방법에 관한 것이다.

배경 기술

[0003] 체액 내에는 다양한 생체분자들(cfDNA, CTC, Exosome 등)이 떠돌아다니는데, 그 중에서 대표적인 예로 세포유리 DNA가 있다. 세포유리 DNA(cell-free DNA, cfDNA)란, 세포 안에서만 존재하지 않고, 세포사멸 등과 같은 다양한 원인에 의하여 세포 밖으로 나와서, 체액 속에 떠돌아다니는 DNA 조각을 의미한다. cfDNA는 우리 몸 전체의 상태를 반영하여 건강상태를 파악하고, 각종 질병을 진단하고, 예후 추정 및 모니터링을 하는 대리 표지자로서 역할을 할 수 있다.

[0004] 생체 내 계놈은 모두 동일하지만, 계놈상의 유전자가 발현되는 시기, 위치, 정도에 따라 세포와 조직에서 기능이 달라지게 된다. DNA 메틸화는 신체를 이루고 있는 각각의 조직, 세포에서 알맞은 기능을 하기 위해 조직과 세포마다 특이적인 메틸화 패턴을 가진다. 그리고 이러한 조절을 일생에 걸쳐 끊임없이 외부환경에 의해 변화하고, 다음 세대로까지 전달이 된다. 이렇게 유전자 서열자체의 변이가 없이 유전자의 기능에 영향을 주는 유전적 현상을 연구하는 학문이 후성유전학(Epigenetics)이다.

[0005] cfDNA는 건강한 사람의 경우 일반적으로 대부분이 혈구세포에서 유래한 cfDNA로 구성되어 있으나, 특정 조직이나 세포가 다양한 원인에 의해 사멸 (Apoptosis)되거나 괴사(Necrosis) 되는 경우에는 해당 조직이나 세포속에 포함된 DNA가 혈액으로 방출되기 때문에 혈액 내 cfDNA 구성 비율이 특정 조직이나 세포에서 유래한 cfDNA가 많아지는 방향으로 변화하게 된다. 이러한 변화는 조직마다 특이적 패턴을 가지는 DNA 메틸화 특성을 이용하여 확인할 수 있고, 해당 연구를 cfDNA deconvolution이라고 하는 구성성분 분석을 통해 수행할 수 있다. 이러한 원발 조직 (Tissue of origin, TOO) 분석이 최근 활발히 진행되고 있다.

[0006] 특히, 암 환자의 체액에는 cell-free DNA(cfDNA)가 증가됨이 보고된 바 있다. 암세포에서 배출된 cfDNA는 각별히 순환종양 DNA(circulating tumor: ctDNA)라고 명명하며, 이 ctDNA는 암세포가 파열되어 사멸하는 경우 그 찌꺼기가 혈류 속으로 방출되는데, 그 속에서 종양의 DNA가 포함되는 것을 말한다. 혈액 속에 cfDNA형태로 떠돌아다니는 ctDNA의 암 관련 유전적 변화의 패턴을 프로파일링하면 한 개 혹은 여러 암의 조기발견을 위해 건강하거나, 위험에 처한 인구집단을 대규모로 스크리닝할 수 있다.

[0007] 최근 다양한 연구 결과에서, cfDNA 길이 및 말단 모티프와 같은 cfDNA 단편 프로파일이 암에서 독특한 패턴을 보이는 것이 보고되었다. 일반적으로 cfDNA의 길이 프로파일은 뉴클레오타이드 구조와 관련하여 대략 166bp (base pair)에서 주요 피크를 나타내는 반면에, 종양 유래의 특징을 지닌 cfDNA의 길이는 건강한 사람보다 길이가 짧으며, 더 많은 양으로 존재한다.

[0008] cfDNA 단편 길이 외에도 말단 모티프라고 하는 cfDNA의 5' 말단은 절단 효소 종류에 따라 원래 조직에 대한 정보가 포함되어 있다. 이러한 cfDNA의 특징들은 유래한 조직 및 세포 종류에 따라 달라지므로, 다양한 임상 응용 가능성을 가지고 있다.

[0009] 뿐만 아니라, cfDNA는 형태 (Topology), 메틸화 상태, 뉴클레오타이드 구조, DNA 변이정보 등 다양한 특성들을 포함하고 있으며, cfDNA 단편의 특성을 정량화 하기 위해 측정할 수 있는 파라미터들이 다수 알려져 있다.

[0010] 유전자 복제수변이(CNV, Copy-number variation)는 구조변이(Structural variation)의 한 종류로 유전체의 특정 영역이 2개 이상으로 복사되거나 결손되는 현상을 의미한다. CNV는 SNP처럼 개체마다 다른 변이를 갖는다. 특정 유전자 영역의 Copy-number는 암 세포 혹은 질병에 걸린 세포에서 특히 크게 변화하며, 암세포에서의 CNV를 특별히 Somatic copy-number alteration (SCNA)라고 하는데, 증식과 관련된 유전자(Oncogene) 발현을 증가(증폭)시키거나, 암 억제 유전자 (Tumor Suppressor Gene) 발현을 감소(억제)시키는 역할을 한다. 2010년 Nature에 보고된 바에 의하면, 약 3천개의 암 시료를 분석한 결과 암세포 유전체의 17%가 초과복제되고 16%가 결손되어 있다고 한다.

[0011] 이처럼 cfDNA는 생체 상태에 대한 다양한 정보를 담고 있으며, 이를 활용하여 효과적으로 질병을 진단하고 건강 상태에 대한 정보를 제공할 수 있는 예측 모델 및 방법에 대한 필요성이 있다.

발명의 내용

해결하려는 과제

- [0013] 이에 본 발명의 발명자는, cfDNA 단편의 다양한 특성을 적용하여 건강상태 및 질병에 대한 정보를 제공할 수 있는 새로운 예측 모델 및 방법에 대하여 연구하였으며, cfDNA 메틸화 정보의 분석방법을 새롭게 구축하고, 타겟 유래 cfDNA 정량화 방법을 제공하며, cfDNA 단편으로부터 추출된 데이터를 단독 또는 앙상블하여 머신 러닝 기법을 통해 제조된 새로운 건강/질병관리 및 암 진단에 대한 정보 제공용 장치, 시스템 및 방법을 확인하고 본 발명을 완성하였다.
- [0014] 또한 본 발명의 발명자는 상기 cfDNA 메틸화 정보를 DNA의 메틸화 밀도(DNA Methylation Density; DMD)로 수치화하고, 이를 이용하여 목적하는 타겟 유래 cfDNA를 정량화 할 수 있는 ctCandi (circulating-target DNA Candidate index)를 정의하고 새로운 모델을 확립하였다.
- [0015] 또한 본 발명의 발명자는 상기 cfDNA 단편으로부터 추출된 데이터를 이용하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어링 방법을 정의하고 새로운 암 진단 장치 및 방법을 확립하였다.

과제의 해결 수단

- [0017] 따라서 본 발명은 1) 생물학적 시료 유래 cfDNA 로 이루어진 서열 DNA를 타겟 특이적 CpG 마커 사이트를 기준으로 정렬하는 단계; 2) 상기 타겟 특이적 CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하고, 정렬된 DNA 중 상기 타겟 특이적 영역 내 존재하는 DNA의 메틸화 밀도(DNA Methylation Density; DMD)를 도출하는 단계; 3) 상기 타겟 특이적 영역에 존재하는 DNA 중 2) 단계에서 도출된 DMD 값이 역치 값을 초과하는 DNA를 ctDNA(circulating-target DNA) 후보로 정의하고 ctDNA 후보 수를 계수하여 정규화하는 단계; 및 4) 각 타겟 특이적 영역에서 도출된 정규화된 ctDNA후보 수의 평균을 ctCandi (circulating-target DNA Candidate index) 로 도출하는 단계; 를 포함하는, 타겟 유래 cfDNA 정량화 방법을 제공한다.
- [0018] 또한 본 발명에서 정의하는 DNA 메틸화 밀도 (DNA Methylation density, DMD)는 단일 CpG 사이트 뿐 만 아니라, 리드(read) 및 단편(fragment)의 메틸화 밀도(methylation density), DNA의 메틸화 밀도(methylation density)를 포함하고, DNA 한 분자 내에 존재하는 일부 또는 전체 영역에 대해서 메틸화된 CpG 의 수를 전체 CpG 수로 나눈 값인, 각 DNA의 메틸화 정도의 수치화 방법이다. 상기 DNA 메틸화 밀도 (DMD) 도출을 위한 메틸화 정도 측정은 차세대 염기서열 분석법, 장서열 분석법, 패널, Chip 및 PCR로 이루어진 군에서 선택된 1종의 메틸화 측정 방법을 이용하여 측정되는 것 일 수 있다.
- [0019] 또한 본 발명은 1) 분석이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받는 단계; 2) 상기 CpG 마커 사이트 정보를 기준으로 상기 입력된 서열 DNA 정보를 정렬하고, CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하며, 상기 타겟 특이적 영역 내 존재하는 서열 DNA 부위의 DNA 메틸화 밀도 (DNA Methylation density, DMD)를 계산하는 단계; 3) 상기 타겟 특이적 영역에 존재하는 DNA중 계산된 DMD 값이 역치값을 초과하는 DNA 수를 ctDNA 후보 수로 정의하여 정규화하는 단계; 4) 정규화된 ctDNA 후보 수의 평균을 계산하여 ctCandi (circulating-target DNA Candidate index) 로 도출하는 단계; 5) 입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 학습된 머신러닝을 이용해 시료 제공자의 건강상태에 대한 확률값을 산출하는 단계; 및 6) 상기 산출된 확률값을 기준으로 상기 시료 제공자의 건강 상태를 식별하는 단계; 를 포함하고, 상기 학습된 머신러닝은 시료 제공자의 생체 정보와 도출된 ctCandi 정보를 입력으로 건강 상태에 대한 확률값을 출력하는 것인, 건강 상태를 식별하는 방법을 제공한다.
- [0020] 또한 본 발명은 분석이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받는 입력부; 상기 CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하며, 상기 입력된 서열 DNA정보를 정렬하고, 상기 타겟 특이적 영역 내 존재하는 서열 DNA의 메틸화 밀도 (DNA Methylation density, DMD)를 계산하고, 상기 타겟 특이적 영역에 존재하는 서열 DNA 중 계산된 DMD 값이 역치값을 초과하는 DNA 수를 ctDNA 후보 수로 정의하여 정규화하며, 정규화된 ctDNA 후보 수의 평균을 계산하여 ctCandi (circulating-target DNA Candidate index) 로 도출하고, 입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 학습된 머신러닝을 이용해 시료 제공자의 건강상태에 대한 확률값을 산출하며, 상기 학습된 머신러닝은 시료 제공자의 생체 정보와 도출된 ctCandi 정보를 입력으로 건강 상태에 대한 확률값을 출력하는 것이고, 상기 산출된 확률값을 기준으로 상기 시료 제공자의 건강 상태를 식

별하는 프로세서; 및 상기 식별된 건강 상태 결과를 출력하는 출력부; 를 포함하는 건강 상태를 식별하는 장치를 제공한다.

[0021] 또한 본 발명은 1) 진단이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받아 상기 건강 상태를 식별하는 방법에 따라 산출된 시료 제공자의 건강상태에 대한 확률값을 데이터로 입력받는 단계; 2) cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델에서 도출된 스코어링 값과 질병 진단 정보의 상관관계에 대한 데이터를 입력받는 단계; 3) 상기 데이터를 정규화 또는 표준화하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어 값을 도출하는 단계; 및 4) 상기 TOF 스코어 값을 이용하여 상기 진단이 필요한 시료의 암 유무 또는 병기를 식별하는 단계; 를 포함하는 암을 진단하는 방법을 제공한다.

[0022] 또한 본 발명은 진단이 필요한 시료 유래 cfDNA 로 이루어진 서열 리드 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받아 상기 건강 상태를 식별하는 방법에 따라 산출된 시료 제공자의 건강상태에 대한 확률값 데이터; 및 cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델에서 도출된 스코어링 값과 질병 진단 정보의 상관관계에 대한 데이터; 를 입력받는 입력부; 상기 입력된 데이터들을 정규화 또는 표준화하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어 값을 도출하고 상기 TOF 스코어 값을 이용하여 상기 진단이 필요한 시료의 암 유무 또는 병기를 식별하는 프로세서; 및 상기 식별된 암 유무 또는 병기에 관한 결과를 출력하는 출력부; 를 포함하는 암 진단 장치를 제공한다.

발명의 효과

[0024] 본 발명에 따르면 체액 속에 존재하는 cfDNA 단편이 갖는 cfDNA의 메틸화 정보를 이용하여, DNA의 메틸화 밀도 (DNA Methylation Density; DMD)를 계산하고 ctCandi (circulating-target DNA Candidate index)를 구하여 타겟 조직 또는 세포 유래 cfDNA를 정량화 함으로써, 체액 시료만으로 건강 및 질병관리에 대한 정보를 제공할 수 있다. 또한, cfDNA의 메틸화 정보, 단편의 특성 및 변이정보를 이용하여, 암을 조기에 예측 및 진단할 수 있으며, 체액 시료만으로 노화와 같은 건강상태 및 질병에 대한 정보를 높은 민감도와 특이도로 모니터링 할 수 있다.

도면의 간단한 설명

[0026] 도 1은 본 발명의 종양 유래 단편 (tumor originated fragment, TOF) 스코어링 모델의 모식도이다. 도 1의 A는 TOF 스코어의 컨셉을 나타내며 미리 결정된 TOF 스코어 값과 시료의 TOF 스코어를 비교하여 암에 대한 정보를 제공할 수 있다. 도 1의 B는 TOF 스코어링 모델의 구축 방법을 나타낸 도이다.

도 2는 폐암 예측에 사용되는 circulating-target DNA 후보수 인덱스 구축 (circulating-target DNA candidate index, ctCandi) 의 모식도이다.

도 2a는 타겟 특이적 영역 내 존재하는 서열 DNA 부위의 DNA 메틸화 밀도 (DNA Methylation density, DMD) 및 ctCandi 의 개념, ctCandi 를 이용하여 건강 상태를 식별하는 방법을 설명한다.

도 2b의 (A)는 circulating-target DNA 후보 수의 히트맵을 나타내며, 상단의 색은 시료 유형을 나타낸다 (왼쪽에서 오른쪽으로 NSCLC의 대조군, 병기 I, 병기 II, 병기 III, 병기 IV). 도 2b의 (B)는 정상 또는 암 병기에 따른 테스트 시료의 ctCandi 를 박스 플롯으로 나타낸 도이다. 도 2b의 (C)는 독립 검증 모델을 이용한, cfDNA 후보 수 기반의 *in silico* 시뮬레이션 결과를 나타낸 도이다.

도 3은 cfDNA 의 메틸화, 말단-모티프 및 단편 크기 특징을 이용한 머신 러닝 모델의 모식도이다.

도 4는 본 발명의 예측 및 진단 머신러닝 모델에 의한 암 진단 결과를 나타낸 도이다. 도 4의 A 및 B는 테스트 세트에서 ctDNA 후보 수 기반 폐암 진단 머신러닝 모델의 암 진단 결과를 나타낸다. 도 4의 C 및 D는 테스트 세트에서 말단-모티프 기반 머신 러닝 모델의 암 진단 결과를 나타낸다. 도 4의 E 및 F는 테스트 세트에서 SFR-기반 머신 러닝 모델의 진단 결과를 나타낸다.

도 5는 ctDNA 후보 수, 말단 모티프 및 SFR 특성에 대한 머신러닝 알고리즘의 일종인 로지스틱 회귀 분석을 이용하는 앙상블 학습 모델에서 도출된 TOF 스코어 기반 암 진단 결과를 나타낸 도이다. 도 5의 A는 테스트 세트에서 TOF-스코어 기반 암 예측 결과를 나타낸 도이다. 도 5의 B는 T 병기에 따른 N 및 M 병기가 혼합된 테스트 시료에서의 TOF 스코어링 모델을 이용한 분류 효과를 나타낸 도이다. 도 5의 C는 비소세포성 폐암 데이터 세트에서도 TOF 스코어링 모델이 우수한 분류 효과를 나타냄을 보여주는 결과이다.

도 6은 TOF 스코어링을 이용하여 정상 대조군, NSCLC 아형인 선암종 (ADC) 및 편평세포 암종 (SQC)를 효과적으로 분류할 수 있음을 보여주는 결과이다.

도 7은 TOF 스코어 기반 위험도 측정 기준의 예시를 나타낸다. 종양 유래 단편 스코어링 학습을 통해 미리 설정된 TOF 스코어 기준 값과, 대상 시료의 TOF 스코어를 비교하여, 대상 시료의 질병 위험도에 대한 정보를 제공할 수 있다.

도 8은 본 발명의 TOF 스코어를 이용한 암 진단, 예측 방법의 모식도이다.

발명을 실시하기 위한 구체적인 내용

- [0027] 본 발명은 타겟 유래 cfDNA 정량화 방법 및 cfDNA 를 이용한 건강 상태를 식별하는 방법, 암 예측, 진단하는 방법 및 장치에 관한 것이다.
- [0028] 본 발명은 DNA 메틸화 밀도 (DNA Methylation density, DMD) 도출 방법을 이용하는 새로운 타겟 유래 cfDNA 정량화 방법, 새롭게 도출되는 ctCandi (circulating-target DNA Candidate index) 를 이용하는 검출 대상 시료의 타겟 유래 cfDNA의 정량화 방법, 다양한 cfDNA 단편 특성을 이용한 새로운 건강 상태를 식별하는 방법 및 장치를 제공한다.
- [0029] 또한 본 발명은 다양한cfDNA 단편 정보를 함께 학습하고, 스코어링하는 단계를 포함하는 종양 유래 단편 (tumor originated fragment, TOF) 스코어링 방법을 제공하며, 상기 스코어링 방법을 이용한 암 진단 방법 및 장치를 제공한다.
- [0031] 이하, 본 발명을 구체적인 실시예를 통해 상세히 설명하나 하기 실시예에 의하여 본 발명의 범위가 제한되는 것은 아니다.
- [0033] 본 발명은 1) 생물학적 시료 유래 cfDNA 로 이루어진 서열 DNA를 타겟 특이적 CpG 마커 사이트를 기준으로 정렬하는 단계; 2) 상기 타겟 특이적 CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 마커 영역으로 정의하고, 정렬된 DNA 중 상기 타겟 특이적 영역 내 존재하는 DNA의 메틸화 밀도(DNA Methylation Density; DMD)를 도출하는 단계; 3) 상기 타겟 특이적 영역에 존재하는 DNA 중 2) 단계에서 도출된 DMD 값이 역치 값을 초과하는 DNA 를 ctDNA(circulating-target DNA) 후보로 정의하고 ctDNA 후보 수를 계수하여 정규화하는 단계; 및 4) 각 타겟 특이적 영역에서 도출된 정규화된 ctDNA후보 수의 평균을 ctCandi (circulating-target DNA Candidate index) 로 도출하는 단계; 를 포함하는, 타겟 유래 cfDNA 정량화 방법에 관한 것이다.
- [0034] 본 발명에 따르면 시료 내 cfDNA 분자의 메틸화 정도를 수치화하여 정량화된 정보를 제공할 수 있다. cfDNA의 메틸화는 조직, 세포마다 특이한 패턴을 가지므로 이러한 정보를 이용하여 상태 식별을 원하는 타겟 조직, 세포에 대한 정보를 얻을 수 있다.
- [0035] 본 발명에 있어서, "메틸화"는 cfDNA를 구성하는 염기에 메틸기가 부착되는 것을 의미하며 DNA 상의 CpG 사이트의 사이토신에서 일어나는 메틸화 여부 또는 메틸화 정도를 의미할 수 있다. DNA 메틸화는 유전자 발현 조절에 중추적인 역할을 하는 후성유전적 변형으로서, 사이토신의 5번 탄소에 메틸기(-CH₃)가 붙어 5-메틸사이토신 형태로 변형된다. DNA 메틸화는 주로 CpG 디뉴클레오타이드의 사이토신에서 발생하며, CpG가 집중적으로 몰려있는 DNA 영역은 CpG 섬이라고 불리며 유전자의 스위치에 해당하는 프로모터 위치에 이런 구조가 집중되어 있어 유전자 발현 조절에 주요 기능으로 연구되고 있다. 생체 내 게놈은 모두 동일하지만, 게놈상의 유전자가 발현되는 시기, 위치, 정도에 따라 세포와 조직에서 기능이 달라지게 된다. DNA 메틸화는 신체를 이루고 있는 각각의 조직, 세포에서 알맞은 기능을 하기 위해 조직과 세포마다 특이적인 메틸화 패턴을 가진다. 본 발명에서는 이러한 생물학 원칙을 근거하여 액상 시료 내 존재하는 특정 조직 또는 세포에서 유래된 cfDNA의 메틸화 패턴을 이용하여 타겟 조직 또는 세포에서 유래된 cfDNA를 정량화 하는 방법을 제공하고, 이를 이용한 건강상태 예측, 질병 예측 및 질병 상태 예측에 응용하는 방법을 제공한다.
- [0036] DNA 메틸화 정보를 활용하는 질병 연구의 한 예로써 암질환을 살펴보면, 정상적인 세포의 DNA에는 전체적으로 사이토신에 메틸화가 되어 있지만, 기능을 하는 프로모터 부위에는 거의 메틸화가 되어 있지 않다. 한편 암이 진행되는 단계에서 프로모터 부위에 집중적으로 메틸화가 진행되고(Promoter hypermethylation), DNA 전체적으로는 메틸화가 오히려 감소되는 것(Global hypomethylation)으로 알려져 있다. 특히 프로모터의 과메틸화(hypermethylation)는 종양억제 유전자의 스위치를 끄게 되므로 암이 발생하고, DNA 전체의 저메틸화(hypomethylation)은 DNA 구조의 불안정성을 가중시켜 암을 일으키게 되는 것으로 알려져 있다. 그렇지만 암 중에 따라, 암의 진행단계에 따라 메틸화 양상이 다양하게 나타난다. 암 세포의 DNA 메틸화는 암 발생의 초기에

일어나는 것으로 보고되어 있어, 암의 진단, 특히 조기 진단에 유용한 타겟이 될 수 있는 것으로 알려져 있다. 특히 암종마다 관여하는 종양 억제 유전자의 메틸화가 암의 진행단계에서 중요한 요소로써 연구되고 있다. 뿐만 아니라, DNA 메틸화는 질병치료의 예후나 약물반응성, 재발 여부를 관찰하는데도 널리 연구되고 있다.

[0037] 최근에는 조직 또는 세포 특이적 DNA 메틸화 특성을 이용하여 노화나 치매, 대사증후군, 비만 등과 같은 건강관리 측면에서 많은 연구들이 진행되고 있다. 뿐만 아니라, 산모의 영양과 환경적 요인은 태아의 DNA에 영향을 주는 것이 잘 알려져 있는데, 특히 태아 상태에서 산모의 환경적 변화가 태아가 태어났을 때의 질병 감수성에 영향을 미친다는 연구들이 활발히 보고되고 있으며, 여기에 핵심이 되는 기전이 DNA 메틸화이다.

[0039] 이와 같이 기본적으로 장기별, 조직별, 세포별로 다른 DNA 메틸화 특성(Tissue specific methylation)을 바탕으로, 최근 액체 생검기술의 발달과 함께 체액 내 cfDNA의 메틸화 정도를 측정하여 cfDNA가 체액으로 유리된 근원 조직 또는 세포를 찾는 기술 (Tissue of origin, TOO)들이 발달하고 있으며, 이에 따라 질병 진단을 포함하는 건강관리 기술들이 활발히 개발되고 있다. 많은 연구들에서 조직 DNA와 체액 cfDNA의 상관관계가 통계적 유의성이 있음을 보고한 바 있다.

[0040] 따라서 본 발명의 새로운 타겟 유래 cfDNA 정량화 방법을 이용하여 도출되는 cfDNA 정보를 이용하면 cfDNA 메틸화와 관련성이 보고되거나, 보고될 각종 질병 및 건강 상태를 식별할 수 있다.

[0041] 본 발명에 있어서, 조직 또는 세포 종류에 특이적으로 나타나는 DNA 메틸화 상태가 알려져 있거나, 도출될 수 있는 CpG 영역을 “타겟 특이적 CpG 마커 사이트” 로 정의하였으며, 이 중 특히 질병 상태를 반영할 수 있는 질병 특이적으로 메틸화 차이를 보이는 CpG 마커 사이트는 “질병 특이적 CpG 마커 사이트” 로 정의하였다. 상기 타겟 특이적 CpG 마커 사이트, 질병 특이적 CpG 마커 사이트는 하나 또는 인접하여 존재할 수 있으며, CpG 마커 사이트를 하나 또는 여러 개 포함하는 영역을 “타겟 특이적 영역” 이라고 정의한다. 따라서 본 발명의 타겟 특이적 영역은 하나의 CpG 사이트 또는 인접한 여러 개의 CpG 사이트로 이루어질 수 있으며, 특정 범위로 제한되는 것은 아니다.

[0043] 타겟 특이적 CpG 마커 사이트는 특정 조직 또는 세포, 질병, 건강 상태에 대한 정보를 제공할 수 있는 마커로 활용될 수 있는 부위이다. 타겟 특이적 CpG 마커 사이트가 유래된 조직 또는 세포 종류에 대한 정보를 포함하고 있는 경우, cfDNA 메틸화 정량화를 통해 유래된 조직 또는 세포에 대한 정보를 제공할 수 있으며, 질병 특이적 정보를 포함하고 있는 경우 시료의 건강상태, 즉 질병에 대한 정보를 제공할 수 있다.

[0044] 상기 타겟 특이적 CpG 마커 사이트는 종래 공지된 DB, 문헌에서 수득할 수 있는 정보일 수 있으며, 다음과 같은 방법으로 수득될 수 있다.

[0046] (a) 유전체 메틸화 정보를 제공하는 공용 데이터베이스에서 질병 조직의 후성 유전체 영역의 메틸화 정량값을 도출하는 단계; (b) 유전체 메틸화 정보를 제공하는 공용 데이터베이스에서 정상 조직의 후성 유전체 영역의 메틸화 정량값을 도출하는 단계; (c) 상기 (a) 및 (b) 단계에서 도출된 질병 조직 및 정상조직의 메틸화 정량값의 차이를 도출하는 단계; (d) 유전체 메틸화 정보를 제공하는 공용 데이터베이스에서 정상 체액 특이적 후성유전체 영역의 메틸화 정량값을 도출하는 단계; (e) 체액의 노이즈 시그널을 제거하는 단계; 및 (f) 상기 (c) 단계에서 도출된 메틸화 정량값의 차이; 및 (d) 단계에서 도출된 정상 체액 특이적 후성유전체 영역의 메틸화 정량값이 미리 설정된 각각의 임계값을 초과하는 후성 유전체 영역을 타겟 특이적 CpG 마커 사이트로 선택하는 단계. 이때 상기 공용 데이터베이스(DB)는 TCGA(The Cancer Genome Atlas) 데이터베이스, ENCODE(Encyclopedia Of DNA Elements), 또는 GEO(Gene Expression Omnibus)의 데이터베이스를 사용할 수 있으나, 특정 DB에 국한되지 않고, 후성 유전체 영역의 메틸화 정보를 담고 있는 한, 기 보고된 논문의 데이터 및 공용 데이터베이스를 제한없이 사용 가능하다.

[0048] 본 발명에서는 상기 알려진 또는 도출된 타겟 특이적 CpG 마커 사이트에 검사 대상 생물학적 시료 유래 cfDNA로 이루어진 서열 DNA, 리드를 정렬하는 단계를 포함한다.

[0049] 본 발명에서 "리드(reads)"는, 당업계에 알려진 다양한 방법을 이용하여 서열정보를 분석하여 추출되는 핵산 단편정보를 의미한다. cfDNA의 추출은 당업계에 공지된 방법에 따라 수행될 수 있으며, 특정 기술에 한정하지 않고, 본 발명의 일 구현예에서는 QIAamp Circulating Nucleic Acid Kit (QIAGEN, 55114) 또는 Apostle MiniMax™ High Efficiency Isolation Kit (Beckman Coulter Life Sciences, C40603) 를 이용하여 제조사 프로토콜에 따라 추출하였다.

[0050] cfDNA 메틸화 정보는 메틸화 정도를 측정할 수 있는 기술인 차세대 염기서열 분석법, 장서열 분석법, 패널, Chip, PCR 등의 방법들이 사용될 수 있으며, 특정 기술에 한정하지 않고, DNA 메틸화 정보를 도출할 수 있는 측

정법을 제한없이 사용 가능하다.

- [0052] 본 발명은 타겟 특이적 CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하고, 정렬된 DNA 중 상기 타겟 특이적 영역 내 존재하는 DNA의 메틸화 밀도(DNA Methylation Density; DMD)를 도출하는 단계를 포함한다.
- [0053] 타겟 특이적 영역은 타겟 특이적 CpG 마커 사이트를 기준으로 미리 설정된 영역일 수 있으며, 그 길이는 목적에 따라 적절하게 설정될 수 있다. 예컨대 본 발명의 바람직한 일 구현예에서는 타겟 특이적 CpG 마커 사이트를 기준으로 상류 100bp, 하류 100bp 부위를 포함하는 영역을 타겟 특이적 영역으로 정의하여 DNA 메틸화 밀도 계산에 사용하였으나, 이에 제한되는 것은 아니다.
- [0054] 각 CpG 마커 사이트를 기준으로 인접한 주변 DNA 영역을 타겟 특이적 영역으로 정의하고, 정렬된 DNA 중 상기 타겟 특이적 영역 내 존재하는 DNA의 메틸화 밀도 (DNA Methylation density, DMD)를 도출할 수 있다.
- [0055] 본 발명에서 정의하는 DNA 메틸화 밀도 (DMD)는 단일 CpG 사이트뿐만 아니라, 리드(read) 및 단편(fragment)의 메틸화 밀도(methylation density), DNA의 메틸화 밀도(methylation density)를 포함하고, DNA 한 분자 내에 존재하는 일부 또는 전체 영역에 대해서 메틸화된 CpG의 수를 전체 CpG 수로 나눈 값인, 각 DNA의 메틸화 정도의 수치화 방법을 의미한다. DMD 값을 측정하는 방법은 메틸화 정도를 측정할 수 있는 기술인 차세대 염기서열 분석법, 장서열 분석법, 패널, Chip, PCR 등을 포함할 수 있으며, 이에 한정되지 않는다. 따라서 DNA 메틸화 밀도 (DMD)는 DNA 한 분자 내에 존재하는 일부 또는 전체 영역에 대하여 메틸화된 CpG의 수를 전체 CpG 수로 나눈 값일 수 있다.
- [0056] 예컨대 DNA 조각에 포함된 전체 CpG가 메틸화된 경우, DMD는 1의 값을 가지고, 전체 CpG가 비메틸화된 경우 0의 값을 가지는 것으로 정의된다. 일 예로 PCR로 모두 메틸화된 DNA를 타겟하는 경우 DMD가 1에 해당하는 DNA를 정량화 하는 것과 동일하고, 모두 메틸화되지 않은 DNA를 타겟하는 경우는 DMD가 0에 해당하는 DNA를 정량화 하는 것과 동일하다.
- [0058] 본 발명의 3) 단계는 상기 타겟 특이적 영역에 존재하는 DNA 중 2) 단계에서 도출된 DMD 값이 역치값을 초과하는 DNA를 ctDNA(circulating-target DNA) 후보로 정의하고 그 수를 계수하여 정규화하는 단계; 이다.
- [0059] 본 발명에 있어서, “circulating-target DNA(ctDNA)”는 체액 내 존재하는 cfDNA 중 타겟 조직 또는 세포에서 시료, 예컨대 혈액과 같은 액체 시료로 유리된 cfDNA로 정의하였고, 타겟 조직 또는 세포는 혈구세포, 암세포, 질병조직, 정상조직 등 신체를 구성하는 핵산을 가진 세포 또는 그 집단을 의미하며, 특정 조직 또는 세포에 한정하지 않는다. 즉, ctDNA는 체액 내 존재하는 cfDNA 중 타겟 조직 또는 세포에서 유리된 cfDNA를 의미할 수 있다.
- [0060] DMD 값이 역치값을 초과하는 DNA는 ctDNA 후보군이 될 수 있다. 상기 정규화는 각 타겟 특이적 영역 상의 ctDNA 후보 수를 백만개의 맵핑된 리드 당 수(counts per million mapped reads, CPM)를 이용하여 정규화하는 방법으로 수행될 수 있다.
- [0061] 상기 DMD의 역치값은 미리 설정된 값일 수 있으며, 목적하는 진단, 식별, 검출에 따라 적절하게 설정될 수 있다. 본 발명의 일 구현예에서는 DMD의 역치값을 0.6으로 설정하여 ctDNA 후보군 수를 도출하는 방법을 도 2a에 도시하였다.
- [0063] 본 발명에서는 각 타겟 특이적 영역에서 도출된 정규화된 ctDNA 후보 수의 평균을 ctCandi (circulating-target DNA Candidate index)로 도출하는 단계를 포함한다.
- [0064] 상기 ctCandi는 본 발명의 발명자에 의하여 최초로 제안되는 개념으로 각 DNA 한 분자 내에 존재하는 일부 또는 전체 영역의 메틸화 정보를 스코어링한 값이며, 개인별 타겟 조직 또는 세포유래 DNA 단편의 양을 수치화하여 나타낼 수 있다. 예컨대 본 발명의 일 구현예에서는 6243개의 CpG 마커 사이트 각각에 존재하는 정규화된 ctDNA 후보수를 도출하고 이들의 평균을 계산하여 ctCandi를 도출하였다.
- [0066] 본 발명의 ctCandi(circulating-target DNA Candidate index)를 이용하면, 타겟을 어떤 조직 또는 세포로 선정하는지에 따라 생물학적 시료 유래 cfDNA가 어떤 조직 또는 어떤 세포로부터 유래된 것인지 다양하게 예측 및 분류할 수 있다. 즉, 본 발명은 5) ctCandi (circulating-target DNA Candidate index) 값에 대하여 cfDNA 디콘볼루션(deconvolution)을 수행하는 단계를 더 포함할 수 있다. 이때 예측 및 분류되는 대상은 정상 조직, 정상 세포, 질병 조직, 질병 세포와 같이 질병 유무 뿐만 아니라, 초기 설정되는 타겟 마커셋에 따라 타겟 조직 또는

세포에서 유래한 cfDNA 를 분류하는 것을 제한없이 포함한다.

- [0067] 예컨대 도출된 ctCandi 가 높을수록, ctDNA 가 많이 존재하는 것으로 판단할 수 있으므로 타겟 유래 cfDNA가 체액내 존재하는 양이 많은 것으로 예측해 볼 수 있으므로, 검출 대상 시료인 타겟 조직 또는 세포의 사멸 또는 손상 정도를 유추할 수 있다. 따라서 본 발명의 타겟 유래 cfDNA 정량화 방법은 검출 대상이 되는 생물학적 시료의 건강 상태나, 다양한 질병, 예컨대 암 유무를 진단하거나 질병 정도를 예측하는데 활용할 수 있다.
- [0068] 본 발명에서 검출 대상이 되는 생물학적 시료는 액체 생검 시료일 수 있으며, 정상인 또는 환자 유래의 혈액, 혈청, 혈장, 타액, 눈물, 소변, 대변, 질액, 소화액, 뇌척수액 및 콧물로 이루어진 군에서 선택된 1종 이상을 포함하는 신체에서 유래되는 모든 액체화될 수 있는 물질들을 포함할 수 있고 바람직하게는 혈액, 혈장, 소화액, 소변 또는 대변일 수 있다.
- [0070] 본 발명은 또다른 양태로 다음과 같은 cfDNA 를 이용한 건강 상태를 식별하는 방법을 제공한다:
- [0071] 1) 분석이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받는 단계; 2) 상기 CpG 마커 사이트 정보를 기준으로 상기 입력된 서열 DNA 정보를 정렬하고, CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하며, 상기 타겟 특이적 영역 내 존재하는 서열 DNA 부위의 DNA 메틸화 밀도 (DNA Methylation density, DMD)를 계산하는 단계; 3) 상기 타겟 특이적 영역에 존재하는 DNA중 계산된 DMD 값이 역치값을 초과하는 DNA수를 ctDNA 후보 수로 정의하여 정규화하는 단계; 4) 정규화된ctDNA 후보 수의 평균을 계산하여 ctCandi (circulating-target DNA Candidate index) 로 도출하는 단계; 5) 입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 학습된 머신러닝을 이용해 시료 제공자의 건강상태에 대한 확률값을 산출하는 단계; 및 6) 상기 산출된 확률값을 기준으로 상기 시료 제공자의 건강 상태를 식별하는 단계; 를 포함하고, 상기 학습된 머신러닝은 시료 제공자의 생체 정보와 도출된 ctCandi 정보를 입력으로 건강 상태에 대한 확률값을 출력하는 것인, 건강 상태를 식별하는 방법.
- [0072] 상기 건강상태는 시료의 질병 유무, 질병의 진행 정도, 예컨대 질병이 암인 경우 암의 병기에 대한 상태를 의미할 수 있다.
- [0073] 또한 상기 생체 정보는 시료 제공자의 성별, 나이, 질병 상태 등의 정보를 포함할 수 있다.
- [0074] 본 발명의 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트에 있어서 타겟은 조직 또는 세포를 의미할 수 있다. 예컨대 암 세포 특이적 CpG 마커로 추정된 CpG 마커 사이트를 본 발명의 방법에 이용하는 경우, 암에 관한 건강 상태를 식별할 수 있다.
- [0075] 본 발명에서 차세대 염기서열 분석법, 장서열 분석법, 패널, Chip, PCR와 같은 당 분야에 널리 이용할 수 있는 기술을 이용하여 측정된 시료 유래 cfDNA에서 얻을 수 있는 메틸화 정보를 서열 DNA 정보로써 입력받을 수 있으며, 메틸화 정도를 측정하여 도출되는 정보의 형태를 한정하지 않는다. 즉, 입력받는 cfDNA 메틸화 정보는, 공지의 DB 문헌의 정보를 활용하거나, 별도의 임상 시료를 이용하여 얻은 정보일 수 있다. 본 발명의 일 구현예에서는 종양에 대한 정보를 알고 있는 임상연구에서 수득된 임상시료 유래 cfDNA 및 한국 게놈 프로젝트(KGP)에서 수득된 정상 대조군 시료 유래 cfDNA를 차세대 염기서열 분석법, 예컨대 EM-seq (enzymatic methyl-sequencing)하여 각 시료의 cfDNA에 대한 정보를 생성하였다.
- [0076] 본 발명의 방법은 타겟 조직 또는 세포에 대한 메틸화 정보값을 갖는 CpG 마커 사이트에 대한 정보를 이용하여 수행될 수 있다. 즉, 상기 타겟 특이적 CpG 마커 사이트로 추정된 CpG 마커 사이트는, 타겟에 대한 정보를 제공할 수 있는 마커로 활용될 수 있는 부위이다. 타겟 조직 또는 세포에 대한 정보값을 갖는 CpG 마커 사이트에 대한 정보는 공지의 DB, 문헌의 정보를 활용하거나, 별도의 임상 시료를 이용한 학습과정을 통해 얻어진 정보를 활용할 수 있다.
- [0077] 본 발명에서는 도출된 ctCandi 정보를 이용하여 시료 제공자의 건강상태에 대한 확률값을 산출할 수 있다. 여기에서 건강 상태는 질병의 진단 또는 질병의 진행 정도(예: 암 병기)에 대한 확률을 의미할 수 있다.
- [0079] 상기 건강 상태에 대한 확률값은 5) 입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 학습된 머신러닝을 이용해 시료 제공자의 건강상태에 대한 확률값을 산출하는 단계; 및 6) 상기 산출된 확률값을 기준으로 상기 시료 제공자의 건강 상태를 식별하는 단계; 를 포함하고, 상기 학습된 머신러닝은 시료 제공자의 생체 정보와 도출된 ctCandi 정보를 입력으로 건강 상태에 대한 확률값을 출력하는 방법을 통해 출력될 수 있다.
- [0080] 보다 구체적으로 상기 확률값의 산출은 학습된 머신러닝을 이용하여 산출되며, 상기 학습된 머신러닝은 건강상

태에 대한 정보, 예를 들면 정상 또는 질병군에 대한 기존 보유 데이터를 학습 데이터로 이용하여 학습된 것일 수 있다. 이와 같이 학습된 머신러닝에 입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 하여 시료 제공자의 건강상태에 대한 확률값을 산출할 수 있다. 이때 시료 제공자의 생체 정보는 테스트 데이터로 이용될 수 있다.

[0081] 본 발명에 있어서 상기 머신러닝 알고리즘은 나이브 베이즈(Naive Bayes), KNN(K Nearest Neighbors), 랜덤 포레스트, 로지스틱 회귀 분석, 서포트 벡터 머신, 의사결정나무, 연관성 규칙 마이닝, 인공신경망, 선형 회귀, 순환 신경망 및 딥러닝 중 적어도 하나일 수 있다.

[0082] 본 발명에서는 ctCandi 스코어와 시료제공자의 생체정보를 입력값으로 하여 상호 관련성을 분석함으로써, ctCandi 스코어와 건강상태, 질병의 진단 또는 질병의 진행 정도에 대한 확률을 산출할 수 있다. 예컨대, 복수 개의 시료정보를 이용하여 ctCandi 스코어와 건강상태, 질병의 진단 또는 질병의 진행 정도에 대한 확률을 도출하고, 분석이 필요한 시료의 ctCandi 값을 도출하여 이를 상기 확률에 산입하여 시료의 ctCandi에 따른 건강상태, 질병의 진단 또는 질병의 진행 정도(예: 암 병기)에 대한 확률을 산출할 수 있다.

[0084] 또한 본 발명은 상기 건강 상태 식별 방법을 구현할 수 있는 장치를 제공한다.

[0085] 상기 건강 상태를 식별하는 장치는 분석이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받는 입력부; 상기 CpG 마커 사이트를 기준으로 인접한 주변 영역을 타겟 특이적 영역으로 정의하며, 상기 입력된 서열 DNA 정보를 정렬하고, 상기 타겟 특이적 영역 내 존재하는 서열 DNA의 메틸화 밀도 (DNA Methylation density, DMD)를 계산하고, 상기 타겟 특이적 영역에 존재하는 서열 DNA 중 계산된 DMD 값이 역치값을 초과하는 DNA수를 ctDNA 후보 수로 정의하여 정규화하며, 정규화된 ctDNA 후보 수의 평균을 계산하여 ctCandi (circulating-target DNA Candidate index)로 도출하고, 입력받은 시료 제공자의 생체정보와 ctCandi 정보를 입력으로 학습된 머신러닝을 이용해 시료 제공자의 건강상태에 대한 확률값을 산출하며, 상기 학습된 머신러닝은 시료 제공자의 생체 정보와 도출된 ctCandi 정보를 입력으로 건강 상태에 대한 확률값을 출력하는 것이고, 상기 산출된 확률값을 기준으로 상기 시료 제공자의 건강 상태를 식별하는 프로세서; 및 상기 식별된 건강 상태 결과를 출력하는 출력부;를 포함하는 건강 상태를 식별하는 장치일 수 있다.

[0086] 본 발명 장치의 입력부를 통해 입력된 분석이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보는 프로세서를 통해 처리된다. 상기 프로세서는 분석이 필요한 시료 유래 cfDNA의 ctCandi 정보를 도출하며, 이와 같이 도출된 ctCandi 정보는 분석이 필요한 시료의 건강 상태에 대한 정보, 예컨대 질병의 유무 또는 질병의 진행 정도에 대한 정보를 출력부를 통해 제공할 수 있다. 예컨대 미리 결정된 기준 값 (임계값) 이하의 ctCandi 값을 갖는 분석이 필요한 시료는 정상 시료인 것으로 판별하거나, 미리 결정된 기준 값을 초과하는 ctCandi 값을 갖는 분석이 필요한 시료는 건강 상태가 정상범위에서 벗어나거나 질병으로 진단, 예측될 가능성이 높은 시료인 것으로 식별 또는 분류할 수 있다.

[0087] 상기 장치에 있어서, 상기 식별된 건강 상태 결과는 타겟 조직의 건강 상태, 질병 유무 진단 또는 질병의 진행 정도에 대한 정보일 수 있고, 암의 유무 또는 암의 병기에 대한 정보일 수 있다.

[0089] 또한 본 발명은 상기 cfDNA 메틸화 정보를 상호 보완하는 방법으로, cfDNA 단편의 다양한 정보들을 상호 교차 및 조합하여 활용하는 종양 유래 단편 (tumor originated fragment, TOF) 스코어링 방법을 이용한 암 진단 방법 및 암 진단 장치를 포함한다.

[0090] 보다 구체적으로 상기 암을 진단하는 방법은 다음을 포함한다:

[0091] 1) 진단이 필요한 시료 유래 cfDNA 로 이루어진 서열 DNA 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받아 상기 건강 상태를 식별하는 방법으로 산출된 시료 제공자의 건강상태에 대한 확률값을 데이터로 입력받는 단계; 2) cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델에서 도출된 스코어링 값과 질병 진단 정보의 상관관계에 대한 데이터를 입력받는 단계; 3) 상기 데이터를 정규화 또는 표준화하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어 값을 도출하는 단계; 및 4) 상기 TOF 스코어 값을 이용하여 상기 진단이 필요한 시료의 암 유무 또는 병기를 식별하는 단계;를 포함하는 암을 진단하는 방법.

[0092] 상기 cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델은 당 분야에 공지된 cfDNA 단편 특성과 암에 대한 정보를 입력 값으로 하여 도출된 정보를 포함하는 스코어링 모델을 제한없이 포함할 수 있다. 예컨대, 상기

cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델은 cfDNA 리드의 5' 말단의 4개 염기서열 모티프의 빈도를 표준화(Standardization)하고, 도출된 표준화 값과 암 진단에 대한 정보를 머신러닝 알고리즘을 통해 학습시켜, 상기 표준화된 값을 입력 데이터로 생성하는, 말단 모티프 특성을 이용한 스코어링 모델; cfDNA 단편의 SFR(short fragment ratio) 값을 표준화(Standardization)하고, 도출된 표준화 값과 암 진단에 대한 정보를 머신러닝 알고리즘을 통해 학습시켜, 상기 표준화된 값을 입력 데이터로 생성하는, SFR 특성을 이용한 스코어링 모델; Jagged end length, Preferred end coordinates, Oriented end density, Motif Diversity Score, Window Protection Score, cfDNA integrity 또는 Nucleosomal footprinting을 스코어링 하여 정규화 또는 표준화하고, 상기 도출된 값과 암 진단에 대한 정보를 머신러닝 알고리즘을 통해 학습시켜, 상기 도출된 값을 입력 데이터로 생성하는, cfDNA 단편특성을 이용한 스코어링 모델 중 하나 이상일 수 있다.

[0093] 본 발명의 TOF 스코어 값을 이용하여 암을 진단하는 방법은 복수개의 모델에서 생성되는 입력 데이터를 모두 통합하여 하나의 결과값인 TOF 스코어를 도출하는 것을 특징으로 한다. 본 발명에서 암의 진단을 위해 사용하는 TOF 스코어 값은 값은 상기 1) 및 2) 단계에서 입력받은 데이터 정보를 입력으로 한 머신러닝 알고리즘 학습을 통해 획득된 값일 수 있고, 각 개별 입력값을 정규화 또는 표준화하고, 이를 머신 러닝하여 얻어진 정보를 통해 획득된 값일 수 있다. 이 때 상기 머신러닝 알고리즘은 앙상블 학습 모델일 수 있다. TOF 스코어에 대한 모식도는 도 8에 나타내었다. 복수개 모델의 입력 데이터는 미리 결정된 가중치를 부여하여 처리될 수 있으며, 이를 통해 복수개의 cfDNA 단편 정보와 암 정보와의 관련성을 포함하는 TOF 스코어가 도출될 수 있다.

[0094] 본 방법의 일 구현예에서는 별도의 임상 시료를 이용한 학습과정을 통해 얻어진 정보를 활용할 수 있다. 본 발명의 일 구현예에서는 종양에 대한 정보를 알고 있는 임상시료 유래 cfDNA 및 한국 게놈 프로젝트(KGP)에서 수득된 정상 대조군 시료 유래 cfDNA를 차세대 염기서열 분석, 예컨대 EM-seq (enzymatic methyl-sequencing)하여 각 시료의 cfDNA에 대한 정보와 종양과의 관련성에 대한 정보를 생성하였다.

[0095] 상기 정규화 및 표준화는 당 분야에 공지된 데이터 정규화 및 표준화 방법을 제한없이 이용할 수 있으나, Min-Max scaling, MaxAbs scaling, Standard scaling, Robust scaling으로 이루어진 군에서 선택되는 1종 이상일 수 있다.

[0096] 상기 말단 모티프 특성을 이용한 스코어링 모델은, cfDNA의 5' 말단의 1~20개 염기서열 모티프는 256가지 조합이 존재할 수 있고, 건강한 사람과 질병 환자의 말단 모티프의 조합 패턴은 차이를 나타낸다는 사실에 기초하여 5' 말단 모티프 비율(ratio)을 스코어링하는 모델이다. 복수개의 cfDNA 리드의 5' 말단의 1~20개 염기서열 모티프의 빈도의 표준화 값과 암 유무 또는 암 진행정도에 대한 정보가 학습되며, 이를 통해, 특정 값(스코어)과 암 진단 정보의 관련성이 도출될 수 있다.

[0097] 상기 cfDNA 단편의 SFR(short fragment ratio) 특성을 이용한 스코어링 모델은, cfDNA 길이 프로파일은 통상 약 166bp에서 주요 피크를 나타내는 반면, 질병 환자의 cfDNA 길이는 건강한 대조군 대비 짧은 길이 값을 갖는다는 사실에 기초하여 스코어링하는 모델이다. SFR 스코어는 짧은 단편의 수를 긴 단편의 수로 나눈 값으로 정의되며 짧은 단편은 100bp 이상이면 150bp 이하, 긴 단편은 151bp 이상이면 220bp 보다 짧은 단편으로 정의된다. 복수개 시료의 SFR 스코어의 표준화 값과 암 유무 또는 암 진행정도에 대한 정보가 학습되며, 이를 통해, 특정 값(스코어)과 암 진단 정보의 관련성이 도출될 수 있다.

[0098] 상기 cfDNA 단편특성을 이용한 스코어링 모델은, cfDNA 단편과 그 집합이 갖는 형태 (Topology), 뉴클레오솜 구조, DNA 변이정보 등의 특성들이 건강한 사람과 질환자의 패턴차이를 나타낸다는 사실에 기초하여 다양한 cfDNA 특성 각각을 대변하는 스코어링 모델이다. 복수개 시료의 cfDNA 단편특성 스코어의 표준화 값과 암 유무 또는 암 진행정도에 대한 정보가 학습되며, 이를 통해, 특정 값(스코어)과 암 정보의 관련성이 도출될 수 있다.

[0100] 또한 본 발명은 상기 암을 진단하는 방법을 구현하는 암 진단 장치를 제공한다.

[0101] 본 발명의 암 진단 장치는 진단이 필요한 시료 유래 cfDNA 로 이루어진 서열 리드 정보, 시료 제공자의 생체정보 및 타겟 특이적 CpG 마커로 추정된 CpG 마커 사이트의 정보를 입력받아 상기 건강 상태를 식별하는 방법에 기재된 방법으로 산출된 시료 제공자의 건강상태에 대한 확률값 데이터; 및 cfDNA 단편 특성을 이용한 타겟 유래 단편 스코어링 모델에서 도출된 스코어링 값과 질병 진단 정보의 상관관계에 대한 데이터;를 입력받는 입력부; 상기 입력된 데이터들을 정규화 또는 표준화하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어 값을 도출하고 상기 TOF 스코어 값을 이용하여 상기 진단이 필요한 시료의 암 유무 또는 병기를 식별하는 프로세서; 및 상기 식별된 암 유무 또는 병기에 관한 결과를 출력하는 출력부를 포함할 수 있다.

[0103] 이하, 본 발명을 구체적인 실험예 및 실시예에 의해 설명하나, 본 발명의 범주는 이에 제한되지 않는다.

[0105] **실험예 1. 샘플 수집 및 가공**

[0106] 혈장 cfDNA를 추출하기 위하여, 암으로 진단받은 시점의 폐암 환자로부터 전혈을 Cell-Free DNA BCT tube (Streck)에 수집하였다. 일련의 과정은 CNUHH에서 승인한 IRB 프로토콜 (IRB no. CNUHH-2019-127)에 따라 수행하였다. 모든 환자 유래 시료들은 암 전문의에 의해 진단된 환자로부터 수득하였으며, 환자들은 모두 연구 프로젝트에 참여하는 것에 동의하였다. 혈액 시료에서 혈장을 분리하기 위하여 3,000rpm으로 10분 동안 4℃에서 원심분리하였으며, 혈장 내 세포 및 debris를 4℃에서 16,000g, 10분 조건의 추가적인 원심분리 과정을 통해 제거하였다. 혈장 시료는 이후 실험을 위하여 -80℃에서 보관하였다.

[0107] 암 전문의가 암으로 진단한 폐암 환자 유래의 조직 시료를 동남권 원자력 의학원(DIRAMS)에서 승인한 IRB 프로토콜 (IRB no. D-2012-014-002)에 따라 동남권원자력 의학원으로부터 수집하였다. 조직 DNA 추출을 위하여, 조직을 액체 질소에서 막자 사발로 분쇄하고 분말을 세포 용해 용액 (2% CTAB, 1.4 M NaCl, 100 mM Tris-Cl (pH 8.0), 20 mM EDTA, β -mercaptoethanol(사용 직전 첨가, 100ul/10 ml))에서 균질화하였다. 이 후 이들을 혼합하고 단백질 분해효소 K 를 첨가한 후 65℃에서 1.5 시간 동안 배양하였다. Phenol-chloroform-isoamylalcohol (25:24:1, PGI)을 동일한 부피로 용해물에 첨가하고 실온에서 10분 동안 12,000 rpm으로 원심분리하였다. 최상단 수용상을 분리하고 이에 1/12 부피의 5M NaCl 및 2 부피의 100% 에탄올을 첨가하였다. -20℃에서 30 분 동안 배양한 후, DNA 펠렛을 수집하고 원심분리하였다. 그 후 DNA 펠렛을 70% 에탄올로 세척하고 100ul의 이온교환 초정제수에 용해시켰다.

[0108] 건강한 대조군의 혈장 시료는 암으로 진단받은 적이 없고 호흡기 질환이 없는 환자로부터 수집하였다. 임신한 여성은 대조군에서 제외하였으며, 모든 대조군 시료는 UNIST 의IRB 승인 (IRB No.: UNISTIRB-21-66-A)에 따라 한국 계놈 프로젝트(KGP)에서 수집하였다. KGP에서 실온에서 10분동안, 1,500g 및 4℃에서 3,000g 로 10분 동안 원심분리를 거쳐 전혈로부터 혈장시료를 분리하였고, 실험에 사용하기 전까지 -80℃에서 보관하였다.

[0109] cfDNA 는 3 내지 5ml 의 혈장에서 QIAamp Circulating Nucleic Acid Kit (QIAGEN, 55114) 또는 Apostle MiniMax™High Efficiency Isolation Kit (Beckman Coulter Life Sciences, C40603) 를 이용하여 제조사 프로토콜에 따라 추출하였다. cfDNA 의 농도는 Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) 로 측정하였으며, cfDNA 의 퀄리티는 4150 TapeStation system (Agilent Technologies)를 이용하여 평가하였다. cfDNA의 순도가 80% 이상이고, 총 5ng 이상인 시료만 선별하여 이후 실험에 사용하였다.

[0111] **실험예 2. 메틸화 검출을 위한 시퀀싱 라이브러리 제작**

[0112] 제조사의 프로토콜에 따라 NEBNext®Enzymatic Methyl-seq Kit (NEB)를 이용하여 효소 변형된 DNA 라이브러리를 구축하였다. 먼저 5 내지 10ng의 cfDNA를 단편화하고, 말단 수선 및 A-테일링 후, 증폭 어댑터를 이용하여 결합시켰다. 그 후, 효소 변환의 첫번째 단계로 상기 어댑터 결합된 DNA를 TET2 및 산화 인헨서를 이용하여 산화시켜, 5'-메틸시토신과 5'-하이드록시메틸시토신을 이후에 수행될 탈아민화로부터 보호하였다. 이 후, 두번째 효소적 전환 단계로 APOBEC을 이용하여 모든 시토신을 우라실로 탈아미노화하여 전환하였다. 이 과정에서, 5'-메틸시토신과 5'-하이드록시메틸시토신은 탈아미노화되지 않았다. 다음으로 효소 변환된 DNA를 PCR을 통해 증폭하였고, DNA를 분리하였다. 모든 라이브러리는 정량 및 정성 분석을 거쳤다. 라이브러리의 Paired-end 150bp 리드를 Illumina Novaseq 6000 platform으로 시퀀싱하였다. Illumina NovaSeq 6000 system을 이용하여 cfDNA EMseq 라이브러리로부터 FASTQ 파일을 생성하였다. Illumina adapter sequences 및 폴리-g-테일은 fastp (0.20.1) 로 트리밍하였고, 평균 Phred 퀄리티 스코어가 20bp 보다 낮거나, 20bp 보다 짧은 저품질의 리드는 fastp(0.20.1) 로 필터링하였다. 이러한 공정을 통해 전처리된 리드는 Bismark(0.22.3)을 이용하여 bisulfite로 변환된 hg38 참조 게놈 서열에 대하여 정렬하였다. 정렬 단계에서, 정렬되지 않거나 고유하게 맵핑되지 않은 리드들은 Bismark (0.22.3)로 제거하였다. 최종적으로 중복 리드들을 MarkDuplicates (2.25.0)로 제거하였다. 이와 같이 전처리된 bam 파일들을 이후 분석에 이용하였다.

[0113] 한편 상기 EM-Seq에 사용된 139개의 NSCLC 혈장 시료는 전남국립대학교 화순병원(CNIHH)에서 NSCLC 환자 유래 혈장 시료를 수집하여 이용하였다. 대상자 선정 기준은 하기와 같다.

[0114] 1) 18세 이상의 남녀

[0115] 2) CNUHH의 바이오뱅크 기부 기증 동의서 서명

[0116] 3) 폐암 진단 목적으로 2019 년 8월부터 2021년 5월까지 CNUHH를 방문한 환자 및

[0117] 4) 조직학적 또는 세포학적으로 NSCLC 또는 SCLC 인 것으로 진단된 환자

[0118]

[0119] 대상자 제외 기준은 하기와 같다.

[0120] 1) 의료 기록이 접근 불가능한 환자

[0121] 2) 임신 또는 수유중인 여성

[0122] 3) 폐 병변이 조직학 또는 세포학적으로 악성으로 확인되지 않은 환자

[0124] 건강한 혈장 시료 (n=97)은 한국 게놈 프로젝트(KGP)에서 건강한 대조군 97명의 혈장 cfDNA 시료를 이용하였다. 건강한 대조군의 기준은 KGP 건강검진보고서 및 질환 자기보고 기준에 따랐다. 가능한 대조군과 실험 케이스의 나이를 매칭할 수 있도록 하였다.

[0126] **실험예 3. 바이오인포메틱스 및 통계 소프트웨어**

[0127] 데이터 분석을 위한 스크립트는 python ver. 3. 8.5를 이용하였다. Python package Pysam (0.16.0.1)를 사용하여 bam 파일에서 리드의 메틸화, 단편 크기 및 말단-모티프 특징을 계산하였다. Python package Pandas (1.4.2)는 테이블 형식 데이터를 구성하는데 활용하였다. Python package Scipy (1.6.1)를 피어슨 상관관계수 및 Wilcoxon Rank-Sum test를 포함한 통계 분석에 이용하였으며, python package Scipy (1.6.1)는 머신러닝 구축에 이용하였다. 모든 그래프 및 도면은 R package ggplot2 (3.3.6), grid (4.2.1) 및 data.table (1.14.2) packages를 이용한 R version 4.2.1. 를 이용하여 작성하였다.

[0129] 메틸화 패턴, 단편의 크기 및 말단 모티프 서열 유형에 대한 분석 결과를 활용하여 종양 유래 단편 (tumor originated fragment, TOF) 스코어를 도출하고, 이를 이용하여 환자에서 초기 폐암을 예측 및 진단하기 위한 새로운 진단 방법을 수립하였으며, 일련의 과정을 도 1에 나타내었다.

[0130] 이하에서 각 진단 모델 및 통합 모델인 TOF 스코어링 모델에 대하여 상세히 설명한다.

[0132] **실시예 1. cfDNA 메틸화 패턴을 이용한 진단 모델 구축**

[0133] **1.1 폐종양 특이적 CpG 마커 선발**

[0134] cfDNA 메틸화 패턴을 이용한 진단 모델을 구축하기 위하여, ctDNA 생리학적 특징을 고려하여 폐암 환자에서 리드-기반 ctDNA를 검출하고 계수하고자 하였다. 즉 ctDNA의 메틸화 정도를 스코어링하여 이를 질병 진단에 사용한다.

[0135] 먼저, 이를 위하여 공개된 DB인 TCGA(The Cancer Genome Atlas) 데이터베이스의 Infinium MethylationEPIC 어레이 데이터와 GEO(Gene Expression Omnibus)의 Infinium Methylation 450K 어레이 데이터를 사용하여 종양 특이적 메틸화 마커를 선별하였다. TCGA에서 219개의 폐 종양 조직 샘플과 190개의 정상 폐 조직 샘플을 비교 분석하여 종양 특이적 마커를 확인하였다. 종양 조직 샘플과 정상 조직 샘플 사이의 평균 베타 값의 차이가 0.15보다 높은 CpG 사이트를 초기에 선택하였다. 다음으로 GEO GSE40279의 건강한 혈액 샘플 656개를 이용하여 종양 조직 및 혈액 세포 모두에서 과메틸화된 CpG 사이트를 제외하였다. 또한, 건강한 혈액 샘플에서 평균 베타 값이 0.2보다 높은 CpG 사이트는 선택된 마커에서 제외하였다. 위의 과정에서 각 그룹의 샘플 중 절반에서 베타 값을 사용할 수 없는 CpG 사이트가 분석에서 제외되었다. 상기와 같은 과정을 거쳐 최종적으로 6243개의 CpG 사이트를 폐암 특이적 메틸화 마커로 선택하였다.

[0136] CpG 마커 선발의 과정을 도 1의 B CpG Marker selection에 나타내었다.

[0138] **1.2 ctDNA 후보 수 인덱스 구축 (ctCandi)**

[0139] 각 개인의 시퀀싱 리드에 대한 메틸화 상태를 이용하여 폐 종양 기원 단편의 수준을 추정할 수 있는 ctDNA 후보 수 인덱스(ctDNA candidate count index, ctCandi) 라는 새로운 cfDNA 메틸화 정량화 분석 방법을 고안하였다. ctCandi 도출 과정을 도 2a 및 도 1의 B 중간에 나타내었다.

[0140] 상기 1.1에서 선택된 6243개의 종양 특이적 메틸화된 CpG 마커 사이트에서 EM-seq 리드를 정렬하고 CpG사이트를 기준으로 상류 및 하류 100bp 부위에 위치한 리드들을 선별하여 ctCandi 계산에 사용하였다.

[0141] 실험예 2의 방법으로 제조되는 모든 건강한 혈장 시료 및 종양 유래 cfDNA에서 유래된 EM-seq(enzymatic methyl-sequencing) 리드는 ctDNA 후보를 포함한다. 각 ctDNA 후보들은 6243개의 폐종양 특이적 메틸화 CpG 마커 사이트에서 분석되고, 각 ctDNA 후보군 수는 백만개의 맵핑화된 리드 당 수로 표준화되었다. 보다 구체적인

로 선택된 6243 개의 폐종양 특이적 메틸화된 CpG 마커 사이트에서 ctDNA 후보군을 포함하는 EM-seq 리드를 정렬하였다. 선택된 폐종양 특이적 메틸화 마커 위치에서 상류 및 하류 100bp를 포함하는 영역의 각 리드에 존재하는 메틸화 밀도를 계산하였다. 각 리드 내 CpG 사이트의 평균 메틸화 비율을 DMD (DNA Methylation density)로 정의하고, 0.6을 초과하는 DMD 값을 갖는 리드들은 ctDNA 리드 후보군으로 분류하였다. 각 CpG 사이트에 대한 후보 ctDNA 리드 수를 백만개의 맵핑화된 리드 당 수(counts per million mapped reads, CPM)을 이용하여 정규화하였다. ctCandi는 총 폐 종양조직 특이적 마커인 6243 개의 CpG 마커 사이트에 존재하는 정규화된 ctDNA 후보 수의 평균으로 계산하였다. 이러한 방법을 통해 개인별 질병 또는 정상 조직 유래 DNA 단편의 양을 수치화할 수 있으며, ctCandi 가 높을수록, ctDNA 가 많이 존재하는 것으로 판단할 수 있으므로 암 가능성이 높을 것으로 예측해 볼 수 있다. 일련의 ctCandi 도출의 예는 도 2a에 상세하게 기술되어 있다.

[0142] ctDNA 후보군의 수를 이용한 히트맵 결과를 도 2b의 (A)에 나타내었다.

[0143] 도 2b의 (A)에 나타난 바와 같이, 각 개인에 대하여 ctCandi 를 계산한 결과, NSCLC 군이 대조군과 비교하여 이전에 선택된 6243개의 CpG 사이트에서 평균적으로 더 많은 수의 ctDNA 후보군을 갖는다는 것을 확인하였다.

[0144] 정상 또는 암 병기에 따른 시험 시료에 대한 ctCandi의 boxplot 결과를 도 2b의 (B)에 나타내었다.

[0145] 도 2b의 (B)에 나타난 바와 같이 ctCandi 스코어는 NSCLC의 후기 병기에서 더 증가하는 것을 확인하였고, 건강한 대조군과 폐암 환자 사이에서 유의적인 차이를 나타내었다. P 값은 Wilcoxon rank sum test에 의해 계산하였다.

[0146] 추가적인 독립 검증 모델로, 동남권 원자력 의학원에서 제공받은 7개의 폐암 조직 샘플을 이용하였다. 모든 실험은 CNUHH의 IRB에서 승인되었다. ctCandi 스코어링 모델이 암과 정상군을 구분할 수 있는지를 확인하기 위하여, 조직 시료에 적용하였다. 먼저 ctCandi 점수를 시뮬레이션하고 예측하기 위하여 ctDNA 비율 측면에서 서로 다른 폐 종양 정도 수준 (0, 0.1%, 0.2%, 0.5% 및 1%)의 5개 세트를 구축하였다. 그 후 ctCandi 스코어링 방법을 이용한 종양 레벨 예측과 조직 시료의 결과가 잘 일치하는지 확인하였고, 그 결과를 도 2b의 (C)에 나타내었다.

[0147] 도 2b의 (C)에 나타난 바와 같이, ctDNA 비율을 이용한 *in silico*시뮬레이션 결과, 7개의 실제 폐 종양 시료에 대하여 높은 상관관계를 나타낸다는 것을 확인하였다: $R=0.95$, $P = 8.4 \times 10^{-8}$.

[0149] 실시예 2. cfDNA 메틸화 및 단편 특징을 이용한 진단 모델 구축

[0150] 건강한 대조군과 NSCLC 환자를 분류할 수 있는 머신 러닝 모델을 구축하기 위하여 다음 3가지 인자를 활용하였다: cfDNA의 메틸화, 말단-모티프 및 단편 크기 특징. 각 인자 기반 모델의 모식도를 도 3에 나타내었다.

[0151] TCGA 에는 SCLC 메틸화 데이터가 존재하지 않으므로 케이스 그룹으로 NSCLC 환자 시료를 이용하였다. 메틸화 모델의 독립적인 검증을 위하여 NSCLC 환자의 조직 시료는 시뮬레이션 테스트에서만 사용하였다. NSCLC 및 건강한 대조군 시료의 2/3 ($N=166$, 70%)는 기계 학습 모델 구축을 위한 트레이닝 세트로 사용하였고, 나머지 시료 ($N=70$, 30%)는 테스트 세트로 사용하였다. SCLC 환자 유래 혈장 시료는 적용가능성을 확장하기 위하여 TOF 스코어링 방법에 적용되었다. NSCLC 환자 유래 혈장 시료 및 SCLC 환자 유래 혈장 시료는 선정은 실험예 2에 기재된 바와 같다.

[0152] 모든 케이스 시료들은 조직학적으로 검사하였고, 다른 병기 단계에 따라 카테고리화하였다. NSCLC 시료들은 비슷한 비율로 선암종 및 편평세포 암종 하위 타입으로 분류되었다. 76개의 선암종 및 63개의 편평세포 암종이 확인되었다. 201 개의 암 시료 중 NSCLC에서 62건 및 SCLC에서 63건은 각각 IV 기 및 ED 로 분류된 전이성 환자였다.

[0153] 대조군으로, 한국 게놈 프로젝트(KGP)에서 건강한 대조군 97명의 혈장 cfDNA 시료를 이용하였다. 건강한 대조군의 기준은 KGP 건강검진보고서 및 질환 자기보고 기준에 따랐다. 가능한 대조군과 실험 케이스의 나이를 매칭할 수 있도록 하였다.

[0154] 환자의 혈액 또는 정상 조직 시료의 전체 게놈 시퀀싱을 수행하지 않았기 때문에, 상기 케이스들에 대해 중요한 분자 및 유전적 이상이 있는지에 대해서는 확인할 수 없었다. 모든 대조군 시료들은 전체 게놈 시퀀싱을 수행하였고, 유전적 이상을 발견하지 않았다.

[0155] 본 실험에 사용된 실험군의 정보를 표 1에 나타내었다.

[0157] [표 1]

	Model construction and validation			Independent test	
	NSCLC (n = 139)	Healthy Control (n = 97)	P value ^a	SCLC (n = 62)	Simulation test for ctCandi model NSCLC (n = 7)
Sample Type	Plasma	Plasma		Plasma	Tissue
Age, avg. (min, max)	69.2 (42,86)	65.6 (48,80)	3.012 × 10 ⁻⁴	69.2 (38,84)	70.9 (63,79)
Male, n (%)	92 (66.2)	65 (67)	1	54 (87.1)	5 (71.4)
Female, n (%)	47 (33.8)	32 (33)		8 (12.9)	2 (28.6)
Smoker, n (%)	20 (14.4)	6 (9.7)		10 (16.1)	5 (71.4)
Previous Smoker, n (%)	69 (49.6)	22 (25.5)		45 (72.6)	0 (0)
Never Smoking, n (%)	50 (36)	34 (34.8)		7 (11.3)	2 (28.6)
NA, n	0	35		0	0
Previous Cancer, n (%)	17 (12.2)	0 (0)		8 (12.9)	0 (0)
Cancer Subtypes	ADC 76 (54.7)				ADC 4 (57.1)
	SQC 63 (45.3)				SQC 3 (42.9)
Stage, n (%)	I 19 (13.7)			LD 23 (33.9)	I 4 (57.1)
	II 23 (16.5)			ED 36 (58.1)	II 0 (0)
	III 35 (25.2)				III 3 (42.9)
	IV 62 (44.6)				IV 0 (0)

^aP values were calculated for mean ages using a Wilcoxon rank sum test and for sex distribution using a χ^2 test. ctCandi, ctDNA candidate count index; NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer; ADC, Adenocarcinoma; SQC, Squamous cell carcinoma; LD, limited stage of SCLC; ED, extensive stage of SCLC.

[0158]

[0159]

67 개의 건강한 환자 및 99개의 NSCLC 환자의 학습/검증 데이터세트를 이용하여 각 cfDNA 특성에 기초한 로지스틱 회귀 분석 모델을 학습시키고, 검증하였다. Ridge regularization을 학습 단계에 적용하였다. 로지스틱 회귀 모델을 평가하기 위하여 5배 교차 검증(5-fold cross-validation)을 수행하였다.

[0161]

cfDNA 후보군 기반 머신 러닝 모델

[0162]

cfDNA 메틸화 특징에 기초한 분류 모델을 구축하기 위하여, 6243개의 CpG 사이트에서 실시예 1과 같이 정규화된 ctDNA 후보 수를 MinMax scaler를 이용하여 스케일링하고, 스케일링된 값을 입력 특징으로 이용하였다. 이때 스케일값은 최소 0 내지 최대 1이다.

[0163]

테스트 세트에서 ctDNA 후보군 수 기반 폐암 진단 머신러닝 모델의 암 진단 결과를 도 4의 A 및 B에 나타내었다.

[0164]

도 4에 나타난 바와 같이, ctDNA 후보군 수 기반 폐암 진단 모델은 0.908 의AUC값을 나타내어, 암 진단에 효과적임을 확인하였다. 도 4B에 따르면 후기 암 (II-IV) 및 건강한 대조군은 예측 값에서 통계적으로 유의한 차이를 나타내었다 ($P = 3.1 \times 10^{-7}$).

[0166]

단편 말단-모티프 기반 머신 러닝 모델

[0167]

특이적 단편 말단-모티프 (specific fragment end-motif) 특징에 기초한 분류 모델은 ctDNA 의 단편적 특성을 활용하여 암을 진단하고자 하는 모델이다. Emseq리드의 5' 말단에서부터 4개의 염기서열로 이루어지는 4mer 말단 모티프 빈도를 계산하였다. 4⁴인 256 개의 말단-모티프 빈도를 평균 0과 단위 분산의 스케일 값을 갖는 Standard scaler로 스케일링하였다. 상기 스케일링된 값을 분류 모델에서 입력 특징으로 사용하였다.

[0168]

테스트 세트에서 말단-모티프 기반 머신 러닝 모델의 암 진단 결과를 도 4의 C 및 D에 나타내었다.

[0169]

도 4C 및 D에 나타난 바와 같이, 건강한 대조군과 폐암 환자를 0.892 AUC 값으로 분류할 수 있음을 확인하였다.

[0171]

SFR(short fragment ratio) 기반 머신 러닝 모델

[0172]

단편 크기에 기초한 분류 모델을 구축하기 위하여, DELFI 방법에 따라 SFR(short fragment ratio)을 정의하였다. SFR 은 짧은 단편의 수를 긴 단편의 수로 나눈 값으로 정의된다. 여기에서 짧은 단편은 100bp 이상 이면서 150bp 이하, 긴 단편은 151bp 이상이면서, 220bp 보다 짧은 단편으로 정의하였다. Hg38 참조 게놈을 5Mb bins으로 쪼개고, 해당 bin에서 SFR을 계산하였다. 587 bin의 SFR 은 평균 0과 단위 분산의 스케일 값을 갖는 Standard scaler로 스케일링하였다. 상기 스케일링된 값은 분류 모델에서 입력값으로 사용하였으며, 테스트 세트에서 SFR-기반 머신 러닝 모델의 진단 결과를 도 4의 E 및 F에 나타내었다.

[0173]

도 4에 나타난 바와 같이, 테스트 세트에 대하여 획득한 SFR 기반 머신 러닝 모델의 평균 AUC 값은 평균 0.908 이었다. 또한 폐암 환자는 건강한 대조군과 비교하여 혈장 cfDNA의 짧은 단편 비율 (SFR) 이 상대적으로 높은 것을 확인하였다. 또한 암 분류 머신 러닝 모델에서 게놈-wide SFR 이 단독으로도 사용될 수 있음을 확인하였다.

[0175]

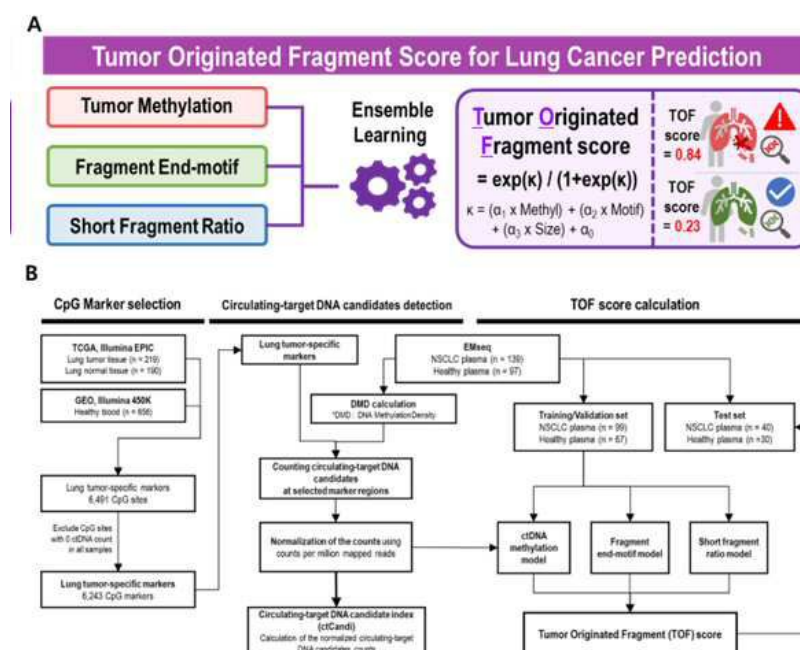
말단-모티프 및 단편 크기와 같은 cfDNA 단편학적 특성은 통계적으로 유의하게 정상 대조군 샘플과 초기 암 샘플에서 차이를 나타내었다($P = 1 \times 10^{-6}$ in the 5' end-motif-based model 및 $P = 0.0039$ in the SFR-based

model).

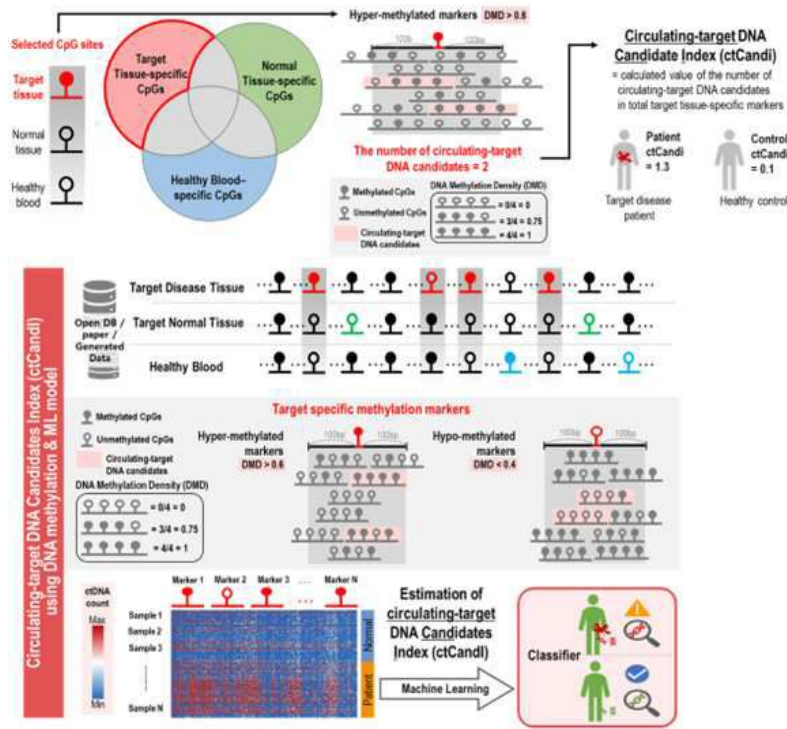
- [0176] 따라서 이러한 cfDNA 단편 특성을 이용한 머신 러닝 모델의 분류 결과와 ctDNA 메틸화 특성을 이용한 머신 러닝 분류 결과를 조합하면 초기 폐암 환자 분류를 포함하여 더욱 정확한 폐암 시료의 구분이 가능할 것으로 예상하였다.
- [0178] **실시예 3. 중앙 유래 단편 (Tumor originated fragment, TOF) 스코어링 모델의 구축**
- [0179] 본 발명의 3가지 cfDNA 특성에 기초한 각 모델의 결과를 통합하는 분류 모델을 구축하였다.
- [0180] 입력특징으로 세가지 모델의 확률을 이용한 학습 및 검증 데이터 세트로 로지스틱 회귀 모델을 학습시켰다. 각 모델의 구축에 5배 교차 검증을 수행하였기 때문에, 각 배수에 대한 5개의 통합된 로지스틱 회귀분석 모델을 얻었다. TOF 스코어는 5개 회귀 모델에서 예측된 값의 평균으로 정의하였다.
- [0181] TOF 스코어는 ctDNA 후보군 리드 수, 말단 모티프 및 SFR 특성에 대한 로지스틱 회귀 분석을 이용하는 앙상블 학습 모델에 기초하여 도출되었으며, 이에 기초한 암 진단 결과를 도 5에 나타내었다.
- [0182] 도 5의 A에 나타난 바와 같이, TOF를 이용한 암 진단 테스트에서 평균 AUC 값은 0.98을 나타내었다(민감도-95%, 특이도=96.7%). 특히 도 5의 B에 나타난 바와 같이 TOF 스코어는 T1 단계의 N0/M0 NSCLC 환자군도 대조군에서 유의적인 차이를 나타내었고 ($P = 0.004$), 이러한 결과는 초기 단계 암 환자의 진단에 TOF 스코어링 모델이 유용할 수 있음을 나타내는 결과이다. 또한 도 6에 나타난 바와 같이, TOF 스코어링은 건강한 대조군과 비교하여 NSCLC 아형인 선암종 (ADC) 및 편평세포 암종 (SQC)에서 유의적인 차이를 나타내었다.
- [0183] 한편 SCLC 포함하는 확장된 케이스 군에 TOF 스코어 컨셉을 추가적으로 적용하였으며, 도 5의 C에 나타난 바와 같이 놀랍게도 NSCLC 를 기초로 도출된 TOF 스코어링 모델이 SCLC에도 동일하게 적용될 수 있음을 확인하였다. SCLC에 적용된 모델에서 TOF 스코어는 limited stage 와 extensive stage를 포함하는 SCLC 환자군과 건강한 환자군을 구분할 수 있었다($P = 1.7 \times 10^{-12}$ 및 $P = 3.5 \times 10^{-19}$, 각각). 따라서 본 발명의 스코어링 모델이 광범위한 폐암 환자 분류에 적용될 수 있음을 확인하였다.
- [0184] 이와 같이 도출된 TOF 스코어를 기반으로 폐암 가능성에 대한 위험도를 예측할 수 있으며, 도 7에 예시적으로 나타난 바와 같이, 스코어 값 0 내지 1을 기준으로 0 내지 0.3은 질병 가능성이 낮고, 0.3 내지 0.4는 주의, 0.4 내지 1은 질병 위험성이 있는 것으로 예측할 수 있게 된다.

도면

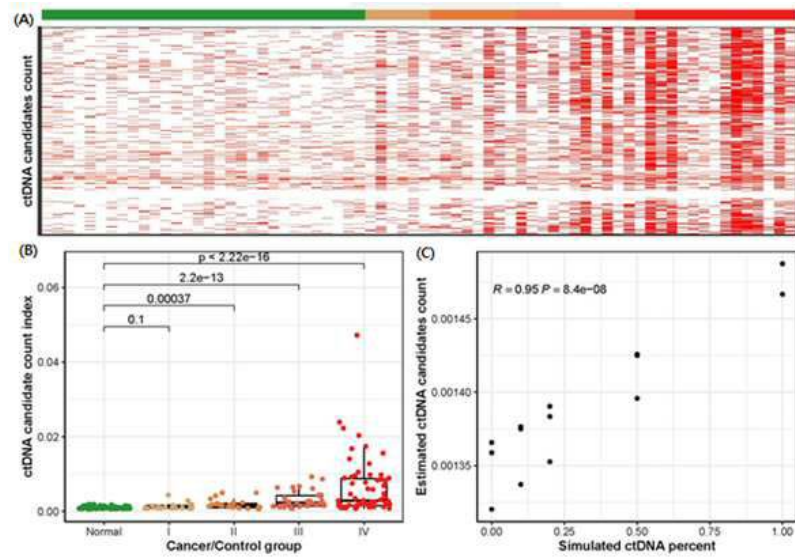
도면1



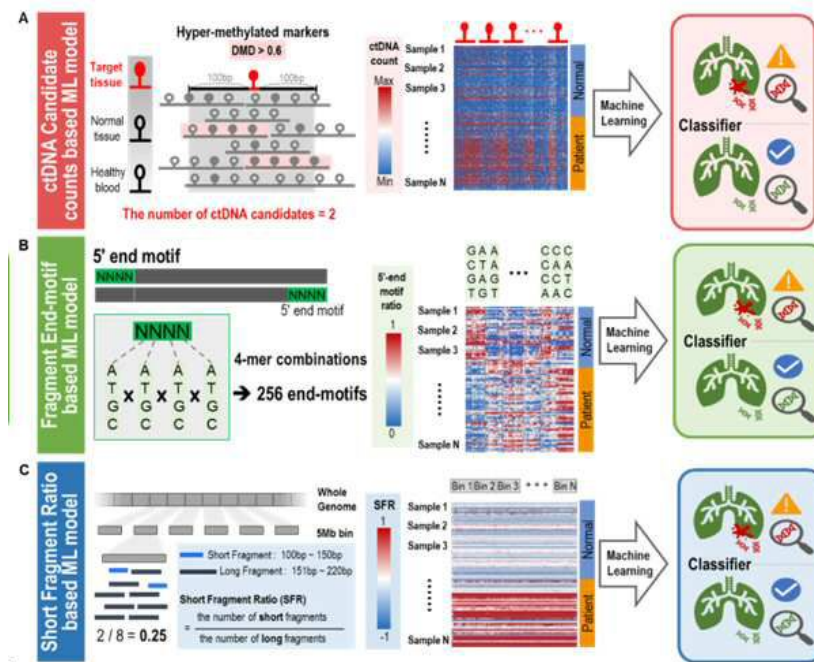
도면2a



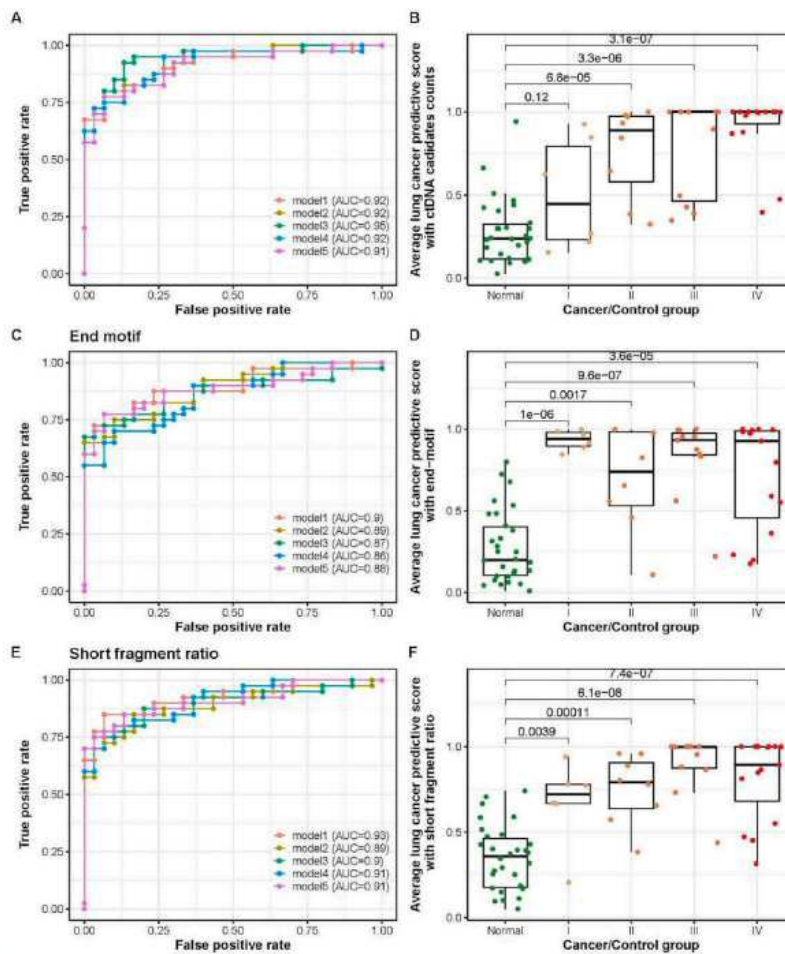
도면2b



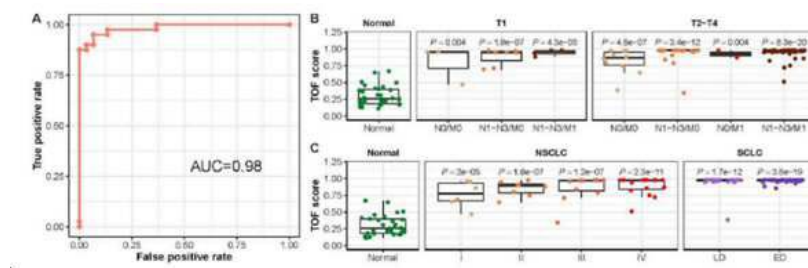
도면3



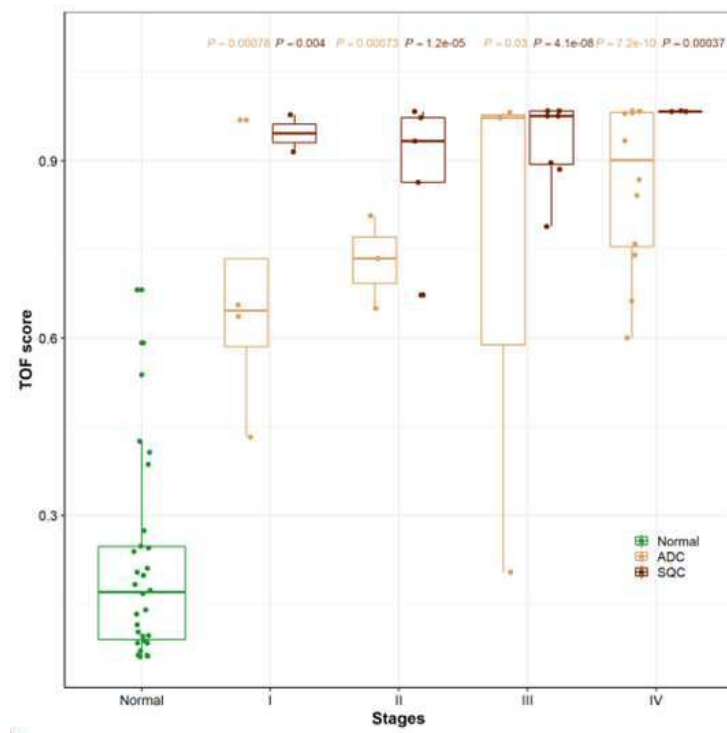
도면4



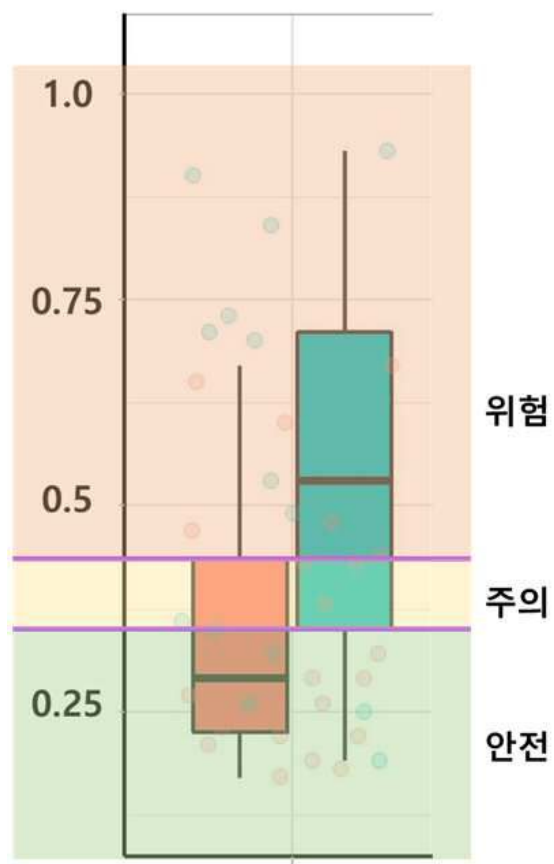
도면5



도면6



도면7



도면8

