(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0059423
(43) 공개일자 2023년05월03일

(51) 국제특허분류(Int. Cl.)
G16B 5/00 (2019.01) *C12Q 1/6886* (2018.01)
G06N 3/04 (2023.01) *G06N 3/08* (2023.01)
G16B 30/10 (2019.01) *G16B 35/10* (2019.01)
G16B 40/00 (2019.01)
(52) CPC특허분류
G16B 5/00 (2019.02)
C12Q 1/6886 (2022.01)
(21) 출원번호 10-2021-0143610
(22) 출원일자 2021년10월26일
심사청구일자 2022년12월05일

(71) 출원인
주식회사 지씨지놈
경기도 용인시 기흥구 이현로30번길 107 (보정동)
(72) 발명자
조은해
경기도 용인시 기흥구 이현로 30번길 107 (보정동)
안진모
경기도 용인시 기흥구 이현로 30번길 107 (보정동)
(뒷면에 계속)
(74) 대리인
이처영, 장제환

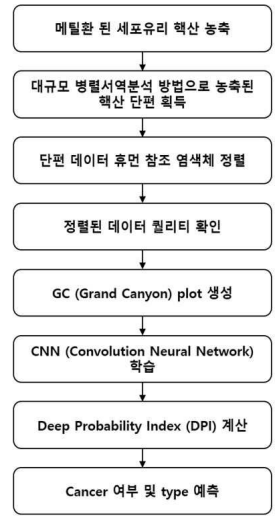
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 **메틸화된 무세포 핵산을 이용한 암 진단 및 암 종 예측방법**

(57) 요 약

본 발명은 메틸화된 무세포 핵산을 이용한 암 진단 및 암 종 예측방법에 관한 것으로, 보다 구체적으로는 생체시료에서 메틸화된 핵산을 추출하여, 서열정보를 획득하여 정렬한 리드를 기반으로 핵산단편의 벡터화된 데이터를 생성한 다음, 이를 학습된 인공지능 모델에 입력하여 계산된 값을 분석하는 방법을 이용한 암 진단 및 암 종 예측방법에 관한 것이다. 본 발명에 따른 메틸화된 무세포 핵산을 이용한 암 진단 및 암 종 예측방법은, 기존의 리드 개수(read count) 기반으로 염색체 양을 결정하는 단계를 이용하는 방식 또는 정렬된 리드(reads) 사이의 거리 개념을 이용하는 검출 방법 등에서 리드와 관련된 값을 하나하나의 정형화된 값으로 활용하는 데 비하여 벡터화된 데이터를 생성하여 AI 알고리즘을 이용하여 분석하기 때문에, 리드 커버리지가 낮더라도 유사한 효과를 발휘할 수 있어 유용하다

대 표 도 - 도1



(52) CPC특허분류
G06N 3/042 (2023.01)
G06N 3/08 (2023.01)
G16B 30/10 (2019.02)
G16B 35/10 (2019.02)
G16B 40/00 (2019.02)
(72) 발명자
기창석
경기도 용인시 기흥구 이현로 30번길 107 (보정동)
이준남
경기도 용인시 기흥구 이현로 30번길 107 (보정동)

명 세 서

청구범위

청구항 1

(a) 생체시료에서 핵산을 추출하여 메틸화 정보를 포함하는 서열정보를 획득하는 단계;

(b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

(c) 상기 정렬된 서열정보(reads) 기반의 핵산단편(fragments)을 이용하여 벡터화된 데이터를 생성하는단계;

(d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및

(e) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 2

제1항에 있어서, 상기 (a) 단계는 다음의 단계를 포함하는 방법으로 수행되는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법:

(a-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 및 이의 혼합물에서 메틸화 정보가 포함된 핵산을 수득하는 단계;

(a-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 수득하는 단계;

(a-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing) 라이브러리(library)를 제작하는 단계;

(a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

(a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계.

청구항 3

제2항에 있어서, 상기 (a-i) 단계의 메틸화 정보는 바이설파이트 전환법(bisulfite conversion) 또는 메틸화 DNA 면역침강법(Methylated DNA Immunoprecipitation, MeDIP)으로 수득한 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 4

제1항에 있어서, 상기 (c) 단계의 벡터화된 데이터는 그랜드 캐년 플롯(Grand Canyon plot, GC plot) 인 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 5

제4항에 있어서, 상기 GC plot은 정렬된 핵산단편의 염색체 구간 별 분포를 구간 별 수(count) 또는 핵산단편

(fragment) 사이의 거리를 계산하여 벡터화된 데이터로 생성하는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 6

제5항에 있어서, 상기 염색체 구간 별 분포를 핵산단편의 수로 계산하는 것은 하기의 단계를 포함하여 수행하는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법:

i) 염색체를 일정구간(bin)으로 구분하는 단계;

ii) 각 구간에 정렬된 핵산단편의 수를 결정하는 단계;

iii) 각 구간에 결정된 핵산단편 수를 샘플의 전체 핵산단편 수로 나누어 정규화(normalization)하는 단계; 및

iv) 각 구간의 순서를 X 축 값으로 하고, 상기 iii) 단계에서 계산한 정규화 값을 Y축 값으로 하여 GC plot을 생성하는 단계.

청구항 7

제5항에 있어서, 상기 염색체 구간 별 분포를 핵산단편 사이의 거리로 계산하는 것은 하기의 단계를 포함하여 수행하는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법:

i) 염색체를 일정구간(bin)으로 구분하는 단계;

ii) 각 구간에 정렬된 핵산단편 사이의 거리(Fragments Distance, FD)값을 계산하는 단계;

iii) 각 구간별로 계산된 거리값을 기반으로 각 구간의 거리의 대표값(RepFD)을 결정하는 단계;

iv) 상기 iii) 단계에서 계산된 대표값을 전체 핵산단편 사이의 거리 값의 대표값으로 나누어 정규화(normalization)하는 단계; 및

iv) 각 구간의 순서를 X 축 값으로 하고, 상기 iv) 단계에서 계산한 정규화 값을 Y축 값으로 하여 GC plot을 생성하는 단계.

청구항 8

제7항에 있어서, 상기 대표값은 핵산단편 사이의 거리의 합, 차, 곱, 평균, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대편차, 변동계수, 이들의 역수값 및 이들의 조합으로 구성된 군에서 선택되는 하나 이상인 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 9

제1항에 있어서, 상기 (d) 단계의 인공지능 모델은 정상인 벡터화된 데이터와 암이 있는 벡터화된 데이터를 구별할 수 있도록 학습하는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 10

제9항에 있어서, 상기 인공지능 모델은 합성곱 신경망(convolutional neural network, CNN), 심층 신경망(Deep Neural Network, DNN), 순환 신경망(Recurrent Neural Network, RNN) 및 오토 인코더(autoencoder)로 구성된 군에서 선택되는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 11

제9항에 있어서, 상기 인공지능 모델이 CNN이고, binary classification 을 학습할 경우, 손실함수는 하기 수식 1로 표시되며, 상기 인공지능 모델이 CNN이고, Multi-class classification을 학습할 경우, 손실함수는 하기 수식 2로 표시되는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법:

수식 1:

$$\text{loss(model(x),y)} = -\frac{1}{n}[\sum_{i=1}^{n}(y_i \log(model(x_i)) + (1-y_i)\log(1-model(x_i)))]$$

$\text{model}(x_i) = i\text{번째 input에 인공지능 model output}$

y = 실제 label 값

n = input data 수

수식 2:

$$\text{loss(model(x),y)} = -\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{c}(y_{ij}\log(model(x_i))_j)$$

$\text{model}(x_i)_j = i\text{번째 input에 j번째 인공지능 model output}$

y = 실제 label 값

n = input data 수

c = class 수

청구항 12

제1항에 있어서, 상기 (d) 단계의 인공지능 모델이 입력된 벡터화된 데이터를 분석하여 출력하는 결과값은 DPI(Deep Probability Index)값인 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 13

제1항에 있어서, 상기 (d) 단계의 기준값은 0.5이며, 0.5 이상일 경우, 암인 것으로 판정하는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 14

제1항에 있어서, 상기 (e) 단계의 출력 결과값 비교를 통해 암 종을 예측하는 단계는 출력 결과값 중, 가장 높은 값을 나타내는 암 종을 샘플의 암으로 판정하는 단계를 포함하는 방법으로 수행하는 것을 특징으로 하는 암 진단 및 암 종 예측을 위한 정보의 제공방법.

청구항 15

생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 해독하는 해독부;

해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부;

정렬된 서열 기반의 핵산단편을 이용하여 벡터화된 데이터를 생성하는 데이터 생성부;

생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부; 및

출력된 결과값을 분석하여 암 종을 예측하는 암 종 예측부를 포함하는 암 진단 및 암 종 예측 장치.

청구항 16

컴퓨터 판독 가능한 저장 매체로서, 암 진단 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,

(a) 생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 획득하는 단계;

(b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

(c) 상기 정렬된 서열정보(reads) 기반의 핵산단편을 이용하여 벡터화된 데이터를 생성하는단계;

(d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및

(e) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 통하여, 암 유무 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 메틸화된 무세포 핵산을 이용한 암 진단 및 암 종 예측방법에 관한 것으로, 보다 구체적으로는 생체시료에서 핵산을 추출하여, 메틸화 정보가 포함된 서열정보를 획득하여 정렬한 리드를 기반으로 핵산단편의 벡터화된 데이터를 생성한 다음, 이를 학습된 인공지능 모델에 입력하여 계산된 값을 분석하는 방법을 이용한 암 진단 및 암 종 예측방법에 관한 것이다.

배 경 기 술

[0003] 임상에서의 암 진단은 통상적으로 병력 조사, 물리적 검사 및 임상적 평가 후 조직 생검(tissue biopsy)을 수행하여 확인하고 있다. 임상 실험에 의한 암 진단은 암 세포의 수가 10억 개 이상이고 암의 직경이 1cm 이상일 경우에만 가능하다. 이 경우, 암 세포는 이미 전이능력을 가지고 있으며, 적어도 이들 중 반은 이미 전이된 상태이다. 또한, 조직생검은 침습적이어서 환자에게 상당한 불편함을 주고, 암 환자를 치료하다 보면 조직생검을 수행할 수 없는 경우도 자주 있다는 문제점이 있다. 이외에, 암 스크리닝에 있어서 암으로부터 직접 또는 간접적으로 생산되는 물질을 모니터링하기 위한 종양 마커가 사용되고 있지만, 암이 존재하는 경우에도 종양 마커 스크리닝 결과 반 이상이 정상으로 나타나고, 암이 없는 경우에도 자주 양성으로 나타나기 때문에, 그 정확성에 한계가 있다.

[0004] 이와 같은 통상적인 암 진단 방법의 문제점을 보완할 만한 비교적 간편하고 비침습적이며 높은 민감도 및 특이도를 가진 암 진단 방법의 요구에 따라, 최근 암의 진단, 추적 검사로 환자의 체액을 활용하는 액체생검(liquid biopsy)이 많이 이용되고 있다. 액체생검은 비침습적(non-invasive)인 방법으로, 기존의 침습적인 진단 및 검사 방법의 대안으로 주목 받고 있는 진단기술이다.

[0005] 최근에는 액체생검에서 획득한 세포 유리 DNA (cell free DNA)을 이용하여 암 진단 및 암 종 감별을 수행하는 방법이 개발되고 있으며(US 10975431, Zhou, Xionghui et al., bioRxiv, 2020.07.16.201350), 특히, 세포 유리 핵산의 메틸화 패턴을 이용하여 암 진단/종류를 결정하는 방법이 알려져 있다(Li, Jiaqi et al., bioRxiv, 2021.01.12.426440, US 2020-0131582, KR 10-2148547).

[0006] 한편, 인공 신경망이란 연결선으로 연결된 많은 수의 인공 뉴런들을 이용하여 생물학적인 시스템의 계산 능력을 모방하는 소프트웨어나 하드웨어로 구현된 연산모델을 나타낸다. 인공 신경망에서는 생물학적인 뉴런의 기능을 단순화시킨 인공 뉴런을 사용하게 된다. 그리고 연결강도를 갖는 연결선을 통해 상호 연결시켜 인간의 인지작용이나 학습과정을 수행하게 된다. 연결강도는 연결선이 갖는 특정 값으로, 연결가중치라고도 한다. 인공신경망의 학습은 지도 학습과 비지도 학습으로 나눌 수 있다. 지도 학습이란 입력 데이터와 그에 대응하는 출력 데이터를 함께 신경망에 넣고, 입력 데이터에 대응하는 출력 데이터가 출력되도록 연결선들의 연결강도를 갱신시키는 방법이다. 대표적인 학습 알고리즘으로는 델타규칙(Delta Rule)과 오류 역전파 학습(Back propagation Learning)이 있다. 비지도 학습이란 목표 값 없이 입력 데이터만을 사용하여 인공신경망이 스스로 연결강도를 학습시키는 방법이다. 비지도 학습은 입력 패턴들 사이의 상관관계에 의해 연결가중치들을 갱신시켜 나가는 방법이다.

[0007] 기계학습에서 적용되는 많은 데이터는 복잡해지고 차원이 늘어남에 따라 차원의 저주(curse of dimensionality)의 문제가 발생한다. 즉 이는, 필요한 데이터의 차원이 무한으로 갈수록 임의의 두 점간의 거리가 무한대로 발산하며 데이터의 존재량, 즉 밀도가 고차원의 공간에서는 다소 낮아져 데이터의 특성(Feature)을 제대로 반영하지 못하게 되는 것이다(Richard Bellman, Dynamic Programming, 2003, chapter 1). 최근 심층신경망(deep learning)의 발달은 입력층(input layer)과 출력층(output layer) 사이에 숨겨진 층(hidden layer)이 있는 구조로, 입력층으로부터 전달되는 변수 값의 선형 결합(linear combination)을 비선형 함수로 처리하면서 이미지, 영상, 신호데이터 등의 고차원의 데이터에서의 분류기(classifier)의 성능을 크게 향상시켰다고 보고되었다(Hinton, Geoffrey, et al., IEEESignal Processing Magazine Vol. 29.6, pp. 82-97, 2012).

[0008] 이러한 인공신경망을 이용하여 바이오 분야에 활용하는 다양한 특허(KR 10-2018-124550, KR 10-2019-7038076, KR 10-2019-0003676, KR 10-2019-0001741)가 존재하고 있으며, 본 발명자들은 혈액 내 무세포 DNA(cell-free DNA, cfDNA)의 서열분석 정보를 기반으로 인공신경망 분석을 통해 염색체 이상을 검출하는 방법에 대해 특허를 출원한 바 있다(KR 10-2021-0067931).

[0009] 하지만, 메틸화된 세포 유리 핵산의 정보를 이미지화하여 분석한 사례는 없으며, 또한 전장 유전체 단위의 메틸화된 패턴을 표현하는 사례는 없었다.

[0010] 이에, 본 발명자들은 상기 문제점들을 해결하고, 높은 민감도와 정확도의인공지능 기반 암 진단방법을 개발하기 위해 예의 노력한 결과, 메틸화된 무세포 핵산단편의 거리 또는 양에 기반하여 벡터화된 데이터를 생성하고, 이를 학습된 인공지능 모델로 분석할 경우, 높은 민감도와 정확도로 암 진단 및 암 종 판별을 수행할 수 있다는 것을 확인하고, 본 발명을 완성하였다.

발명의 내용

해결하려는 과제

[0012] 본 발명의 목적은 메틸화된 무세포 핵산을 이용한 암 진단 및 암 종 예측방법을 제공하는 것이다.

[0013] 본 발명의 다른 목적은 메틸화된 무세포 핵산을 이용한 암 진단 및 암 종 예측장치를 제공하는 것이다.

[0014] 본 발명의 또 다른 목적은 상기 방법으로 암 진단 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공하는 것이다.

과제의 해결 수단

[0016] 상기 목적을 달성하기 위하여, 본 발명은 (a) 생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 획득하는 단계; (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; (c) 상기 정렬된 서열정보(reads) 기반의 핵산단편(fragments)을 이용하여 벡터화된 데이터를 생성하는단계; (d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및 (e) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측을 위한 정보의 제공방법을 제공한다.

[0017] 본 발명은 또한, 생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 해독하는 해독부; 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부; 정렬된 서열 기반의 핵산단편을 이용하여 벡터화된 데

이터를 생성하는 데이터 생성부; 생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부; 및 출력된 결과값을 분석하여 암 종을 예측하는 암 종 예측부를 포함하는 암 진단 및 암 종 예측 장치를 제공한다.

[0018] 본 발명은 또한, 컴퓨터 판독 가능한 저장 매체로서, 암 진단 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되, (a) 생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 획득하는 단계; (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계; (c) 상기 정렬된 서열정보(reads) 기반의 핵산단편을 이용하여 벡터화된 데이터를 생성하는단계; (d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및 (e) 상기 출력값 비교를 통해 암 종을 예측하는 단계를 통하여, 암 유무 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체를 제공한다.

발명의 효과

[0020] 본 발명에 따른 메틸화된 무세포 핵산을 이용한 암 진단 및 암 종 예측방법은, 기존의 리드 개수(read count) 기반으로 염색체 양을 결정하는 단계를 이용하는 방식 또는 정렬된 리드(reads) 사이의 거리 개념을 이용하는 검출 방법 등에서 리드와 관련된 값을 하나하나의 정형화된 값으로 활용하는 데 비하여 벡터화된 데이터를 생성하여 AI 알고리즘을 이용하여 분석하기 때문에, 리드 커버리지가 낮더라도 유사한 효과를 발휘할 수 있어 유용하다

도면의 간단한 설명

[0022] 도 1은 본 발명의 인공지능 기반 염색체 이상을 판정하기 위한 전체 흐름도이다.

도 2는 본 발명의 일 실시예에 따라 메틸화된 cfDNA를 기반으로 생성한 GC plot의 예시로서, X축은 구간별 염색체이며, Y축은 각 구간에 해당되는 핵산단편 사이의 수를 의미한다.

도 3은 본 발명의 일 실시예에 따라, 메틸화된 cfDNA를 이용한 핵산단편 사이의 수를 기반으로 생성한 GC plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 neuroblastoma 판정의 정확도를 확인한 결과이며,

도 4는 본 발명의 일 실시예에 따라, 메틸화된 cfDNA를 이용한 핵산단편 사이의 수를 기반으로 생성한 GC plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 neuroblastoma 판정의 각 데이터 세트별 확률분포를 나타낸 결과로서, (A)는 Training set, (B)는 Validation set, (C)는 Test set을 의미한다.

도 5는 본 발명의 일 실시예에 따라 cfDNA를 기반으로 생성한 GC plot의 예시로서, X축은 구간별 염색체이며, Y축은 각 구간에 해당되는 핵산단편 사이의 수를 의미한다.

도 6은 본 발명의 일 실시예에 따라, cfDNA를 이용한 핵산단편 사이의 수를 기반으로 생성한 GC plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 neuroblastoma 판정의 정확도를 확인한 결과이며,

도 7은 본 발명의 일 실시예에 따라, cfDNA를 이용한 핵산단편 사이의 수를 기반으로 생성한 GC plot 이미지 데이터를 학습한 딥러닝 모델에 대하여 neuroblastoma 판정의 각 데이터 세트별 확률분포를 나타낸 결과로서, (A)는 Training set, (B)는 Validation set, (C)는 Test set을 의미한다.

발명을 실시하기 위한 구체적인 내용

[0023] 다른 식으로 정의되지 않는 한, 본 명세서에서 사용된 모든 기술적 및 과학적 용어들은 본 발명이 속하는 기술 분야에서 숙련된 전문가에 의해서 통상적으로 이해되는 것과 동일한 의미를 갖는다. 일반적으로 본 명세서에서 사용된 명명법 및 이하에 기술하는 실험 방법은 본 기술 분야에서 잘 알려져 있고 통상적으로 사용되는 것이다.

[0024] 본 발명에서는, 샘플에서 추출한 메틸화된 무세포 핵산에서 획득한 서열 분석 데이터를 참조 유전체에 정렬한 다음, 정렬된 핵산단편을 기반으로 벡터화된 데이터를 생성한 다음, 학습된 인공지능 모델에서 DPI값을 계산하여 기준값과 비교하여 암을 검출할 경우, 높은 민감도와 정확도로 암을 검출할 수 있다는 것을 확인하고자 하였다.

[0026] 즉, 본 발명의 일 실시예에서는, 혈액에서 추출한 DNA를 메틸화 정보가 포함되도록 시퀀싱 한 뒤, 참조 염색체에 정렬한 다음, 핵산단편 사이의 거리 또는 양을 일정한 염색체 구간 별로 계산하고, 각 유전영역을 X축으로 하고, 핵산단편 사이의 거리 또는 양을 Y축으로 하는 벡터화된 데이터를 생성한 다음, 이를 딥러닝 모델에 학습시켜 DPI 값을 계산하였으며, DPI 값이 기준값 이상일 경우, 암이 있다고 결정하고, 다수의 DPI 값 중, 가장 높은 값을 나타내는 암 종을 실제 암 종으로 결정하는 방법을 개발하였다(도 1)

[0027] 따라서, 본 발명은 일관점에서,

[0028] (a) 생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 획득하는 단계;

[0029] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

[0030] (c) 상기 정렬된 서열정보(reads) 기반의 핵산단편(fragments)을 이용하여 벡터화된 데이터를 생성하는단계;

[0031] (d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및

[0032] (e) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 포함하는 암 진단 및 암 종 예측을 위한 정보의 제공방법에 관한 것이다.

[0034] 본 발명에 있어서, 상기 핵산 단편은 생체시료에서 추출한 핵산의 조각이면 제한없이 이용할 수 있으나, 바람직하게는 세포 유리 핵산 또는 세포 내 핵산의 조각일 수 있으나, 이에 한정되는 것은 아니다.

[0035] 본 발명에 있어서, 상기 핵산 단편은 통상의 기술자에게 알려진 모든 방법으로 얻을 수 있으며, 바람직하게는 직접 서열분석하거나, 차세대 염기서열 분석을 통해 서열분석하거나 또는 비특이적 전장 유전체 증폭(non-specific whole genome amplification)을 통해 서열분석하여 얻거나, 프로브 기반 서열분석을 통해 얻을 수 있으나, 이에 한정되는 것은 아니다.

[0036] 본 발명에서 상기 핵산 단편은 차세대 염기서열 분석을 이용할 경우에는 리드를 의미할 수 있다.

[0038] 본 발명에서 상기 암은 고형암 또는 혈액암일 수 있으며, 바람직하게는 비호지킨 림프종 (non-Hodgkin lymphoma), 호지킨 림프종 (non-Hodgkin lymphoma), 급성 골수성 백혈병 (acute-myeloid leukemia), 급성 림프구성 백혈병 (acute-lymphoid leukemia), 다발성 골수종 (multiple myeloma), 경부암 (head and neck cancer), 폐암, 교모세포종 (glioblastoma), 대장/직장암, 췌장암, 유방암, 난소암, 흑색종 (melanoma), 전립선암, 간암, 갑상선암, 위암, 담낭암, 담도암, 방광암, 소장암, 자궁경부암, 원발부위불명암, 신장암, 식도암, 신경모세포종 및 중피종 (mesothelioma)으로 구성된 군에서 선택될 수 있으며, 더욱 바람직하게는 신경모세포종 (neuroblastoma)일 수 있으나, 이에 한정되는 것은 아니다.

[0040] 본 발명에 있어서,

[0041] 상기 (a) 단계는

[0042] (a-i) 혈액, 정액, 질 세포, 모발, 타액, 소변, 구강세포, 태반세포 또는 태아세포를 포함하는 양수, 조직세포 및 이의 혼합물에서 메틸화 정보가 포함된 핵산을 수득하는 단계;

[0043] (a-ii) 채취된 핵산에서 솔팅-아웃 방법(salting-out method), 컬럼 크로마토그래피 방법(column chromatography method) 또는 비드 방법(beads method)을 사용하여 단백질, 지방, 및 기타 잔여물을 제거하고 정제된 핵산을 수득하는 단계;

[0044] (a-iii) 정제된 핵산 또는 효소적 절단, 분쇄, 수압 절단 방법(hydroshear method)으로 무작위 단편화(random fragmentation)된 핵산에 대하여, 싱글 엔드 시퀀싱(single-end sequencing) 또는 페어 엔드 시퀀싱(pair-end sequencing) 라이브러리(library)를 제작하는 단계;

[0045] (a-iv) 제작된 라이브러리를 차세대 유전자서열검사기(next-generation sequencer)에 반응시키는 단계; 및

[0046] (a-v) 차세대 유전자서열검사기에서 핵산의 서열정보(reads)를 획득하는 단계;를 포함하는 것을 특징으로 할 수 있다.

[0047] 본 발명에서, 상기 (a) 단계의 서열정보를 획득하는 단계는 분리된 무세포 DNA를 1백만 내지 1억 리드 깊이로 전장 유전체 시퀀싱을 통해 획득하는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0049] 본 발명에 있어서, 상기 생체시료는 개체로부터 얻어지거나 개체로부터 유래된 임의의 물질, 생물학적 체액, 조직 또는 세포를 의미하는 것으로, 예를 들면, 전혈(whole blood), 백혈구(leukocytes), 말초혈액 단핵 세포(peripheral blood mononuclear cells), 백혈구 연층(buffy coat), (혈장(plasma) 및 혈청(serum)을 포함하는) 혈액, 객담(sputum), 눈물(tears), 점액(mucus), 세비액(nasal washes), 비강 흡인물(nasal aspirate), 호흡(breath), 소변(urine), 정액(semen), 침(saliva), 복강 세척액(peritoneal washings), 골반 내 유체액(pelvic fluids), 낭종액(cystic fluid), 뇌척수막 액(meningeal fluid), 양수(amniotic fluid), 선액(glandular fluid), 췌장액(pancreatic fluid), 림프액(lymph fluid), 흉수(pleural fluid), 유두 흡인물(nipple aspirate), 기관지 흡인물(bronchial aspirate), 활액(synovial fluid), 관절 흡인물(joint aspirate), 기관 분비물(organ secretions), 세포(cell), 세포 추출물(cell extract), 정액, 모발, 타액, 소변, 구강세포, 태반 세포, 뇌척수액(cerebrospinal fluid) 및 이의 혼합물을 포함할 수 있으나, 이에 한정되는 것은 아니다.

[0051] 본 발명에서 용어, ”참조집단”은 표준 염기서열 데이터베이스와 같이 비교할 수 있는 기준(reference) 집단으로, 현재 특정 질환 또는 병증이 없는 사람의 집단을 의미한다. 본 발명에 있어서, 상기 참조집단의 표준 염색체 서열 데이터베이스에서 표준 염기서열은 NCBI 등의 공공보건기관에 등록되어 있는 참조 염색체일 수 있다.

[0053] 본 발명에서, 상기 (a) 단계의 핵산은 무세포 DNA 일 수 있으며, 보다 바람직하게는 순환종양세포 DNA(circulating tumor DNA, ctDNA) 일 수 있으나, 이에 한정되는 것은 아니다.

[0054] 본 발명에 있어서, 상기 메틸화 정보가 포함된 핵산은 공지의 다양한 방법으로 수득할 수 있으며, 바람직하게는 바이설파이트 전환법(bisulfite conversion) 또는 메틸화 DNA 면역침강법(Methylated DNA Immunoprecipitation, MeDIP)으로 수득한 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0055] 본 발명에서, DNA 메틸화를 검출할 수 있는 방법은 제한효소 기반 검출 방법이 추가로 있는데, 이는 methylation restriction enzyme(MRE)를 이용하여 메틸화 되지 않은 핵산을 절단하거나, 메틸화 여부에 상관없이 특정 서열(recognition site)를 절단하여 hybridization 방법 또는 PCR과 결합해 분석하는 방법이다.

[0056] 본 발명에서 바이설파이트 치환에 기반한 방법은 Whole-Genome Bisulfite Sequencing(WGBS), Reduced-Representation Bisulfite Sequencing (RRBS), Methylated CpG Tandems Amplification and Sequencing (MCTA-seq), Targeted Bisulfite Sequencing, Methylation Array 및 Methylation-specific PCR (MSP) 등이 있다.

[0057] 본 발명에서, 메틸화 DNA를 풍부화(enrichment)하여 분석하는 방법은 Methylated DNA Immunoprecipitation Sequencing (MeDIP-seq), Methyl-CpG Binding Domain Protein Capture Sequencing (MBD-seq) 등이 있다.

[0058] 본 발명에서 메틸화 DNA를 분석할 수 있는 또 다른 방법은 5-hydroxymethylation profiling이 있으며, 그 예시로는 5hmC-Seal (hMe-Seal), hmC-CATCH, Hydroxymethylated DNA Immunoprecipitation Sequencing (hMeDIP-seq), Oxidative Bisulfite Conversion 등이 있다.

[0060] 본 발명에서, 상기 차세대 유전자서열검사기(next-generation sequencer)는 당업계에 공지된 임의의 시퀀싱 방법으로 사용될 수 있다. 선택 방법에 의해 분리된 핵산의 시퀀싱은 전형적으로는 차세대 시퀀싱(NGS)을 사용하여 수행된다. 차세대 시퀀싱은 개개의 핵산 분자 또는 고도로 유사한 방식으로 개개의 핵산 분자에 대해 클론으로 확장된 프록시 중 하나의 뉴클레오타이드 서열을 결정하는 임의의 시퀀싱 방법을 포함한다(예를 들어, 105개 이상의 분자가 동시에 시퀀싱된다). 일 실시형태에서, 라이브러리 내 핵산 종의 상대적 존재비는 시퀀싱 실험에 의해 만들어진 데이터에서 그것의 동족 서열의 상대적 발생 수를 계측함으로써 추정될 수 있다. 차세대 시퀀싱

방법은 당업계에 공지되어 있고, 예를 들어 본 명세서에 참조로서 포함된 문헌(Metzker, M. (2010) Nature Biotechnology Reviews 11:31-46)에 기재된다.

[0061] 일 실시형태에서, 차세대 시퀀싱은 개개의 핵산 분자의 뉴클레오타이드 서열을 결정하기 위해 한다(예를 들어, 헬리코스 바이오사이언스(Helicos BioSciences)의 헬리스코프 유전자 시퀀싱 시스템(HeliScope Gene Sequencing system) 및 퍼시픽바이오사이언스의 팩바이오 알에스 시스템(PacBio RS system)). 다른 실시형태에서, 시퀀싱, 예를 들어, 더 적지만 더 긴 리드를 만들어내는 다른 시퀀싱 방법보다 시퀀싱 단위 당 서열의 더 많은 염기를 만들어내는 대량병렬의 짧은-리드 시퀀싱(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer)) 방법은 개개의 핵산 분자에 대해 클론으로 확장된 프록시의 뉴클레오타이드 서열을 결정한다(예를 들어, 캘리포니아주 샌디에고에 소재한 일루미나 인코포레이티드(Illumina Inc.) 솔렉사 시퀀서(Solexa sequencer); 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재) 및 아이온 토렌트(Ion Torrent)). 차세대 시퀀싱을 위한 다른 방법 또는 기계는, 이하에 제한되는 것은 아니지만, 454 라이프 사이언스(Life Sciences)(코네티컷주 브랜포드에 소재), 어플라이드 바이오시스템스(캘리포니아주 포스터 시티에 소재; SOLiD 시퀀서), 헬리코스 바이오사이언스 코포레이션(매사추세츠주 캠브릿지에 소재) 및 에멀젼 및 마이크로 유동 시퀀싱 기법 나노 점적(예를 들어, 지누바이오(GnuBio) 점적)에 의해 제공된다.

[0062] 차세대 시퀀싱을 위한 플랫폼은, 이하에 제한되는 것은 아니지만, 로슈(Roche)/454의 게놈 시퀀서(Genome Sequencer: GS) FLX 시스템, 일루미나(Illumina)/솔렉사(Solexa) 게놈 분석기(Genome Analyzer: GA), 라이프(Life)/APG의 서포트 올리고(Support Oligonucleotide Ligation Detection: SOLiD) 시스템, 폴로네이터(Polonator)의 G.007 시스템, 헬리코스 바이오사이언스의 헬리스코프 유전자 시퀀싱 시스템(Helicos BioSciences' HeliScope Gene Sequencing system), 옥스포드 나노포어 테크놀로지스(Oxford Nanopore Technologies)의 PromethION, GriION, MinION 시스템 및 퍼시픽 바이오사이언스(Pacific Biosciences)의 팩바이오알에스(PacBio RS) 시스템을 포함한다.

[0064] 본 발명에서, 상기 (b) 단계의 서열 정렬은 컴퓨터 알고리즘으로서 게놈에서 리드 서열(예를 들어, 차세대 시퀀싱으로부터의, 예를 들어 짧은-리드 서열)이 대부분 리드 서열과 기준 서열 사이의 유사성을 평가함으로써 유래될 가능성이 있는 경우로부터 동일성에 대해 사용되는 컴퓨터적 방법 또는 접근을 포함한다. 서열 정렬 문제에 다양한 알고리즘이 적용될 수 있다. 일부 알고리즘은 상대적으로 느리지만, 상대적으로 높은 특이성을 허용한다. 이들은, 예를 들어 역동적 프로그래밍-기반 알고리즘을 포함한다. 역동적 프로그래밍은 그것들이 더 간단한 단계로 나누어짐으로써 복잡한 문제를 해결하는 방법이다. 다른 접근은 상대적으로 더 효율적이지만, 전형적으로 철저하지 않다. 이는, 예를 들어 대량 데이터베이스 검색을 위해 설계된 휴리스틱(heuristic) 알고리즘 및 확률적(probabilistic) 방법을 포함한다.

[0065] 전형적으로, 정렬 과정에 두 단계가 있을 수 있다: 후보자 검사 및 서열 정렬. 후보자 검사는 가능한 정렬 위치의 더 짧은 열거에 대해 전체 게놈으로부터 서열 정렬을 위한 검색 공간을 감소시킨다. 용어가 시사하는 바와 같이 서열 정렬은 후보자 검사 단계에 제공된 서열을 갖는 서열을 정렬시키는 단계를 포함한다. 이는 광역 정렬(예를 들어, 니들만-분쉬(Needleman-Wunsch) 정렬) 또는 국소 정렬(예를 들어, 스미스-워터만 정렬)을 사용하여 수행될 수 있다.

[0066] 대부분의 속성 정렬 알고리즘은 색인 방법에 기반한 3가지 유형 중 하나를 특징으로 할 수 있다: 해쉬 테이블(예를 들어, BLAST, ELAND, SOAP), 접미사트리(예를 들어, Bowtie, BWA) 및 병합 정렬(예를 들어, 슬라이더(Slider))에 기반한 알고리즘. 짧은 리드 서열은 정렬을 위해 전형적으로 사용된다.

[0067] 본 발명에 있어서, 상기 (b) 단계의 정렬단계는 이에 제한되지는 않으나, BWA 알고리즘 및 Hg19 서열을 이용하여 수행되는 것일 수 있다.

[0068] 본 발명에 있어서, 상기 BWA 알고리즘은 BWA-ALN, BWA-SW 또는 Bowtie2 등이 포함될 수 있으나 이에 한정되는 것은 아니다.

[0070] 본 발명에 있어서, 상기 b) 단계의 서열정보(reads)의 길이는, 5 내지 5000 bp이고, 사용하는 서열정보의 수는 5천 내지 500만개가 될 수 있으나, 이에 한정되는 것은 아니다.

[0072] 본 발명에 있어서, 상기 (c) 단계의 벡터화된 데이터는 정렬된 핵산단편을 기반으로 생성할 수 있는 벡터화된 데이터이면 제한없이 이용가능하며, 바람직하게는 그랜드 캐년 플롯(Grand Canyon plot, GC plot)인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0074] 본 발명에서 벡터화된 데이터는 이에 한정되지는 않으나 바람직하게는 이미지화된 것을 특징으로 할 수 있다. 이미지는 기본적으로 픽셀로 구성되는데, 픽셀로 구성된 이미지를 벡터화 시키면, 이미지의 종류에 따라서 1차원 2D 벡터(흑백), 3차원 2D 벡터(color(RGB)) 또는 4차원 2D 벡터(color(CMYK))로 표현될 수 있다.

[0075] 본 발명의 벡터화된 데이터는 이미지에 한정되지 않고, 예를 들어 n개의 흑백 이미지 여러 장으로 쌓아서 n차원의 2D 벡터(Multi-dimensional Vector)를 이용하여 인공지능 모델의 입력 데이터로 사용할 수 있다.

[0076] 본 발명에서 GC plot은 특정 구간을 (일정한 bin 또는 크기가 다른 bin) X축으로 두고 핵산단편 사이의 거리 또는 수와 같은 핵산단편으로 표현 할 수 있는 수치를 Y축으로 생성한 plot이다.

[0078] 본 발명에서, 상기 (c) 단계를 수행하기에 앞서 정렬된 핵산단편의 정렬 일치도 점수(mapping quality score)를 만족하는 핵산단편을 따로 분류하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.

[0079] 본 발명에서 상기 정렬 일치도 점수(mapping quality score)는 원하는 기준에 따라 달라질 수 있으나, 바람직하게는 15-70점, 더욱 바람직하게는 50~70점 일 수 있고, 가장 바람직하게는 60점일 수 있다.

[0081] 본 발명에 있어서, 상기 (c) 단계의 GC plot은 정렬된 핵산단편의 염색체 구간 별 분포를 구간 별 핵산단편의 수 또는 핵산단편 사이의 거리를 계산하여 벡터화된 데이터로 생성하는 것을 특징으로 할 수 있다.

[0082] 본 발명에서 핵산단편의 수 또는 핵산단편 사이의 거리 계산값을 벡터화하는 방법은 계산값을 벡터화하는 공지된 기술이면 제한없이 이용할 수 있다.

[0083] 본 발명에 있어서, 상기 정렬된 서열정보의 염색체 구간 별 분포를 핵산단편의 수로 계산하는 것은 하기의 단계를 포함하여 수행하는 것을 특징으로 할 수 있다:

[0084] i) 염색체를 일정구간(bin)으로 구분하는 단계;

[0085] ii) 각 구간에 정렬된 핵산단편의 수를 결정하는 단계;

[0086] iii) 각 구간에 결정된 핵산단편 수를 샘플의 전체 핵산단편 수로 나누어 정규화(normalization)하는 단계; 및

[0087] iv) 각 구간의 순서를 X 축 값으로 하고, 상기 iii) 단계에서 계산한 정규화 값을 Y축 값으로 하여 GC plot을 생성하는 단계.

[0089] 본 발명에 있어서, 상기 정렬된 서열정보의 염색체 구간 별 분포를 핵산단편 사이의 거리로 계산하는 것은 하기의 단계를 포함하여 수행하는 것을 특징으로 할 수 있다:

[0090] i) 염색체를 일정구간(bin)으로 구분하는 단계;

[0091] ii) 각 구간에 정렬된 핵산단편 사이의 거리(Fragments Distance, FD)값을 계산하는 단계;

[0092] iii) 각 구간별로 계산된 거리값을 기반으로 각 구간의 거리의 대표값(RepFD)을 결정하는 단계;

[0093] iv) 상기 iii) 단계에서 계산된 대표값을 전체 핵산단편 거리 값의 대표값으로 나누어 정규화(normalization)하는 단계; 및

[0094] iv) 각 구간의 순서를 X 축 값으로 하고, 상기 iv) 단계에서 계산한 정규화 값을 Y축 값으로 하여 GC plot을 생성하는 단계.

[0095] 본 발명에서 상기 GC plot은 염색체 1번부터 22번까지의 GC plot을 Y축으로 정렬하여 하나의 이미지를 생성하거

나, 1번부터 22번까지 생성한 이미지를 z 축으로 합쳐서 사용할 수 있다.

[0097] 본 발명에 있어서, 상기 대표값(RepFD)은 핵산단편 사이의 거리의 합, 차, 곱, 평균, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대편차, 변동계수 이들의 역수값 및 이들의 조합으로 구성된 군에서 선택되는 하나 이상인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0098] 본 발명에 있어서, 상기 일정구간(bin)은 1Kb 내지 3Gb인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0100] 본 발명에서는 핵산단편을 그룹화하는 단계를 추가로 사용할 수 있으며, 이때 그룹화 기준은, 정렬된 핵산단편의 어댑터 서열을 바탕으로 수행할 수 있다. 정방향으로 정렬된 핵산단편과 역방향으로 정렬된 핵산단편으로 별도로 구분하여서 선별된 서열정보에 대해서 핵산단편 사이의 거리를 계산할 수 있다.

[0101] 본 발명에서, 상기 FD 값은 수득한 n개의 핵산 단편에 대하여, i 번째 핵산 단편과 i+1 내지 n 번째 핵산 단편에서 선택되는 어느 하나 이상의 핵산 단편의 기준값 사이의 거리로 정의되는 것을 특징으로 할 수 있다.

[0102] 본 발명에서, 상기 FD 값은 수득한 n개의 핵산 단편에 대하여, 제1핵산 단편과 제2내지 제n개 핵산 단편으로 구성된 군에서 선택되는 어느 하나 이상의 핵산 단편의 기준값과의 거리를 계산하여 이들의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값과, 가중치가 포함된 계산 결과 및 이에 한정되지 않는 통계치를 FD 값으로 사용할 수 있으나 이에 한정되는 것은 아니다.

[0103] 본 발명에서 “하나 이상의 값 및/또는 하나 이상의 이들의 역수값”이라는 기재는 앞서 기재된 수치값들 중에서 하나 또는 2 이상이 조합되어 사용될 수 있다는 의미로 해석된다.

[0104] 본 발명에서, 상기 “핵산 단편의 기준값”은 핵산 단편의 중앙값으로부터 임의의 값을 더하거나 뺀 값인 것인 것을 특징으로 할 수 있다.

[0106] 상기 FD 값은 수득한 n개의 핵산 단편에 대하여, 다음과 같이 정의 할 수 있다.

[0107] FD = Dist(Ri~Rj) (1<i<j<n),

[0108] 여기서 Dist 함수는 선별된 Ri와 Rj 두 핵산 단편 사이에 포함되는 모든 핵산 단편의 정렬 위치값 차이의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값과, 가중치가 포함된 계산 결과 및 이에 한정되지 않는 통계치를 계산한다.

[0109] 즉, 본 발명에서 FD 값(Fragment Distance Value)는 정렬된 핵산 단편 사이의 거리를 의미한다. 여기서 거리 계산을 위한 핵산 단편의 선별 경우의 수는 다음과 같이 정의 할 수 있다. 총 N개의 핵산 단편이 존재할 경우 $\sum_{k=1}^{n-1} k$ 개의 핵산 단편 간 거리 조합이 가능하다. 즉, i가 1일 경우, i+1은 2가 되어 2 내지 n 번째 핵산 단편에서 선택되는 어느 하나 이상의 핵산 단편과의 거리를 정의 할 수 있다.

[0110] 본 발명에 있어서, 상기 FD 값은 상기 i 번째 핵산 단편 내부의 특정 위치와 i+1 내지 n 번째 중 어느 하나 이상의 핵산 단편 내부의 특정 위치 사이의 거리를 계산하는 것을 특징으로 할 수 있다.

[0111] 예를 들어 어떤 핵산 단편의 길이가 50bp 이며, 염색체 1번의 4,183 위치에 정렬 되었다고 하면, 이 핵산 단편의 거리 계산에 사용할 수 있는 유전적 위치값은 염색체 1번의 4,183 ~ 4,232 이다.

[0112] 상기 핵산 단편과 인접한 50bp 길이의 핵산 단편이 염색체 1번의 4,232번째 위치에 정렬되면, 이 핵산 단편의 거리 계산에 사용할 수 있는 유전적 위치값은 염색체 1번의 4,232 ~ 4,281이고, 두 핵산 단편 사이의 FD 값은 1에서 99가 될 수 있다.

[0113] 또 다른 인접한 50bp 길이의 핵산 단편이 염색체 1번의 4123번째 위치에 정렬되면, 이 핵산 단편의 거리 계산에 사용할 수 있는 유전적 위치값은 염색체 1번의 4,123~4,172이며, 두 핵산 단편 사이의 FD 값은 61에서 159이며,

첫 번째 예시 핵산 단편과의 FD 값은 12에서 110으로, 상기 두 FD 값 범위의 한 값의 합, 차, 곱, 평균, 곱의 로그, 합의 로그, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값, 및 가중치가 포함된 계산결과 및 이에 한정되지 않는 통계치를 FD 값으로 사용할 수 있으며, 바람직하게는 두 FD값 범위의 한 값의 역수값인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다

[0115] 바람직하게는 본 발명에 있어서, 상기 FD 값은 핵산 단편의 중앙값으로부터 임의의 값을 더하거나 뺀 값인 것을 특징으로 할 수 있다.

[0116] 본 발명에서, FD의 중앙값은 계산된 FD 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미한다. 예를 들어 1, 2, 100 과 같이 세 개의 값이 있을 때, 2가 가장 중앙에 있기 때문에 2가 중앙값이 된다. 만약 짝수 개의 FD 값이 있을 경우 가운데 있는 두 값의 평균으로 중앙값을 결정 한다. 예를 들어 1, 10, 90, 200의 FD 값이 있을 경우 중앙값은 10과 90의 평균인 50이 된다.

[0117] 본 발명에 있어서, 상기 임의의 값은 핵산 단편의 위치를 나타낼 수 있으면 제한없이 이용가능하나, 바람직하게는 0 내지 5 kbp 또는 핵산 단편 길이의 0 내지 300%, 0 내지 3 kbp 또는 핵산 단편 길이의 0 내지 200%, 0 내지 1 kbp 또는 핵산 단편 길이의 0 내지 100%, 더욱 바람직하게는 0 내지 500 bp 또는 핵산 단편 길이의 0 내지 50% 일 수 있으나, 이에 한정되는 것은 아니다.

[0118] 본 발명에 있어서, 상기 FD값은 페어드 엔드 시퀀싱(paired-end sequencing)일 경우, 정방향 및 역방향 서열정보(reads)의 위치값을 기반으로 도출하는 것을 특징으로 할 수 있다.

[0119] 예를 들어, 50bp 길이의 페어드 엔드 리드 쌍에서, 정방향 리드는 염색체 1번의 4183번째 위치에 정렬되고, 역방향 리드는 4349번째 위치에 정렬되면, 이 핵산단편의 양 말단은 4183, 4349가 되고, 핵산 단편 거리에 사용할 수 있는 기준값은 4183~4349이다. 이 때 상기 핵산 단편과 인접한 다른 페어드 엔드 리드 쌍에서, 정방향 리드는 염색체 1번의 4349번째 위치에 정렬되고, 역방향 리드는 4515번째 정렬되면, 이 핵산 단편의 위치 값은 4349~4515 이다. 이 두 핵산 단편의 거리는 0~333이 될 수 있고, 가장 바람직하게 각 핵산 단편의 중앙값의 거리인 166이 될 수 있다.

[0120] 본 발명에 있어서, 상기 페어드 엔드 시퀀싱으로 서열정보를 수득할 경우, 서열정보(reads)의 정렬 점수가 기준값 미만인 핵산 단편의 경우, 계산과정에서 제외하는 단계를 추가로 포함하는 것을 특징으로 할 수 있다.

[0122] 본 발명에 있어서, 상기 FD 값은 싱글 엔드 시퀀싱(single-end sequencing)일 경우, 정방향 또는 역방향 서열정보(read)의 위치값 한 종류를 기반으로 도출하는 것을 특징으로 할 수 있다.

[0123] 본 발명에 있어서, 상기 싱글 엔드 시퀀싱의 경우, 정방향으로 정렬된 서열정보를 기반으로 위치값을 도출할 경우에는 임의의 값을 더해주며, 역방향으로 정렬된 서열정보를 기반으로 위치값을 도출할 경우에는 임의의 값을 빼주는 것을 특징으로 할 수 있으며, 상기 임의의 값은 FD 값이 핵산 단편의 위치를 명확하게 나타내도록 하는 값이면 제한없이 이용가능하나, 바람직하게는 0 내지 5kbp 또는 핵산 단편 길이의 0 내지 300%, 0 내지 3kbp 또는 핵산 단편 길이의 0 내지 200%, 0 내지 1kbp 또는 핵산 단편 길이의 0 내지 100%, 더욱 바람직하게는 0 내지 500bp 또는 핵산 단편 길이의 0 내지 50% 일 수 있으나, 이에 한정되는 것은 아니다.

[0125] 본 발명에서 분석하고자 하는 핵산은 시퀀싱 되어 리드(reads)라는 단위로 표현될 수 있다. 이 리드는 시퀀싱 방법에 따라 싱글 엔드 시퀀싱(single end sequencing read, SE) 및 페어드 엔드 시퀀싱(paired end sequencing read, PE)으로 나눌 수 있다. SE 방식의 리드는 핵산 분자의 5`과 3` 중 한 곳을 랜덤한 방향으로 일정한 길이만큼 시퀀싱 한 것을 의미하고, PE 방식의 리드는 5`과 3` 을 모두 일정 한 길이만큼 시퀀싱 하게 된다. 이러한 차이 때문에, SE 모드로 시퀀싱 할 경우 한 개의 핵산 단편으로부터 1개의 리드가 생기고, PE 모드에서는 1개의 핵산 단편으로부터 2개의 리드가 쌍으로 생성되는 것은 통상의 기술자에게 잘 알려진 사실이다.

[0126] 핵산 단편 사이의 정확한 거리를 계산 하기 위한 가장 이상적인 방식은 핵산 분자를 처음부터 끝까지 시퀀싱하고, 그 리드를 정렬하고, 정렬된 값의 중앙값(센터)을 이용하는 것이다. 그러나 기술적으로 위 방식은 시퀀싱 기술의 한계 및 비용적인 측면 때문에 제약이 있는 실정이다. 따라서 SE, PE와 같은 방식으로 시퀀싱을 하게 되

는데, PE 방식의 경우 핵산 분자의 시작과 끝 위치를 알 수 있기 때문에 이 값들의 조합을 통해 핵산 단편의 정확한 위치(중앙값)를 파악 할 수 있으나, SE 방식의 경우 핵산 단편의 한쪽 끝 정보만을 이용할 수 있기 때문에 정확한 위치(중앙값) 계산에 한계가 있다.

[0127] 또한 정방향, 역방향의 양 "붬袖막* 시퀀싱 된(정렬된), 모든 리드의 말단 정보를 이용해 핵산 분자의 거리 계산시, 시퀀싱 방향 이라는 요소 때문에 정확하지 않은 값이 계산 될 수 있다.

[0128] 따라서, 시퀀싱 방식의 기술적 이유로 정방향 리드의 5`말단은, 핵산 분자의 중심 위치 보다 작은 위치 값을 갖고, 역방향 리드의 3`말단은 큰 값을 갖게 된다. 이러한 특징을 이용해, 정방향 리드의 경우 임의의 값(Extended bp)을 더해주고, 역방향 리드는 빼주게 되면 핵산 분자의 중심 위치에 가까운 값을 추정 할 수 있다.

[0129] 즉, 임의의 값(Extended bp)은 사용하는 시료에 따라 달라질 수 있으며, 세포유리 핵산의 경우 그 핵산의 평균 길이가 166bp 정도로 알려져 있기 때문에 약 80bp 정도로 설정을 할 수 있다. 만일 단편화 장비(ex; sonication)를 통해 실험이 진행 된 경우 단편화 과정에서 설정한 타겟 길이의 절반 정도를 extended bp으로 설정 할 수 있다.

[0130] 본 발명에 있어서, 상기 대표값(RepFD)은 FD 값의 합, 차, 곱, 평균, 중앙값, 분위수, 최소값, 최대값, 분산, 표준편차, 중앙값 절대 편차 및 변동 계수로 구성된 군에서 선택된 하나 이상의 값 및/또는 하나 이상의 이들의 역수값인 것을 특징으로 할 수 있으며, 바람직하게는 FD 값들의 중앙값, 평균값 또는 이의 역수값인 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0132] 본 발명에 있어서, 상기 (d) 단계의 인공지능 모델은 정상인 이미지와 암이 있는 이미지를 구별할 수 있도록 학습할 수 있는 모델이면 제한없이 사용가능하며, 바람직하게는 딥러닝 모델인 것을 특징으로 할 수 있다.

[0134] 본 발명에 있어서, 상기 인공지능 모델은 인공신경망 기반으로 벡터화된 데이터를 분석할 수 있는 인공신경망 알고리즘이면 제한없이 이용할 수 있으나, 바람직하게는 합성곱 신경망(convolutional neural network, CNN), 심층 신경망(Deep Neural Network, DNN), 순환 신경망(Recurrent Neural Network, RNN) 및 오토 인코더(autoencoder)로 구성된 군에서 선택되는 것을 특징으로 할 수 있으나, 이에 한정되는 것은 아니다.

[0136] 본 발명에 있어서, 상기 순환 신경망은 LSTM(Long-short term memory) 신경망, GRU(Gated Recurrent Unit) 신경망, 바닐라 순환 신경망(Vanilla recurrent neural network) 및 집중적 순환 신경망(attentive recurrent neural network)으로 구성된 군에서 선택되는 것을 특징으로 할 수 있다.

[0138] 본 발명에 있어서, 상기 인공지능 모델이 CNN일 경우, binary classification을 수행하는 손실함수는 하기 수식 1로 표시되는 것을 특징으로 할 수 있고, Multi-class classification을 수행하는 손실함수는 하기 수식 2로 표시되는 것을 특징으로 할 수 있다.

[0139] 수식 1: Binary classification

[0140] $$\text{loss}(\text{model}(x),y) = -\frac{1}{n}[\sum_{i=1}^{n}(y_i \log(model(x_i)) + (1-y_i)\log(1-model(x_i)))]$$

[0141] $\text{model}(x_i) = i$번째 input에 인공지능 model output

[0142] y = 실제 label 값

[0143] n = input data 수

[0144] 수식 2: Multi-class classification

[0145] $$\text{loss(model(x),y)} = -\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{c}(y_{ij}\log(model(x_i))_j)$$

[0146] $\text{model}(x_i)_j = i\text{번째 input에 } j\text{번째 인공지능 model output}$

[0147] y = 실제 label 값

[0148] n = input data 수

[0149] c = class 수

[0150] 본 발명에서, 상기 binary classification은 인공지능 모델이 암 유무를 판별하도록 학습하는 것을 의미하며, multi-class classification은 인공지능 모델이 암 종을 판별하도록 학습하는 것을 의미한다.

[0152] 본 발명에서, 상기 인공지능 모델이 CNN일 경우, 학습은 하기 단계를 포함하여 수행되는 것을 특징으로 할 수 있다:

[0153] i) 생산된 GC plot을 training(학습), validation(검증), test(성능평가) 데이터로 분류하는 단계;

[0154] 이 때, Training 데이터는 CNN 모델을 학습할 때 사용되고, Validation 데이터는 hyper-parameter tuning 검증에 사용되며, Test 데이터는 최적의 모델 생산 후, 성능 평가로 사용되는 것을 특징으로 함.

[0155] ii) Hyper-parameter tuning 및 학습 과정을 통해서 최적의 CNN 모델을 구축하는 단계; 및

[0156] iii) Hyper-parameter tuning을 통해서 얻어진 여러 모델의 성능을 validation data를 이용하여 비교하여, validation data 성능이 가장 좋은 모델을 최적의 모델로 결정하는 단계;

[0158] 본 발명에서, 상기 Hyper-parameter tuning 과정은 CNN 모델을 이루는 여러 parameter(convolution layer 수, dense layer 수, convolution filter 수 등) 값을 최적화 하는 과정으로 Hyper-parameter tuning 과정으로는 Bayesian optimization 및 grid search 기법을 사용하는 것을 특징으로 할 수 있다.

[0159] 본 발명에서, 상기 학습 과정은 정해진 hyper-parameter들을 이용하여 CNN 모델의 내부 parameter(weights)들을 최적화 시켜, Training loss 대비 validation loss가 증가하기 시작하면 모델이 과적합(Overfitting) 되었다 판단하고, 그전에 model 학습을 중단하는 것을 특징으로 할 수 있다.

[0161] 본 발명에 있어서, 상기 (d) 단계에서 인공지능 모델이 입력된 벡터화된 데이터로부터 분석한 결과값은 특정 score 또는 실수이면 제한없이 이용가능하며, 바람직하게는 DPI(Deep Probability Index) 값인 것을 특징으로 할 수 있으나 이에 한정되는 것은 아니다.

[0163] 본 발명에서, Deep probability Index는 인공지능 model의 마지막 layer에 binary classification일 경우 sigmoid function, multi-class classification일 경우 softmax function을 사용하여 인공지능의 output을 0 ~ 1 scale로 조정하여 확률값으로 표현한 값을 의미한다.

[0164] Binary classification일 경우에는 sigmoid function을 이용하여 암 일 경우 DPI 값이 1이 되게끔 학습을 하게 된다. 예를 들어, 신경모세포종 샘플과 정상 샘플이 입력되면, 신경모세포종 샘플의 DPI 값이 1에 가깝도록 학습하는 것이다.

[0165] Multi-class classification 일 경우에는 softmax function을 이용하여, class 개수만큼의 DPI 값을 뽑게 된다. Class 개수만큼의 DPI값의 합은 1이되고, 실제 해당되는 암 종의 DPI값이 1이 되게끔 학습을 하게 된다.

예를 들어, 3개의 class 신경모세포종, 간암, 정상이 있고, 신경모세포종 sample이 들어오면, 유방암 class를 1에 가깝게 학습하게 되는 것이다.

[0166] 본 발명에서 상기 (d) 단계의 출력 결과값은 암 종별로 도출되는 것을 특징으로 할수 있다.

[0168] 본 발명에서, 상기 인공지능 모델은 학습할 때, 암이 있으면 output 결과가 1에 가깝게 학습하고, 암이 없으면 output 결과가 0에 가깝게 학습을 시켜서, 0.5를 기준으로 0.5 이상이면 암이 있다고 판단하고, 0.5 이하이면 암이 없다고 판단하여 performance 측정을 수행하였다(Training, validation, test accuracy).

[0169] 여기서, 0.5의 기준값은 언제든지 바뀔 수 있는 값이라는 것은 통상의 기술자에게 자명한 것이다. 예를 들어서 False positive(위양성)를 줄이고자 하면, 0.5보다 높은 기준값을 설정하여 암이 있다고 판단되는 기준을 엄격하게 가져 갈 수 있고, False Negative(위음성)를 줄이고자 하면 기준값을 더 낮게 측정하여 암이 있다고 판단되는 기준을 조금 더 약하게 가져 갈 수 있다.

[0170] 가장 바람직하게는 학습된 인공지능 모델을 이용하여 unseen data(학습에 training하지 않은 답을 알고 있는 data)를 적용시켜서, DPI값의 probability를 확인하여 기준값을 정할 수 있다.

[0172] 본 발명에 있어서, 상기 (e) 단계의 출력 결과값 비교를 통해 암 종을 예측하는 단계는 출력 결과값 중, 가장 높은 값을 나타내는 암 종을 샘플의 암으로 판정하는 단계를 포함하는 방법으로 수행하는 것을 특징으로 할 수 있다.

[0174] 본 발명은 다른 관점에서, 생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 해독하는 해독부;

[0175] 해독된 서열을 표준 염색체 서열 데이터베이스에 정렬하는 정렬부;

[0176] 정렬된 서열 기반의 핵산단편을 이용하여 벡터화된 데이터를 생성하는 데이터 생성부;

[0177] 생성된 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석하고, 기준값과 비교하여 암 유무를 판정하는 암 진단부; 및

[0178] 출력된 결과값을 분석하여 암 종을 예측하는 암 종 예측부를 포함하는 암 진단 및 암 종 예측 장치를 포함하는 암 진단 및 암 종 예측장치에 관한 것이다.

[0180] 본 발명에서 상기 해독부는 독립된 장치에서 수행될 수 있다. 예를 들어, 본 발명의 해독부는 NGS 장치에서 메틸화 정보가 포함된 서열정보 즉, read를 생산할 수 있다.

[0182] 본 발명은 또 다른 관점에서, 컴퓨터 판독 가능한 저장 매체로서, 암 진단 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하되,

[0183] (a) 생체시료에서 핵산을 추출하여 메틸화 정보가 포함된 서열정보를 획득하는 단계;

[0184] (b) 획득한 서열정보(reads)를 표준 염색체 서열 데이터베이스(reference genome database)에 정렬(alignment)하는 단계;

[0185] (c) 상기 정렬된 서열정보(reads) 기반의 핵산단편을 이용하여 벡터화된 데이터를 생성하는단계;

[0186] (d) 생성된 상기 벡터화된 데이터를 학습된 인공지능 모델에 입력하여 분석한 출력 결과값과 기준값(cut-off value)을 비교하여 암 유무를 판정하는 단계; 및

[0187] (e) 상기 출력 결과값 비교를 통해 암 종을 예측하는 단계를 통하여, 암 유무 및 암 종을 예측하는 프로세서에 의해 실행되도록 구성되는 명령을 포함하는 컴퓨터 판독 가능한 저장 매체에 관한 것이다.

[0189] 다른 양태에서 본원에 따른 방법은 컴퓨터를 이용하여 구현될 수 있다. 일 구현예에서, 컴퓨터는 칩 세트에 연

결된 하나 이상의 프로세서를 포함한다. 또한 칩 세트에는 메모리, 저장 장치, 키보드, 그래픽 어댑터(Graphics Adapter), 포인팅 장치(Pointing Device) 및 네트워크 어댑터(Network Adapter) 등이 연결되어 있다. 일 구현예에서, 상기 칩 세트의 성능은 메모리 컨트롤러 허브(Memory Controller Hub) 및 I/O 컨트롤러 허브에 의하여 가능하다. 다른 구현예에서, 상기 메모리는 칩 세트 대신에 프로세서에 직접 연결되어 사용될 수 있다. 저장 장치는 하드 드라이브, CD-ROM(Compact Disk Read-Only Memory), DVD 또는 기타 메모리 장치를 포함하는 데이터를 유지할 수 있는 임의의 장치이다. 메모리는 프로세서에 의하여 사용된 데이터 및 명령에 관여한다. 상기 포인팅 디바이스는 마우스, 트랙볼 (Track Ball) 또는 다른 유형의 포인팅 디바이스일 수 있고, 키보드와 조합하여 입력 데이터를 컴퓨터 시스템으로 전송하는데 사용된다. 상기 그래픽 어댑터는 디스플레이 상에서 이미지 및 다른 정보를 나타낸다. 상기 네트워크 어댑터는 근거리 또는 장거리 통신망으로 컴퓨터 시스템과 연결된다. 본원에 사용되는 컴퓨터는 하지만 위와 같은 구성으로 제한되는 것은 아니고, 일부 구성이 없거나, 추가의 구성을 포함 할 수 있으며, 또한 저장장치영역네트워크(Storage Area Network, SAN)의 일부일 수 있으며, 본원의 컴퓨터는 본원에 따른 방법의 수행을 위한 프로그램에 모듈의 실행에 적합하도록 구성될 수 있다.

[0191] 본원에서 모듈이라 함은, 본원에 따른 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 모듈은 소정의 코드와 상기 소정의 코드가 수행되기 위한 하드웨어 리소스(Resource)의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나, 한 종류의 하드웨어를 의미하는 것은 아님은 본원 기술분야의 당업자에게 자명한 것이다.

[0193] 본원에 따른 방법은 하드웨어, 펌웨어, 또는 소프트웨어 또는 이들의 조합으로 구현될 수 있다. 소프트웨어로 구현되는 경우 저장매체는 컴퓨터와 같은 장치에 의해 판독가능한 형태의 저장 또는 전달하는 임의의 매체를 포함한다. 예를 들면 컴퓨터 판독가능한 매체는 ROM(Read Only Memory); RAM(Random Access Memory); 자기디스크 저장 매체; 광저장 매체; 플래쉬 메모리 장치 및 기타 전기적, 광학적 또는 음향적 신호 전달 매체 등을 포함한다.

[0195] **실시예**

[0196] 이하, 실시예를 통하여 본 발명을 더욱 상세히 설명하고자 한다. 이들 실시예는 오로지 본 발명을 예시하기 위한 것으로서, 본 발명의 범위가 이들 실시예에 의해 제한되는 것으로 해석되지는 않는 것은 당업계에서 통상의 지식을 가진 자에게 있어서 자명할 것이다.

[0198] **실시예 1. 혈액에서 메틸화된 cfDNA를 추출하여, 차세대 염기서열 분석 수행**

[0199] 정상인 185명 및 신경모세포종(neuroblastoma) 환자 57명의 혈액dmf 채취 후 3000 rpm, 25℃10분의 조건으로 혈장 부분만 1차 원심분리한 다음, 1차 원심분리된 혈장을 16000g, 25℃10분의 조건으로 2차 원심분리하여 침전물을 제외한 혈장 상층액을 분리하였다. 분리된 혈장에 대해 chemagen DNA kit 사용하여 cell-free DNA를 추출하고, Truseq Nano DNA HT library prep kit (Illumina)를 사용해 우선 adaptor ligation 과정까지 수행한 다음, cfMediIP kit (diagnode)의 antibody를 이용해 10 rpm, 4℃17시간 반응하여 5mC immunoprecipitation을 진행한 다음 purification을 진행하고, 다시 Truseq Nano DNA HT library prep kit (Illumina)를 사용해 PCR enrichment를 진행하여, 최종적인 library를 제작하였다. 제작한 library는 Novaseq 6000 (Illumina) 를 150 paired-end 모드로 sequencing을 진행하여 샘플 당 약 30 million 개의 reads 를 생산하였다.

[0201] **실시예 2. 핵산단편 수 기반 GC plot 생성**

[0202] 실시예 1에서 수득한 리드를 bwa(version 0.7.17-r1188) alignment tool을 이용하여 핵산 단편 데이터를 참조 유전체에 정렬한 다음, biobambam2 bammarkduplicates (version 2.0.87) tool을 이용하여 PCR duplicate 핵산 단편을 제거하고, sambamba(version 0.6.6)을 이용하여 mapping quality가 60 이하인 핵산 단편을 제거하였다.

[0203] GC plot은 염색체의 시작부터 말단까지 NGS reads가 정렬된 상태를 표현하는데, 성염색체를 제외한 모든 염색체를 겹치지 않는 100 kilobase bin으로 나눈다음, 각 bin에 할당된 reads의 수를 세었다(read count value). 각

bin에 할당된 reads count 값을 샘플의 전체 reads 수로 나누어 정규화(Normalization) 과정을 진행하였다. 정규화된 bin read count값을 Y값, 각 bin의 순서를 X값으로 놓고 염색체 별로 GC plot 생산하고, 생산한 GC plot을 1번 염색체부터 22번 염색체까지 정렬하여 1개의 이미지를 생산하였다(도 2).

[0205] **실시예 3. GC plot 기반 신경모세포종 딥러닝 모델 구축 및 DPI 계산**

[0206] 3-1. 딥러닝 모델 구축

[0207] 실시예 2에서 생산한 GC plot을 training(학습), validation(검증), test(성능평가) 데이터로 나누어, Training 데이터는 CNN 모델을 학습할 때 사용하고, Validation 데이터는 hyper-parameter tuning 검증에 사용하였으며, Test 데이터는 최적의 모델 생산 후, 성능 평가로 사용하였다.

[0208] Tensorflow (version 2.4.1)을 이용하여 CNN 모델 구축 및 학습에 사용하였는데, CNN 모델에 구조는 convolution layer -> pooling layer -> fully connected layer 순으로 이루어져 있고, convolution layer 다음에는 항상 pooling layer가 삽입되어 있다. Convolution layer 개수 및 fully connected layer에 개수는 hyper-parameter tuning 과정을 통해서 결정하였다. 모델을 학습할 때, 손실함수(loss function)를 최소화 하는 방향으로 학습 진행하였으며, 손실함수는 수식 1 및 수식 2와 같다.

[0209] 최적의 모델을 구하기 위해 scikit-optimize (version 0.7.4) python package를 이용하여 hyper-parameter tuning하였다. Hyper-parameter tuning 과정은 CNN 모델을 이루는 여러 parameter(convolution layer 수, dense layer 수, convolution filter 수 등) 값을 최적화 하는 과정으로, Convolution layer 개수, convolution filter 개수, convolution patch size, fully connected layer 개수, hidden node 개수, activation function, dropout 유무, learning rate를 hyper-parameter로 지정 후 Bayesian optimization 기법을 이용하여 최적의 모델 구축하였으며, 정해진 hyper-parameter로 모델을 학습할 때 Training loss 대비 validation loss가 증가하기 시작하면 모델이 과적합(Overfitting) 되었다 판단되어 그전에 model 학습 중단하였고, Hyper-parameter tuning 과정에서 얻어진 여러 모델의 성능을 validation data를 이용하여 비교 후 성능이 가장 좋은 모델을 최적의 모델이라 판단하고, test data를 이용해 성능 평가 진행하였다.

[0210] 3-2. DPI(Deep Probability Index) 계산

[0211] Hyper-parameter tuning을 통해서 구해진 최적의 모델에 데이터 (GC plot)을 넣어주면, 모델의 output layer를 통해서 확률 값이 출력된다.

[0212] 먼저, Binary classification일 경우에는 모델의 output layer에 sigmoid function을 사용하였다. Sigmod fuction은 하기 수식 3과 같다.

[0213] 수식 3 Sigmoid function: $\frac{1}{1+e^{-x}}$,

[0214] $x = output\ \ layer$

[0215] 수식 3에서 출력되는 확률 값(DPI)은 하나로서, 암의 진단에 사용하였다.

[0216] Multi-class classification일 경우에는 모델의 output layer에 softmax function을 사용하였으며, 수식 4와 같다.

[0217] 수식 4 Softmax function: $\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$,

[0218] $C = \text{Class 개수}, \quad x_i = \text{i 번째 class에 output layer}$

[0219] 수식 4에서 출력되는 확률값(DPI)은 class 개수만큼 출력되므로, 암의 종류 구분에 사용하였다.

[0221] **실시예 4. 메틸화된 cfDNA를 이용한 핵산단편 수 기반 GC plot의 신경모세포종 딥러닝 모델 구축 및 성능 확인**

[0222] 정상 샘플 (n=186)과 신경모세포종 샘플 (n=57) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train,

Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

표 1

[0223]

	Normal	Neuroblastoma	Total
Train	130	39	169
Valid	28	9	37
Test	28	9	37
Total	186	57	243

[0224] 그 결과, 표 2, 도 3 및 4에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 100 %, 92%, 94.1 %인 것을 확인하였고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 1.0, 0.95, 0.99인 것을 확인하였다.

[0225] 도 3은 정확도를 측정하는 방법 중 ROC(Receiver Operating Characteristic) curve 를 활용한 분석으로, 커브 아래 면적의 넓이인 AUC(Area Under the Curve) 값이 높을 수록 정확도가 높다고 해석한다. AUC 값은 0-1 사이의 값을 갖으며, 랜덤으로 레이블 값을 예측했을 때 (baseline) 기대되는 AUC 값이 0.5, 완벽히 정확하게 예측했을 때 기대되는 AUC 값이 1이다.

[0226] 도 4는 본 발명의 인공지능 모델에서 계산된 암이 있을 확률값(DPI 값을 정상 샘플과 신경모세포종 샘플 그룹에서 boxplot으로 나타낸 것으로, 붉은 선이 DPI cutoff인 0.5를 나타낸다.

표 2

[0227]

	Accuracy	AUC
Train	98.8%	1.0
Valid	94.6%	0.95
Test	94.6%	0.94

[0229] **실시예 5. cfDNA를 이용한 핵산단편 수 기반 GC plot의 신경모세포종 딥러닝 모델 구축 및 성능 확인**

[0230] 5-1. 혈액에서 DNA를 추출하여, 차세대 염기서열 분석 수행

[0231] 정상인 186명 및 신경모세포종 환자 57명의 혈액을 10mL씩 채취하여 EDTA Tube에 보관하였으며, 채취 후 2시간 이내에 1200g, 4℃분의 조건으로 혈장 부분만 1차 원심분리한 다음, 1차 원심분리된 혈장을 16000g, 4℃분의 조건으로 2차 원심분리하여 침전물을 제외한 혈장 상층액을 분리하였다. 분리된 혈장에 대해 Tiangenmicro DNA kit (Tiangen)을 사용하여 cell-free DNA를 추출하고, MGIEasy cell-free DNA library prep set kit 를 사용하여 library preparation 과정을 수행 한 다음, DNBseq G400 장비 (MGI) 를 100 base Paired end 모드로 sequencing 하였다. 그 결과, 샘플 당 약 170 million 개의 reads가 생산되는 것을 확인 하였다.

[0232] 5-2. 딥러닝 모델 구축 및 성능확인

[0233] 정상 샘플 (n=186)과 신경모세포종 샘플 (n=57) 을 이용해 DPI 값의 성능을 테스트 하였다. 모든 샘플은 Train, Validation, Test 그룹으로 나눠 진행했고, Train 샘플을 이용하여 model을 구축한 다음 Validation 그룹 및 Test 그룹의 샘플을 이용해서, Train 샘플을 이용해 만든 모델의 성능을 확인하였다.

표 3

[0234]

	Normal	Neuroblastoma	Total
Train	130	39	169
Valid	28	9	37
Test	28	9	37
Total	186	57	243

[0235] 그 결과, 표 4, 도 6 및 7에 기재된 바와 같이, Accuracy 는 Train, Valid, Test 그룹에서 각각 92.9 %, 97.3%, 94.6 %인 것을 확인하였고, ROC 분석 결과인 AUC 값은 Train, Valid, Test 그룹에서 각각 0.98, 0.98, 0.95인 것을 확인하였다.

표 4

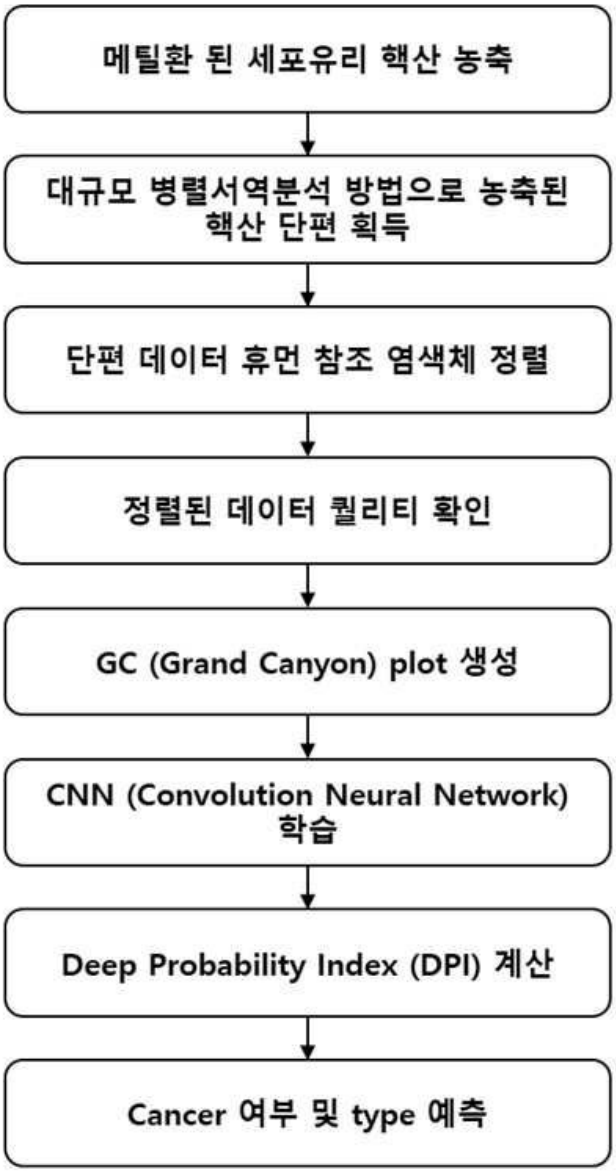
[0236]

	Accuracy	AUC
Train	92.9%	0.98
Valid	97.3%	0.98
Test	94.6%	0.95

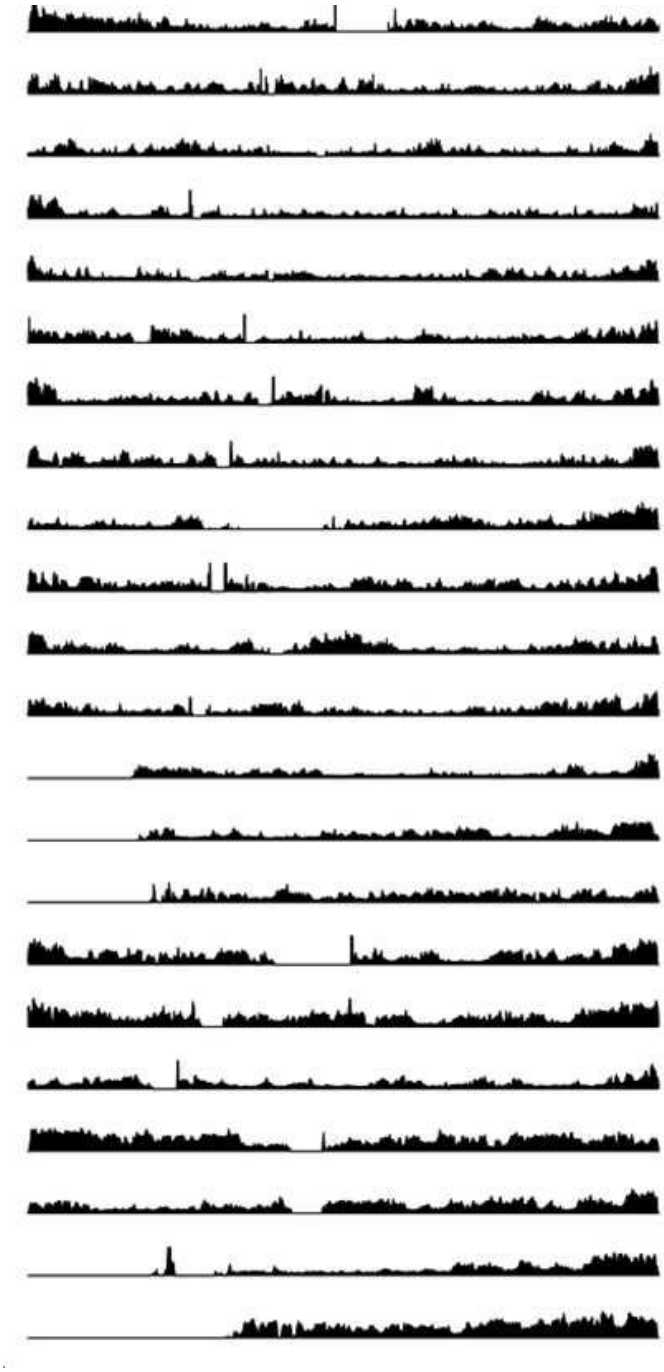
[0237] 이상으로 본 발명 내용의 특정한 부분을 상세히 기술하였는 바, 당업계의 통상의 지식을 가진 자에게 있어서 이러한 구체적 기술은 단지 바람직한 실시 양태일 뿐이며, 이에 의해 본 발명의 범위가 제한되는 것이 아닌 점은 명백할 것이다. 따라서, 본 발명의 실질적인 범위는 첨부된 청구항들과 그것들의 등가물에 의하여 정의된다고 할 것이다.

도면

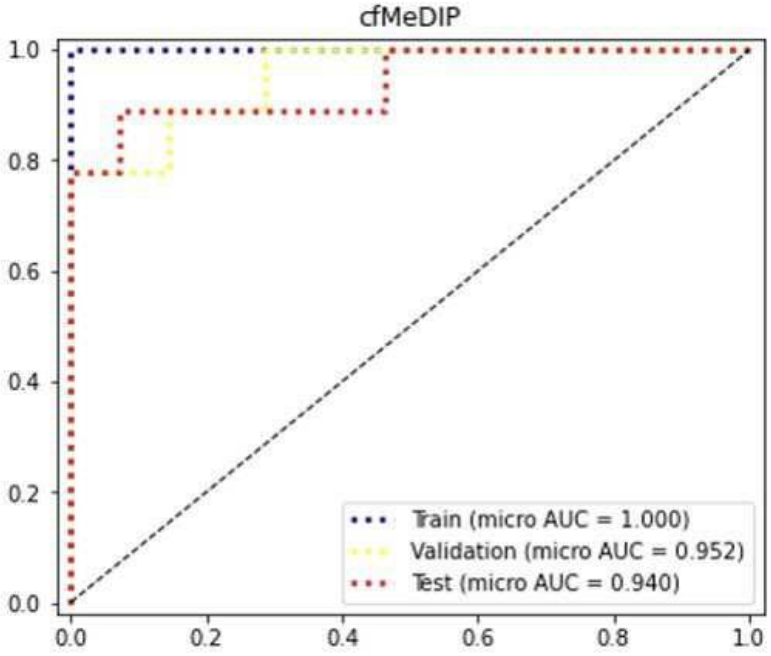
도면1


메틸환 된 세포유리 핵산 농축
대규모 병렬서역분석 방법으로 농축된 핵산 단편 획득
단편 데이터 휴먼 참조 염색체 정렬
정렬된 데이터 퀄리티 확인
GC (Grand Canyon) plot 생성
CNN (Convolution Neural Network) 학습
Deep Probability Index (DPI) 계산
Cancer 여부 및 type 예측

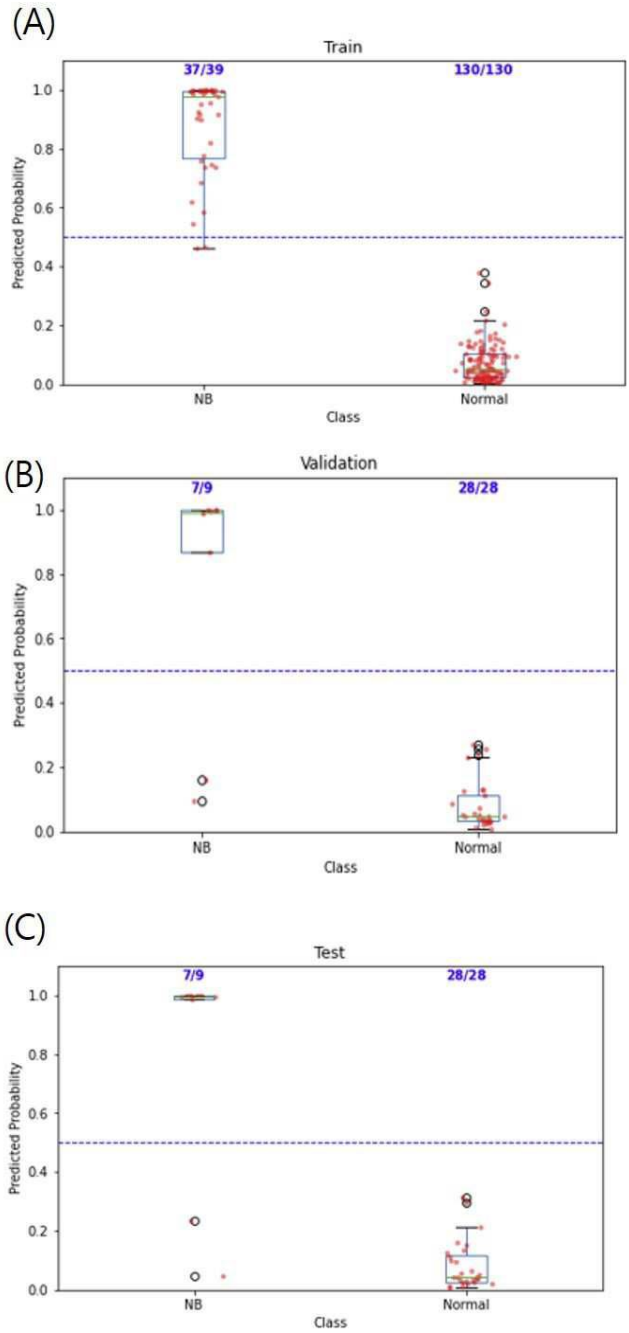
도면2



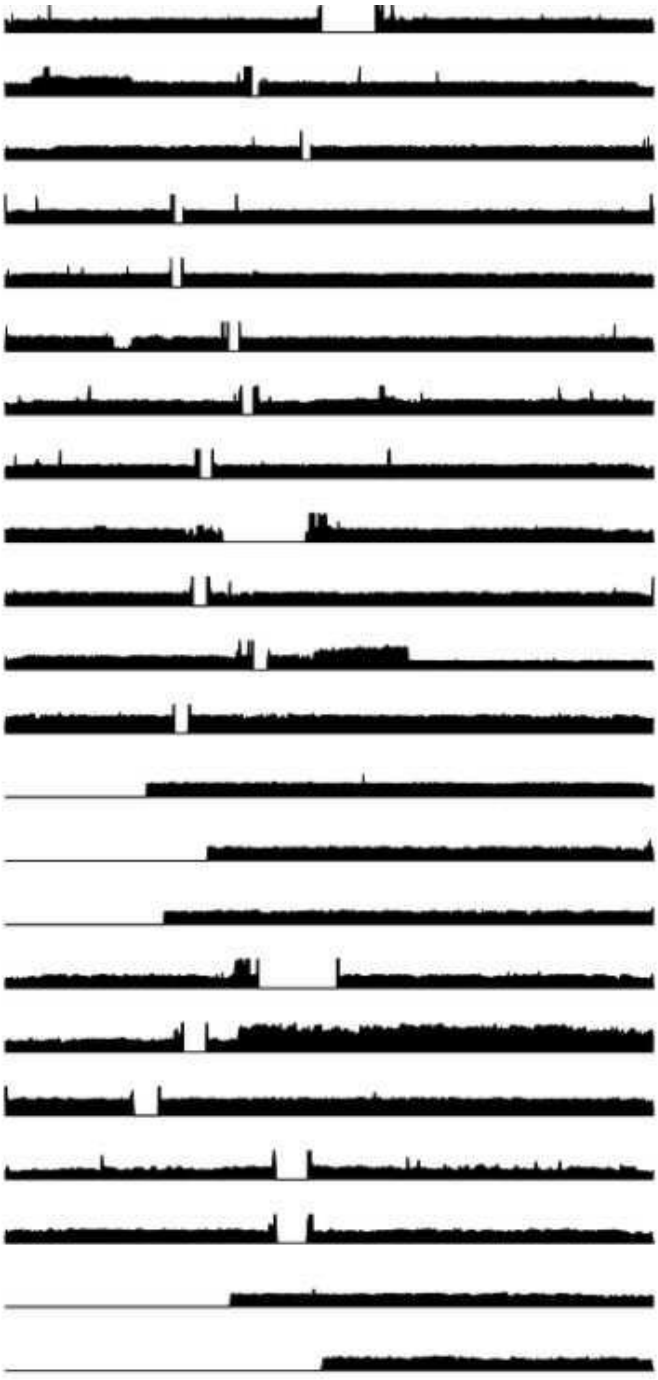
도면3


cfMeDIP
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Train (micro AUC = 1.000)
Validation (micro AUC = 0.952)
Test (micro AUC = 0.940)

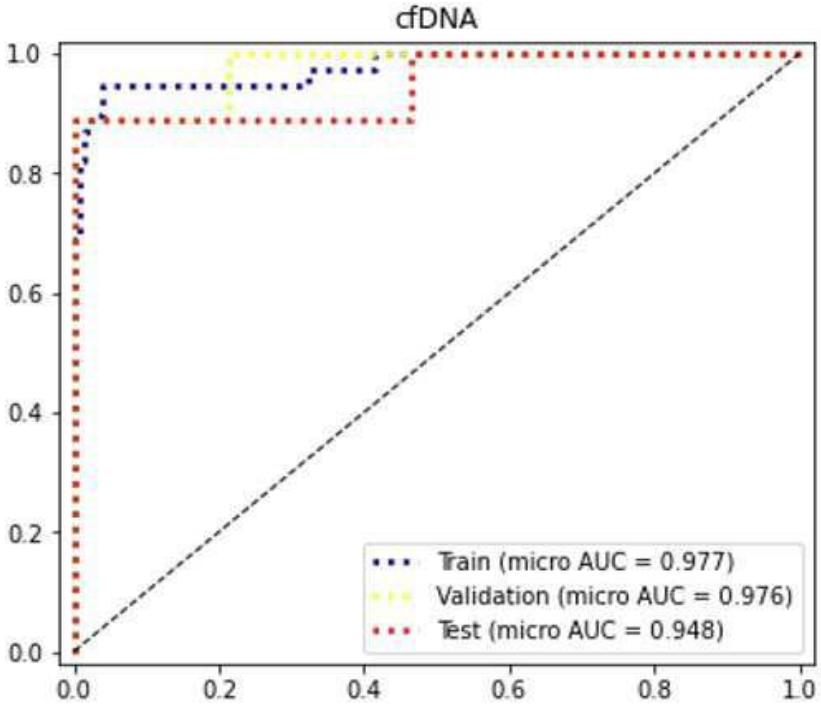
도면4


(A)
Train
37/39
130/130
Predicted Probability
1.0
0.8
0.6
0.4
0.2
0.0
NB
Normal
Class
(B)
Validation
7/9
28/28
Predicted Probability
1.0
0.8
0.6
0.4
0.2
0.0
NB
Normal
Class
(C)
Test
7/9
28/28
Predicted Probability
1.0
0.8
0.6
0.4
0.2
0.0
NB
Normal
Class

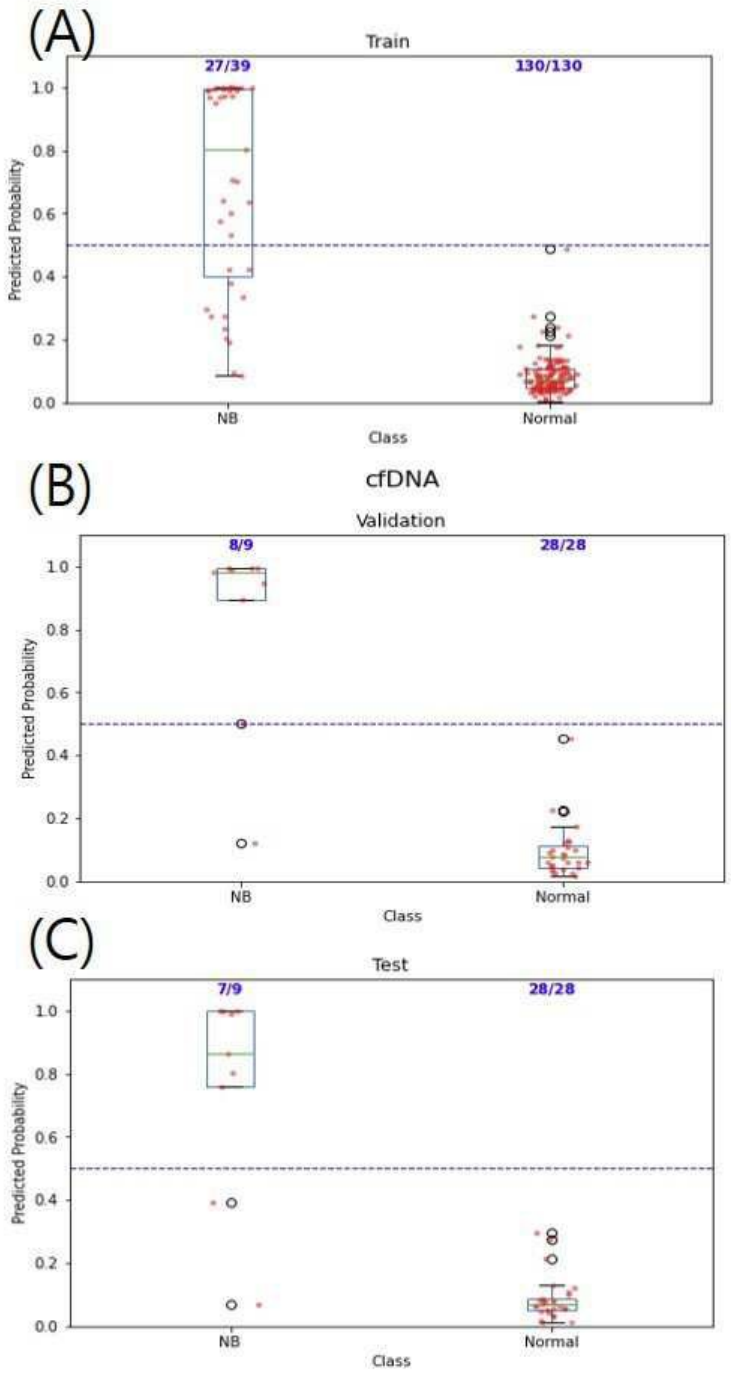
도면5



도면6


cfDNA
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Train (micro AUC = 0.977)
Validation (micro AUC = 0.976)
Test (micro AUC = 0.948)

도면7


(A)
Train
27/39
130/130
Predicted Probability
NB
Normal
Class
(B)
cfDNA
Validation
8/9
28/28
Predicted Probability
NB
Normal
Class
(C)
Test
7/9
28/28
Predicted Probability
NB
Normal
Class